ENCYCLOPÆDIA
BRITANNICA

MACROPÆDIA

The Encyclopædia Britannica
is published with the editorial advice
of the faculties of the University of Chicago;
a committee of persons holding
academic appointments at the universities
of Oxford, Cambridge, London, and Edinburgh;
a committee at the University of Toronto;
and committees drawn from members of the faculties
of the University of Tokyo
and the Australian National University.

THE UNIVERSITY OF CHICAGO

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

The New Encyclopædia Britannica

in 30 Volumes

MACROPÆDIA
Volume 6

Knowledge in Depth

FOUNDED 1768
15TH EDITION

Encyclopædia Britannica, Inc.
William Benton, Publisher, 1943–1973
Helen Hemingway Benton, Publisher, 1973–1974
Chicago/London/Toronto/Geneva/Sydney/Tokyo/Manila/Seoul

©1979
by Encyclopædia Britannica, Inc.
Copyright under International Copyright Union
All rights reserved under Pan American and
Universal Copyright Conventions
by Encyclopædia Britannica, Inc.

Printed in U.S.A.

Library of Congress Catalog Card Number: 77-94292
International Standard Book Number: 0-85229-339-9

Earth, Figure of

The definition of the figure of the Earth—that is, its size and shape—usually does not involve the description of mountains and valleys but, rather, the size and shape of the mean sea-level surface and its continuation under the land. This hypothetical surface, called a geoid, is a reference surface from which topographic heights and ocean depths are measured. Because of the irregular mass distributions inside the Earth, the geoid is not a simple mathematical surface and consequently is not a suitable reference surface for a geometric figure of the Earth.

The geoid

As reference figures of the Earth, but not for its topography, simple geometric forms are used that approximate the geoid. For many purposes an adequate geometrical representation of the Earth is a sphere, for which only the radius of the sphere must be stated. When a more accurate reference figure is required, an ellipsoid of revolution is used as a representation of the shape and the size of the Earth. It is a surface generated by rotating an ellipse 360° about its minor axis. An ellipsoid that is used in geodetic calculations to represent the Earth is called a reference ellipsoid. This ellipsoid of revolution is the shape most frequently used to represent a simple geometric reference surface, although a triaxial ellipsoid has been suggested. The latter is an ellipsoid that has three mutually perpendicular unequal axes and thus requires three parameters (quantities) for definition. Geodetic calculations using a triaxial ellipsoid are unnecessarily complicated, and its use as a reference surface has been limited. Theoretically, more complicated mathematical models could be used as a geometric representation of the Earth, but in practice they have been found to be impractical for ordinary geodetic computations. It is more convenient to represent the deviations from an ellipsoid of revolution and take them into account in computations than it is to use more complicated representation; on the other hand, the presentation of the surface of the Earth in a three-dimensional Cartesian coordinate system is used in some modern geodetic work. When there are three-dimensional coordinates for a large number of points evenly distributed around the world, the information reflects the shape and the size of the Earth.

An ellipsoid of revolution is specified by two parameters, a semi-major axis (equatorial radius for the Earth) and a semi-minor axis (polar radius), or the flattening. Flattening (f) is defined as the difference in magnitude between the semi-major axis (a) and the semi-minor axis (b) divided by the semi-major axis, or $f = (a - b)/a$. For the Earth, the semi-major axis and semi-minor axis differ about 21 kilometres (13 miles) and the flattening is about one part in 300. The departures of the geoid from the best fitting ellipsoid of revolution are about ± 100 metres; the difference between the two semi-axes of the equatorial ellipse in the case of a three-axial ellipsoid fitting the Earth is only about 80 metres. This shows why the simple ellipsoid of revolution is the most used geometric representation of the size and shape of the Earth.

One of the main reasons for knowing the shape and size of the Earth is to establish a relative position of two or more points in the same reference system. This is achieved by establishing a geodetic datum, or reference point, that fixes the origin of the computational system and the surface over which the computations are to be made. A large number of points given in a geodetic datum form what is called a control network. For practical work, such as mapping relatively large areas of the globe, this type of control network is essential. In the construction of interstate highway systems, the location of ships at sea, or the establishment of real property lines, for example, uniform geodetic control plays an important role. In addition, a soundly based geodetic system is important in the general field of navigation and inertial guidance, including space travel and tracking satellites. Once the locations of points in a uniform geodetic system are known, analyses can be made of positional variations of points as a function of time, recording such phenomena as continental drift and sea-floor spreading (*qq.v.*).

The control network

This article provides a historical review of the determination of the Earth's figure and treats world geodetic systems based on gravimetric, satellite, and other kinds of measurements. International agreements on geodetic values are treated in the final section. For further information on the Earth's internal properties that are of concern in physical geodesy, see EARTH, GRAVITATIONAL FIELD OF and EARTH, STRUCTURE AND COMPOSITION OF. Major relief features that cause departure of the Earth's figure from the reference geoid or reference ellipsoid are covered in the article EARTH, PHYSIOGRAPHY OF. See also SURVEYING; MAPS AND MAPPING; and HYDROGRAPHIC CHARTING for treatment of the application of geodetic reference systems in practice.

HISTORICAL REVIEW

Spherical era. *The ancients.* It is astonishing that ancient civilized people such as the Babylonians, who were active in astronomy only a few thousand years ago, believed that the Earth is flat. Some of them surely must have noticed that the Earth's shadow creeping across the Moon at an eclipse is circular. They should have noticed also that by travelling farther south they could see stars that were below the horizon at home. Almost anyone should have noticed that a low island seen from the shoreline disappeared when they went swimming. Homer, however, about 900–800 BC, did describe the Earth as a convex disk surrounded by the Oceanus stream, but, according to others of that time, the Earth is a circular plate supported by four elephants standing on a big sea turtle. Although there is now some suggestion that the Egyptian builders of the Great Pyramid at Giza (*c.* 2550 BC) incorporated information on the spherical size and shape of the Earth into dimensions of the pyramid, credit for the idea that the Earth is spherical is usually given to Pythagoras (born 532 BC) and his school, who reasoned that because the Moon and the Sun are spherical, therefore, the Earth is, too. Notable among other Greek philosophers, Hipparchus and Aristotle (4th century BC) came to the same conclusion. Aristotle devoted a part of his book *De caelo* to the defense of the doctrine. He also estimated that the circumference of the Earth is about 400,000 stadia (one Greek stadium is 185.2 metres, or 607 feet), almost twice the true value. This seems to be the first scientific attempt to estimate the size of the Earth. Eratosthenes (3rd century BC), however, is considered to be one of the founders of geodesy because he was the first to describe and apply a scientific measuring technique for determining the size of the Earth. He used a simple principle of estimating the size of a great circle passing through the North and South Poles (Figure 1). Knowing the length of an arc (l) and size of the corresponding central angle (α) that it subtends, the radius of the sphere can be obtained from the simple proportion that length of arc to size of the great circle (or circumference, $2\pi R$, in which R is the Earth's radius) equals

Eratosthenes' measurement

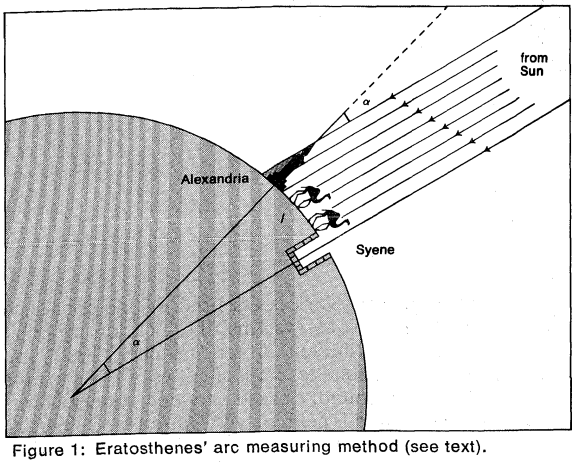


Figure 1: Eratosthenes' arc measuring method (see text).

central angle to the angle subtended by the whole circumference (360°).

$$l : 2\pi R = \alpha^\circ : 360^\circ$$

In order to determine the central angle α, Eratosthenes selected the city of Syene (the modern Aswān on the Nile) because there the Sun in midsummer shone at noon vertically into a well (see Figure 1). He assumed that all sunrays reaching the Earth were parallel to one another, and he observed that the sunrays at Alexandria at the same time (midsummer at noontime) were not vertical but lay at an angle 1/50 of a complete revolution of the Earth away from the zenith. Possibly using the travel time of a camel caravan between Alexandria and Syene, he estimated the distance (l) between these two cities to be 5,000 stadia. From the above equation he obtained for the length of a great circle, $50 \times 5{,}000 = 250{,}000$ stadia, which is 46,300,000 metres (152,000,000 feet). The result is about 15 percent too large in comparison to modern measurements, but his result was extremely good considering the assumptions and the equipment with which the observations were made.

The second known determination of the size of the Earth was made by Poseidonius (1st century BC). He used the distance between Alexandria and Rhodes Island, where the star Canopus was at the horizon at the same time, when it was 1/48 of a complete revolution of the Earth above the horizon in Alexandria. The distance was estimated using the time it took a sailboat to travel between these two points. He encountered many problems, such as those caused by refraction, but his result was only 11 percent too large.

The introduction of triangulation. A new era in determining the size of the Earth began through the introduction of triangulation. The idea of triangulation was apparently conceived by the Danish astronomer Tycho Brahe before the end of the 16th century, but it was developed as a science by a contemporary Dutch mathematician, Willebrord van Roijen Snell. Snell used a chain of 33 triangles to determine the length of an arc essentially in the customary way used today. The resulting size of the Earth, however, was 3.4 percent too small. The idea of triangulation is to establish a network of stations that form connecting triangles. One side of the first triangle in the chain, called the baseline, and all angles of the triangles are accurately measured. Using the law of sines from spherical trigonometry, the lengths of all sides thus can be computed starting from a known baseline. When the lengths and angles are known, coordinates can be computed for each point provided that the coordinates of one point and one azimuth are known. Triangulation points are usually placed on the tops of the hills because the neighbouring points must be clearly visible. Commonly, more complicated figures such as quadrilaterals with diagonals are used in triangulation.

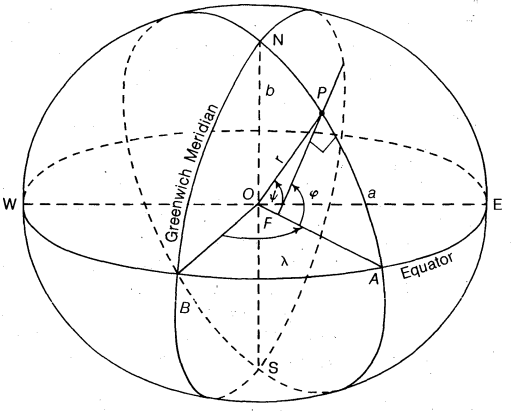
In 1669, Jean Picard, a French astronomer, first used a telescope in the determination of latitude and in measurement of angles in triangulation that consisted of 13 triangles and extended from Paris 1.2° northward. His observations and results were extremely important because his length of arc on a great circle corresponding to 1° was used by Newton in his theoretical calculations to prove that the attraction of the Earth is the principal force governing the motion of the Moon in its orbit.

Ellipsoidal era. The period from Eratosthenes to Picard can be called the spherical era of geodesy. A new ellipsoidal era was begun by Sir Isaac Newton and Christiaan Huygens. In the Ptolemaic astronomy it had seemed natural to assume that the Earth was an exact sphere with a centre that, in turn, all too easily became regarded as the centre of the entire universe. But, with growing conviction that the Copernican system is true—the Earth moves around the Sun and rotates about its own axis—and with the advance in mechanical knowledge due chiefly to Newton and Huygens, it seemed natural to conceive the Earth as an oblate spheroid—that is, a sphere flattened at the poles. In one of the many brilliant analyses in his *Principia*, published in 1687, Newton deduced the Earth's shape theoretically and found that the equatorial semi-axis would be 1/230 longer than the polar semi-axis (true value about 1/300).

First estimate of flattening

Experimental evidence supporting this idea came in 1672 as the result of a French expedition to Guiana. The explorers found that a pendulum clock that kept good time in Paris lost 2½ minutes a day at Cayenne near the Equator. At that time no one knew how to interpret the observation; but Newton's theory that gravity must be larger at the poles (because of closer proximity to Earth's centre) than at the Equator was a logical explanation.

It is possible to determine whether or not the Earth is an oblate spheroid by measuring the length of an arc corresponding to a geodetic latitude difference at two places along the meridian (the ellipse passing through the poles) at different latitudes, which means at different distances from the Equator. This can be seen from Figure 2,



$\overline{OE} = a$ = semi-major axis
$\overline{ON} = b$ = semi-minor axis
$\angle PFA = \varphi$ = geodetic latitude
$\angle POA = \psi$ = geocentric latitude
$\angle BOA = \lambda$ = geodetic longitude
$OP = r$ = radius vector
N and S = poles
FP normal to ellipsoid at P

Figure 2: Elements of a reference ellipsoid (see text).

in which the geodetic latitude at any point (P) is represented by the angle made between a line perpendicular to the ellipsoidal surface at the point P and the equatorial plane. This angle differs from the geocentric latitude that is determined by a line directed from the point P toward the Earth's centre. Such measurements of arc were made by Jean-Dominique Cassini and his son Jacques Dominique in France by continuing the arc of Picard north to Dunkirk and south to the boundary of Spain. Surprisingly, the result of that experiment in 1720 showed the length of a meridian degree north of Paris to be 111,017 metres or 265 metres shorter than one south of Paris (111,282 metres). This suggested that the Earth is a prolate spheroid, not flattened at the poles, but elongated, with the equatorial axis shorter than the polar axis. This was completely at odds with Newton's conclusions.

In order to settle the controversy caused by Newton's theoretical derivations and the measurements of Cassini, the French Academy of Sciences sent two expeditions, one to Peru led by Pierre Bouguer and Charles-Marie de La Condamine to measure the length of a meridian degree in 1735 and another to Lapland in 1736 under Pierre-Louis Moreau de Maupertuis to make similar measurements. Both parties determined the length of the arcs using the method of triangulation. Only one baseline, 14.3 kilometres (8.9 miles) long, was measured in Lapland, and two baselines, 12.2 and 10.3 kilometres (7.6 and 6.4 miles) long, were used in Peru. Astronomic observations for latitude determinations from which the size of the angles was computed were made using the zenith sectors having radii up to four metres (13 feet). The expedition to Lapland returned in 1737, and Maupertuis reported that the length of one degree of the meridian in Lapland was 57,437.9 toise. This result, when compared to the corresponding value of 57,060 toise near Paris, proved that the Earth was flattened at the poles. Later, large errors were found in the measurements, but they were in the "right direction."

After the expedition returned from Peru in 1743, Bouguer and La Condamine could not agree on one common interpretation of the observations, mainly because of the use of two baselines and lack of suitable computing techniques. The mean values of the two lengths calculated by them gave the length of the degree as 56,753 toise, which confirmed the earlier finding of the flattening of the Earth. As a combined result of both expeditions, these values have been reported in the literature: semi-major axis, $a = 6{,}397{,}300$ metres; flattening, $f = 1/216.8$.

Numerous arc measurements were subsequently made, one of which was the historic French measurement used for definition of a unit of length. In 1791 the French National Assembly adopted the new length unit, called the metre and defined as 1:10,000,000 part of the meridian quadrant from the Equator to the pole along the meridian that runs through Paris. For this purpose a new and more accurate arc measurement was carried out between Dunkirk and Barcelona in 1792–98. These measurements combined with those from the Peruvian expedition yielded a value of 6,376,428 metres (20,914,684 feet) for the semi-major axis and for the flattening, 1/311.5 (more modern values are 6,378,160 metres [20,920,365 feet] and 1/298.25), which made the metre 0.02 percent "too short" from the intended definition.

Adoption of the metre unit

The length of the semi-major axis, a, and flattening, f, continued to be determined by the arc method but with some modification for the next 200 years. Gradually, instruments and methods improved and the results became more accurate. Interpretation of the results was made easier through introduction of the statistical method of least squares.

Geoidal era. The above-mentioned method of determination of the size and shape of the Earth would have worked properly if the mass of the Earth were uniformly distributed. The plumb line used as reference in astronomic observation would then be normal (perpendicular) to the mathematical reference surface, the ellipsoid of revolution, because for an ellipsoid with a homogeneous distribution of mass, a plumb bob would hang truly perpendicular to the surface. Mass irregularities exist, however, and therefore the plumb line and the normal to the ellipsoid cannot be considered as the same direction. The angle between the direction that a plumb line actually takes and the direction of a corresponding normal to the reference ellipsoid at the point is called a deflection of the vertical. It is usually divided into two components: ξ, meridian component, north–south direction, and η, a prime vertical component, east–west direction. The plumb line is perpendicular to the geopotential surface going through the point; this means that at sea level a plumb line is perpendicular to mean sea level. The mean sea level and its imagined continuation under the land is a good approximation of the geoid. The deflection of the vertical is caused by mass anomalies—visible, such as mountains or oceans, or invisible mass surpluses or deficiencies. These mass irregularities have an effect on gravity.

When an ellipsoid is considered as an equipotential surface for gravity, defined as the combination of gravitational and centrifugal accelerations, there are two additional parameters to be defined. It has been customary to use the symbols γ_0 and m as these parameters, in which γ_0 is the value of gravity at the Equator and m is the ratio of the centrifugal acceleration at the Equator to γ_0. These parameters are related so that the parameter m is equal to the square of the rate (ω) of rotation of the Earth with respect to inertial space times the semi-major radius (a) divided by the gravity at the Equator; *i.e.*, $\omega^2 a/\gamma_0$. The parameters γ_0 and m are connected to the mass of the Earth (M), together with the constant of gravitation (G) and the flattening (f), by:

$$GM = a^2\gamma_0 \left[1 - f + \frac{3}{2}m - \frac{15}{14}mf + 0(f^3)\right].$$

The product GM is called the geocentric gravitational constant. The term $0(f^3)$ symbolizes a remainder of neglected terms. The formula for the theoretical gravity (γ) at geodetic latitude ϕ is expressed as a sum of sine functions:

$$\gamma = \gamma_0 \left[1 + \beta_2 \sin^2\phi + \beta_4 \sin^2 2\phi + 0(f^3)\right],$$

in which the factors β_2 and β_4 have the values

$$\beta_2 = \frac{5}{2}m - f - \frac{17}{14}mf, \text{ and } \beta_4 = \frac{1}{8}f^2 - \frac{5}{8}mf.$$

Gravity anomalies

By measuring gravity at a point and taking differences between the measured gravity (g) and the value (γ) computed for the same point using a theoretical gravity formula for the reference ellipsoid, and a proper reduction, a gravity anomaly (Δg) is obtained: $\Delta g = g - \gamma$. The basic mathematical formula for computations of undulations of the geoid surface about the ellipsoid surface, N, using gravity anomalies was derived by an English physicist, George Stokes, in 1849. There is no direct relationship between the gravity anomaly Δg at a point and the geoid undulation N at the same point, and an integration over the entire surface of the Earth must be performed. Corresponding formulas for computations of the deflections of the vertical from gravity anomalies were derived by a Dutch geophysicist, Felix Andries Vening Meinesz, in 1928. These derivations are important because they are the basic computational formulas for gravimetric geodesy. Unfortunately, these formulas require that gravity anomalies be known over the entire surface of the Earth; therefore they had hardly any practical value at the time of their derivations.

Because of deflections of the plumb line, latitude differences computed from astronomic observations and from geodetic computations are not equal. The difference between astronomic latitude and geodetic latitude (see above *Ellipsoidal era*) at a point is called the astrogeodetic deflection component in meridian, and the corresponding difference in longitudes, properly scaled, is called the astrogeodetic deflection component in prime vertical. Deflections of the plumb line have caused tremendous difficulties in computations of the size and shape of the Earth using measured arcs and astronomic observations, because the central angle computed from astronomic observations is not the corresponding central angle of the arc on an ellipsoid.

During the triangulation in India close to the Himalayas, the difference in astrogeodetic deflections of the vertical in the meridian at the Kaliana and Kalianpur stations was observed to be 5.24 seconds of arc. The effect of the visible topographic masses was computed to cause a difference of 15.88 seconds of arc. From this, J.H. Pratt in 1854 surmised that the mean density of the crust under the mountains was less than under the lowlands, to compensate the attraction of the mass of the mountains to the plumb-line direction. Pratt explained that mountains rose from the subcrust area like a fermenting dough, the density of which would be smaller as it rose higher. This is one of the principles of isostatic equilibrium that was developed from the measurements in a scientific sense. An English astronomer, George Biddell Airy, came to a similar conclusion in 1855, although

he explained isostatic equilibrium in a different way. He reasoned that the high mountains of Asia had sunk in a more dense substratum and were floating, just as timber or icebergs would float in water. According to Airy's theory, mountains have "roots" and oceans have "anti-roots" to compensate for mass surpluses and mass deficiencies. Airy's theory is more acceptable today than Pratt's explanation. The depth of the compensation according to current evidence is between 30 and 40 kilometres (20 and 25 miles).

In order to improve the accuracy in determinations of the size and shape of the Earth, it was recognized that the deflections of the vertical have to be taken into account in the computations. Because there were not enough gravity observations around the world, isostatic-topographic deflections of the vertical were computed using topographic information and various assumptions of isostasy. Simultaneously, the arc method was expanded to the area method; for example, in 1910 U.S. geodesist J.F. Hayford computed the dimensions of the ellipsoid based on arc measurements covering the entire United States. His results were considered sufficiently accurate to be adopted as the International Ellipsoid by the International Association of Geodesy in 1924, chiefly because of the inclusion of the deflection of the vertical information into computations. Later, it was found that this determination was not the most accurate of the time.

As stated above, there were several determinations of the dimensions of the reference ellipsoid between 1800 and 1960. Some of these results are given in Table 1. The ellipsoids of Bessel, Clarke, Hayford, and Krassowski have been generally used as reference surfaces in geodesy.

Table 1: Early Determinations of Dimensions of the Earth Ellipsoid from Arc Measurements

author	year	*a* (semi-major axis; in metres)	1/*f*
P. Bouguer, P.-L.-M. de Maupertuis	···	6,397,300	216.80
J.-B. Delambre	1800	6,375,653	334.00
G.B. Airy	1830	6,376,542	299.30
F.W. Bessel	1841	6,377,397	299.15
A.R. Clarke	1857	6,378,345	294.26
J.H. Pratt	1863	6,378,245	295.30
A.R. Clarke	1866	6,378,206	295.00
I. Bonsdorff	1888	6,378,444	298.60
J.F. Hayford	1906	6,378,283	297.80
F.A. Helmert	1907	6,378,200	298.30
J.F. Hayford	1910	6,378,388	297.00
W.A. Heiskanen	1926	6,378,397*	—
Krassowski	1938	6,378,245	298.30
H. Jeffreys	1948	6,378,099	297.10
Ledersteger	1951	6,378,298*	—
U.S. Army Map Service	1956	6,378,260*	—

*Value of arc was determined from $f = 1/297.00$.

WORLD GEODETIC SYSTEM

Astrogravimetric method. *Gravimetric geodesy.* As stated above, the irregular deflections of the plumb line have caused problems for geodesists in determination of geodetic positions and, consequently, in all computations related to the size and the shape of the Earth. Even if a reasonable dimension of a reference ellipsoid had been established, deflections would have caused problems in establishing a geodetic datum from a geocentric datum.

Geodetic datum

A geodetic datum, normally referred to simply as datum, is defined by the location of the central point (initial point or datum origin) with respect to the centre of the reference ellipsoid selected. This has been customarily achieved by astronomic latitude, longitude, and azimuth determinations at the initial point; by defining the size and shape of the reference ellipsoid; and, preferably, by fulfilling the condition through the Laplace equation that the rotation axis of the ellipsoid be parallel to the mean rotation axis of the Earth. If the deflection components and undulation of the geoid are absolute (*i.e.*, not relative), then the datum is geocentric. Because the determination of these absolute values at the initial point of a datum is difficult, each country commonly has had its own relative geodetic datum. If a country is not too large in area, any reasonable reference ellipsoid and datum are satisfactory for practical mapping and control system of the country. Because there was no way of determining differences between the geodetic systems located on different continents or islands, it was difficult to utilize all geodetic measurements simultaneously for determination of the shape and size of the Earth. It was necessary, therefore, to find a method that could bring geodetic datums into one worldwide system.

The utilization of isostatic-topographic deflections of the vertical would have brought all geodetic datums into one system provided that the Earth had been in isostatic equilibrium according to the assumption used; however, with observed gravity data it has been shown that this is not the case. The alternative method to be considered was to make gravity observations around the entire surface of the Earth and compute the absolute deflections of the vertical and relative undulations of the geoid for orientation of all geodetic datums. If the gravimetric deflection components ξ_g and η_g are known, the astronomic latitude ϕ', longitude λ', and azimuth A' can be converted to the geodetic quantities ϕ, λ, and A through the formulas:

$$\phi = \phi' - \xi_g,$$
$$\lambda = \lambda' - \eta_g \sec \phi,$$
$$A = A' - \eta_g \tan \phi.$$

This is the principle of the astrogravimetric method for determination of coordinates of control points in a uniform world geodetic system. When these quantities and N, undulations of the geoid, are known at several points in each geodetic datum, the existing geodetic systems can be converted into a single world geodetic system. It was recognized that a common scale for all geodetic measurements was a prerequisite to a world system. This was provided by the light interference method, invented by the Finnish astronomer Yrjö Väisälä, and put into practice in 1923 by the Finnish Geodetic Institute. The standard baselines have been observed at the following locations: Nummela, Finland; Buenos Aires; Loenermark, The Netherlands; Ebersberg, West Germany; Potsdam, East Germany; Matadas Virtudes, Portugal; and Mansfield, Ohio. The reported accuracy of the measurements is up to 1:17 million for 864 metres (2,834 feet) distance.

The astrogravimetric method seemed to be the only feasible way of getting a world geodetic system during the pre-satellite era. This seemed to become possible also because of extensive developments of instrumentation in measuring absolute and relative gravity values. In absolute measurements, it is necessary to measure actual gravity at the observation point. In relative measurements it is necessary to measure only the difference, or the ratio, between g_0 at the base station and the gravity g_i at the field station. In the Stokes and Vening Meinesz formulas only relative gravity values are needed. Until about 1935 all gravity measurements were made with pendulum apparatuses. The principle is a simple one: if the length and the period of a pendulum are measured, the gravity (g) can be obtained from the formula that gravity (g) is equal to a constant ($4\pi^2$) times the pendulum length (l) divided by the period (T) squared; *i.e.*, $g = 4\pi^2 l/T^2$. The period is relatively easy to determine, but the measurement of the length of the pendulum is much more difficult. For relative measurement or for the ratio, pendulum length is eliminated from the formula, and the formula takes the following simplified form: the ratio of the gravity at the field station (g_i) to the gravity at the base station (g_0) is inversely equal to the ratio of their periods (T_i and T_0) squared; *i.e.*, $g_i : g_0 = T_0^2 : T_i^2$. Thus the accuracy of the relative gravity measurements depends only upon the accuracy of the period-of-time measurements.

Pendulum measurements

The invention of an apparatus for gravity observations at sea (the Vening Meinesz pendulum) gave a boost to the gravimetric method, but with the invention of portable gravimeters a new era began in the gravity surveys. With pendulums it took several hours, sometimes even days, to make an observation; with new gravimeters, however, the measuring time was reduced to 10–15 min-

utes. Because of this speed, complete systematic gravity surveys were made in several countries.

Using the vast amount of data available, there were several attempts to compute the shape of the geoid. One of them was the result of the effort started by W.A. Heiskanen at Ohio State University (Columbus) in 1951. In this program all gravity observations around the world were collected and transferred into one uniform system related to the gravity value at Potsdam, East Germany, which was internationally accepted as the principal reference station. The first result from this program was the "Columbus geoid" published in 1957. Some of the results of the program indicated that the flattening of the Earth is less in the Northern Hemisphere than in the Southern Hemisphere, a phenomenon later confirmed by other methods. The attempts to create a world geodetic system, however, were hindered by lack of uniform distribution of gravity observations around the world. There have been tremendous developments in the theoretical side in physical geodesy since World War II, but lack of uniform distribution of gravity observations around the world has made it impossible to compute accurately the shape of the Earth. It has been necessary to develop other techniques in which all available information is combined in the system—information such as astrogeodetic deflections of vertical, astrogeodetic undulations, information from astronomic and geodetic measurements, and all information from gravity measurements. One of the most extensive combinations of this kind was made by the American geodesist William Kaula in 1959.

Other techniques. Parallel to the work in gravimetric geodesy, other developments occurred that have contributed to the determination of the size and shape of the Earth. Techniques in geodetic measurements were improved by the introduction of a new kind of instrumentation. In past years it was easier to get more accurate measurements of directions or angles than distances, but the development of the Geodimeter in the 1920s reversed the situation. It was possible to measure distances with the Geodimeter with an accuracy of one part per million. The instrument measures the time it takes light to travel back and forth between two points. The largest uncertainty in the measuring system comes from propagation anomalies caused by changing atmospheric conditions. This instrument, as well as others based on measuring the velocity of electromagnetic waves, has made a significant contribution to accuracy of geodetic networks.

Geodimeter measurements

Other techniques, such as utilizing the speed of the shadow of the Moon on the Earth (solar eclipse methods) or observing the moment when the stars disappear behind the Moon (occultation or Moon camera methods), were tried to establish a common world geodetic system with somewhat less success. Also, measurements of certain small irregularities in the motion of the Moon and observations of the Earth's axis in relation to the stars were utilized for determination of the Earth's flattening. Table 2 shows some of the results of other methods for the determination of the shape of the Earth.

Satellite measurements. *Dynamic method.* A new era began in October 1957, when the first satellite, Sputnik 1 of the Soviet Union, was launched. Artificial satellites close to the Earth offered a new and far more accurate instrument for measuring the Earth's flattening and size. The first results were better values of the flattening. The key to the technique that gave this kind of information was the observation of perturbations in the orbits of the satellites. If the Earth were a perfect equal-density-layer sphere and had no atmosphere, a satellite would pursue the identical orbit month after month when the small attractions of the Sun and Moon are ignored. Because the Earth is not a sphere, its gravitational pull varies as a function of latitude, and the orbit of the satellite is therefore distorted.

Undulations of the geoid

The geoid, the actual equipotential surface of the Earth, is generally expressed by its undulations (N) above or below the ellipsoid of revolution at the point for which the geodetic latitude and longitude (the geodetic longitude is the angle between the meridian on which the point is located and the Greenwich meridian) and φ and λ. Be-

Table 2: Earth Flattening Determined by Gravimetric or Astronomic Methods

author	year	$1/f$
F.A. Helmert	1884	299.25*
Ivanov	1889	297.2*
F.A. Helmert	1901	298.2*
W. Bowie	1912	298.4*
W. Bowie	1917	297.4*
Berroth	1916	297.4*
F.A. Helmert	1915	296.7*
W.A. Heiskanen	1924	297.4*
W.A. Heiskanen	1928	297.0*
W.A. Heiskanen	1938	298.2*
E. Niskanen	1945	297.8*
U.A. Uotila	1957	297.4*
U.A. Uotila	1961	298.1*
W. de Sitter	···	296.96†
W. de Sitter	···	296.75†
H. Spencer-Jones	1941	296.78†
E.C. Bullard	1948	297.34†
H. Jeffreys	1948	297.34†

*Gravimetrically. †Astronomically.

cause the flattening is only about 1/300 and N/r is, at maximum, 1.8×10^{-5}, simple spherical harmonics can be used to express undulations N:

$$N = \sum_{n=2}^{\infty} \sum_{m=0}^{n} P_{nm}(\sin \varphi)(L_{nm} \cos m\lambda + M_{nm} \sin m\lambda),$$

in which $P_{nm}(\sin \varphi)$ is the Legendre associated function, the solution of a differential equation proposed by the French mathematician Adrien-Marie Legendre, and L_{nm}, M_{nm} are the independent coefficients. Spherical harmonics are written as one equation, representing different solutions of Laplace's differential equation according to the coefficient used. Undulation N is related to the difference between actual gravitational potential (W) at a point and potential (U) of the ellipsoidal reference figure by:

$$N = \frac{W - U}{g} = \frac{T}{g},$$

in which g is the actual gravity on the geoid and T is the disturbing potential. The difference between g and γ is the gravity anomaly, $\Delta g = g - \gamma$. Because the geoid undulations N and the disturbance of the potential T are related, there is a set of potential spherical harmonic coefficients C_{nm}, S_{nm} equivalent to the undulation coefficients L_{nm}, M_{nm}. These coefficients are most convenient when the potential of gravitation is required:

$$V = \frac{GM}{r}\left[1 + \sum_{n=2}^{\infty} \sum_{m=0}^{n} \left(\frac{a}{r}\right)^n P_{nm}(\sin \psi)\right.$$
$$\left.(C_{nm} \cos m\lambda + S_{nm} \sin m\lambda)\right],$$

in which ψ is a geocentric latitude.

An often used alternative notation for C_{n0} is $-J_n$, for which n takes values 1, 2, 3, 4, etc. The most important of these coefficients is J_2, which is customarily related to the flattening f, under the assumption of an ellipsoidal equipotential by

$$J_2 = \frac{2}{3}f[1 - \frac{1}{2}f] - \frac{1}{3}m[1 - \frac{3}{2}m - \frac{2}{7}f] + 0(f^3).$$

The gravitational potential (V) can be used in the equations of motion of a close Earth satellite. Satellites behave according to Newtonian laws of mechanics. The six orbital elements necessary to refer the orbit to Earth are: the orbital inclination—the angle between the orbital plane and the Earth's equator; the orbital period—the elapsed time of one orbit of the Earth; the eccentricity—departure of the orbital ellipse from a circle; the argument of perigee—the angle of the northward crossing of the Earth's equator to perigee measured along the orbital ellipse; the longitude of the ascending node—the angle from the vernal equinox to the point of north passage across the Equator; and the time of perigee passage. When the parameters GM, a, the coefficients $C_{nm}S_{nm}$, ini-

tial position r_0, and velocity dr_0/dt are given, the orbit can be calculated for any specified duration of time, as has been seen, for example, with Gemini rocket flights.

When sufficient numbers of observations of positions and velocity are given, values for GM, a, C_{nm}, S_{nm}, r_0, and dr_0/dt can be inferred from the orbit. Practical limitations of the scheme are that in addition to the gravitational potential, satellite orbit perturbations are also caused by such things as atmospheric drag and attraction of the Sun and Moon, which should be filtered out. Also, inadequate distribution of tracking stations offers only irregular coverage of observations. Furthermore, errors in tracking-station positions affect the observations.

The orbit of Sputnik 1 was not as accurately determined as the orbit of Sputnik 2, which came a month later. Sputnik 2 was a bright object, and an extensive series of accurate observations were made with kinetheodolites and other optical instruments. From these observations, the orbit of the satellite was fairly accurately determined. The first satellites of the U.S., Explorer 1 and Vanguard 1, were launched early in 1958 and were tracked with minitrack radio systems. The satellites were moving eastward, and the flattening of the Earth made their orbits precess westward. The British and American results of observations of this drifting suggested that the flattening of the Earth is closer to 1/298 than to 1/297. This finding is in accordance with present analyses.

Encouraged with the early results, extensive observation programs were set up in the U.S. under the National Aeronautics and Space Administration. Observations for geodetic purposes were carried out mainly by the Smithsonian Astrophysical Observatory, and the satellites' orbits were analyzed for geodetic purposes.

Geometric method. The above application of satellite observations is called a dynamic method. Parallel to this method, the geometric method has been employed. The importance of geometric satellite geodesy has to be assessed by the fact that a geometrical solution of satellite triangulation provides geodesy, for the first time, with the means to create a three-dimensional, world-wide reference system without reference to either the direction or magnitude of the force of gravity. The main objective of geometric satellite geodesy is the establishment of three-dimensional positions of a selected number of points on the physical surface of the Earth. The feasibility of spatial triangulation of a selected number of non-intervisible points on the physical surface of the Earth, using auxiliary target points sufficiently elevated above the surface in connection with the star background, has been demonstrated in the last two decades by numerous scientists. It suffices to describe the basic geometric principles of Yrjö Väisälä's astronomical method of triangulation. The principle of the method is as follows: two conjugate rays emerging from the end points of line AB (see Figure 3) define a plane in space, the spatial orientation of which can be computed from the measured directions of the two rays. Point S_1 is a satellite position. If two such planes containing the line AB are observed, the direction of the line AB can be computed by intersecting the two planes under consideration. Five planes are sufficient for establishing a unique solution for a spatially oriented station triangle. Any one of these planes contains two stations and a specific satellite position. It is obvious that the principle can be applied to any number of triangles forming a three-dimensional triangulation scheme. This particular method calls for observations of satellites simultaneously at several stations but does not require precise knowledge of the satellite orbit.

Satellite triangulation

Using the principle of this method, a worldwide triangulation net has been worked out. The U.S. Echo satellites have been the main targets in this work. All observational work for this net has been finished and the analysis of the data became available in 1971–72. This work has been coordinated by Hellmut Schmid, of the National Ocean Survey (former U.S. Coast and Geodetic Survey). Expected accuracies for the three-dimensional coordinates are approximately plus or minus six metres.

There is also the orbital method, which is used to locate a point on the Earth's surface with reference to the Earth's centre of gravity. It requires continuous tracking and subsequent determinations of a satellite orbit. Tracking from several stations enables the linkup of the stations geocentrically as compared to the geodetic basis of the simultaneous observations. A variation of the orbital method, called a short-arc method, is used to connect geodetically unknown stations to known stations that are too far away for the satellite to be observed from all stations simultaneously. These methods have been employed for extension of the geodetic network by the U.S. Department of Defense and others. Similar programs are under way in eastern and western Europe. Short-arc methods also require continuous prediction of the orbit.

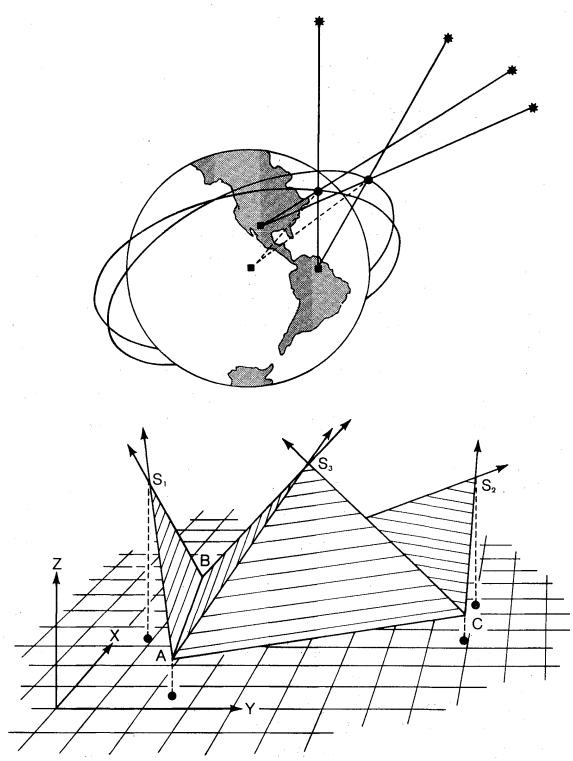


Figure 3: (Top) Simultaneous observation of satellite against star background photographed from two or more known stations and one unknown site. (Below) Principle of satellite triangulation (see text).

Combination of different types of data and methods. The ultimate aim in geodesy, of course, is to get the best uniform world geodetic system. It is logical that the system should be consistent from whatever point of view it is examined. To get this type of system, it may be best to include all available information related to the problem in a mathematical model and to formulate a consistent system. For example, by combining the results of a strictly geometrically executed worldwide satellite triangulation with the results of interpretations of satellite orbits, it is evident that a unified approach to the basic problem of geodesy can then be envisaged. Similarly, all other related data (such as surface gravity data) should be combined with the potential coefficients derived from satellite data. One of the latest results of this kind of combination is the geoid giving the undulations in metres and referring to the best fitting ellipsoid as having a flattening of 1/298.258. This geoid is a result of analyses and computations at the Smithsonian Astrophysical Observatory.

The results obtained from the observations of satellites and from combinations of data have confirmed that the flattening of the Northern Hemisphere is different from that of the Southern Hemisphere, giving a so-called pear-shaped Earth; also, it has been confirmed that the geoid undulations are approximately plus or minus 100 metres, therefore there does not seem to be any other simple mathematical figure that can approximate the geoid significantly better than the ellipsoid of revolution, except a high-order spherical harmonic representation.

INTERNATIONAL REFERENCE SYSTEMS

Current definitions and values. The International Association of Geodesy adopted an International Ellipsoid at the General Assembly in Madrid in 1924. This ellipsoid was defined through adopting numerical values for the semi-major axis of an ellipsoid of revolution: $a = 6{,}378{,}388$ metres, and for the flattening of this ellipsoid: $f = 1{:}297$, which were the numerical values obtained by J.F. Hayford (1909). The formula for theoretical gravity (γ) in terms of the geodetic latitude (φ) on this ellipsoid was adopted by the Association at the General Assembly in Stockholm (1930) as:

$$\gamma = 978.0490\,(1 + 0.0052884\sin^2\varphi - 0.0000059\sin^2 2\varphi)\ \text{gal}.$$

Gal is a unit of acceleration, one centimetre per second per second. The equatorial value of gravity for the formula was computed by a Finnish geodesist, W.A. Heiskanen (1928), from isostatically reduced gravity data, and coefficients were computed by an Italian geodesist, G. Cassinis (1930), from the values corresponding to the ellipsoid.

As time passed, however, the above defined system proved to be less and less adequate for scientific purposes in the light of new observational data and their evaluation by various scientists as described above. The results of the new observations refined the parameters. During each of the General Assemblies of the International Union of Geodesy and Geophysics, new parameters were discussed, but it was found to be advisable to postpone the adoption of new values. Because of large discrepancies between the earlier adopted values and the then available information, however, the International Astronomical Union found it necessary to adopt a new reference figure in 1964. After having obtained the advice of the International Union of Geodesy and Geophysics regarding the choice of the parameters of the figure of the Earth and its gravity field, the International Astronomical Union adopted at its XIIth General Assembly in Hamburg its new system of constants. This system contains the following three basic parameters: equatorial radius of the reference ellipsoid, a; geocentric gravitational constant of the Earth including the atmosphere (gravitational constant times Earth's mass), GM; and the dynamic form factor of the Earth, defined as the coefficient of the second-degree zonal harmonic of the potential, J_2. The adopted numerical values are:

Three basic parameters adopted by the International Astronomical Union

$a = 6{,}378{,}160$ metres (20,920,365 feet)
$GM = 393{,}603 \times 10^9$ metres/second^2
$J_2 = 0.0010827$.

Besides the participation of geodesists in the elaboration of the above system of constants, there soon arose a general feeling that a revision of the international reference system of 1924–30 was necessary. It is true that a reference system serves primarily as a set of approximate parameters to which the observed data may be related, so that the adopted values are arbitrary within certain limits. On the other hand, it is preferable to adopt values that are close to reality, and one of the responsibilities of the International Association of Geodesy is to define the best parameters characterizing the figure of the Earth and its gravity field. Even though the values adopted by the International Union were not the best values available, it was finally necessary, to maintain uniformity and to avoid the proliferation of closely similar geometric and dynamic systems, for the International Association of Geodesy to submit for approval by the International Union of Geodesy and Geophysics the Geodetic Reference System 1967, which had values for a, GM, and J_2 identical to those defined by the International Astronomical Union. This was done during the General Assembly at Lucerne in 1967. Table 3 shows a set of constants derived for the Geodetic Reference System 1967. The gravity formula corresponding to the Geodetic Reference System 1967 is called the Gravity Formula 1967. Normal gravity γ corresponding to the Gravity Formula 1967 may be computed by means of the closed formula:

$$\gamma = \gamma_e \frac{1 + k\sin^2\varphi}{\sqrt{1 - e^2\sin^2\varphi}};$$

or if the extraction of square roots is to be avoided, the Chebyshev approximation:

$$\gamma = 978.03185\,(1 + 0.005\,278\,895\sin^2\varphi + 0.000\,023\,462\sin^4\varphi)\ \text{gal}$$

may be used, for which the maximum error is 0.004 milligal. The conventional formula

$$\gamma = 978.0318\,(1 + 0.005\,3024\sin^2\varphi - 0.000\,0059\sin^2 2\varphi)\ \text{gal}$$

is less suited for direct use because its accuracy is limited (0.1 milligal).

Table 3: Constants Derived for the Geodetic Reference System 1967

Geometric constants	
Semi-major axis	$a = 6{,}378{,}160$ metres (exact)
Semi-minor axis	$b = 6{,}356{,}774.5161$ metres
Polar radius of curvature	$c = 6{,}399{,}617.4290$ metres
e = first eccentricity, $e^2 = \frac{a^2 - b^2}{a^2}$	$e^2 = 0.00669460532856$
Flattening	$f = 1/298.247167427$
Length of meridian quadrant	$Q = 10{,}002{,}001.2313$ metres
Area of surface	$A = 510{,}069{,}262$ sq km
Volume	$V = 1{,}083{,}218{,}915 \times 10^3$ cu km
Mean radius $R_1 = (2a + b)/3$	$R_1 = 6{,}371{,}031.5054$ metres
Radius of a sphere of same surface area as the ellipsoid	$R_2 = 6{,}371{,}029.9148$ metres
Radius of a sphere of same volume as the ellipsoid	$R_3 = 6{,}371{,}023.5234$ metres
Physical constants	
Geocentric gravitational constant	$GM = 398{,}603 \times 10^9$ m^3 sec^{-2} (exact)
Angular velocity	$\omega = 7.2921151467 \times 10^{-5}$ radians per second (exact)
Normal potential at ellipsoid	$U_0 = 6{,}263{,}703.0523$ kilogal-m
Spherical harmonic coefficients	$J_2 = 0.0010827$ (exact)
	$J_4 = -0.00000237126440$
	$J_6 = 0.00000000608516$
	$J_8 = -0.00000000001428$
Normal gravity at Equator	$\gamma_e = 978.03184558$ gal
Normal gravity at pole	$\gamma_p = 983.21772792$ gal
$k = (b\gamma_p - a\gamma_e)/a\gamma_e$	$k = 0.00193166338321$

Recent developments. The results from 1970 measurements, which came mainly from observations of satellites and from combinations of all information, indicate that the values for the best fitting ellipsoid of revolution are:

$a = 6{,}378{,}140$ metres (20,920,299 feet), with standard error of 3 metres (10 feet)
$f = 1/298.258$.

The geodetic measurements have become so accurate that changes can be recorded as a function of time. In precise levelling, for example, results in certain geographical locations have been reduced to a certain epoch (*i.e.*, to a fixed time) in order to have consistent results because of continuous changes in elevation during the time period, such as noticeable uplift of the crust in Scandinavia and in the eastern parts of Canada. Time has become a fourth dimension in most geodetic measurements. Horizontal motions have been monitored, for example, in California, and there are theories about continental drifts. A new generation of instruments and measuring techniques such as very long base line interferometry (VLBI) and geoceivers, now being used, will be capable of detecting continental drift (*q.v.*) in a relatively short period of time. There have been speculations about changes in the universal gravitational constant G. A suggestion has been made that the Earth's polar moment of inertia is increasing and that the gravitational constant is decreasing at the rate of the order of one part in 10^8 per century. These estimates agree in order of magnitude with the rate of upward transport of material at the mid-ocean ridges (see SEA-FLOOR SPREADING), but these should be considered only as conjectures. To discover whether or not gravity is actually changing as a function of time, it has been suggested that the values of gravity be monitored at several locations around the world with the new type of drift-free gravimeters now under development.

Detailed shape of the local geoid can be obtained from local gravity data observed using usual existing gravimeters, airborne gravity instrumentation, or gradient measuring techniques. These undulations also may be detected through use of altimeter data from satellites or data obtained from satellite to satellite tracking. Only the future will show how stable the Earth's size and shape really are. The problems have become more complex and extensive. International cooperation as well as cooperation of scientists in different disciplines is needed.

BIBLIOGRAPHY. The historical background of the determination of the figure of the earth is covered in C.R. BEAZLEY, *The Dawn of Modern Geography*, 3 vol. (1897–1906); A.D. BUTTERFIELD, *A History of the Determination of the Figure of the Earth from Arc Measurements* (1906); J.B.L. DELAMBRE, *Grandeur et figure de la terre* (posthumous work ed. by G. BIGOURDAN, 1912), and *Histoire de l'astronomie ancienne*, 2 vol. (1817). General coverage of geodesy, including modern mathematical methods, as well as the use of satellites, is provided by W.A. HEISKANEN and F.A. VENING MEINESZ, *The Earth and Its Gravity Field* (1958); GUY BOMFORD, *Geodesy*, 2nd ed. (1962); G. VEIS (ed.), *The Use of Artificial Satellites for Geodesy*, 2 vol. (1963–67), proceedings of the first and second international symposia on the use of artificial satellites for geodesy; IVAN I. MUELLER, *Introduction to Satellite Geodesy* (1964); WILLIAM M. KAULA, *Theory of Satellite Geodesy* (1966); W.A. HEISKANEN and HELMUT MORITZ, *Physical Geodesy* (1967); and MARTIN HOTINE, *Mathematical Geodesy* (1969). These works include further references to the specialized literature. Also notable is the *Geodetic Reference System 1967*, a special publication of the *Bulletin geodesique* of the International Association of Geodesy (1970).

(U.A.U.)

Earth, Geological History of

To survey earth history requires an understanding of the vastness of geological time. The year is hardly an adequate unit for measuring geological time. A span of 4,600,000,000 years, the currently accepted age of the earth, is almost beyond comprehension, and earth history is therefore divided into eras, the eras into periods, and the periods into epochs.

The first era, the Precambrian, comprises almost 90 percent of geological time, from the time of origin of the earth about 4,600,000,000 years ago to the beginning of the Paleozoic Era, about 570,000,000 years ago, when evidence of living things first becomes plentiful in the rock record. The geological history of Precambrian time is very incomplete. The succeeding 570,000,000 years, down to the present, are much better known than the earlier times and can be divided into three eras, namely the Paleozoic, Mesozoic, and Cenozoic, which are collectively referred to as the Phanerozoic. During the Paleozoic Era, which lasted about 345,000,000 years, the most abundant animals were marine invertebrates. During the Mesozoic Era, which lasted about 160,000,000 years, dinosaurs and other reptiles were abundant. The Cenozoic Era, comprising the last 65,000,000 years of earth history, is the time when mammals and broadleaf trees became important. The basis for reconstructing Phanerozoic history has been the succession of rock formations and their contained fossils.

The eras of geological time

This article is intended as a broad overview of earth history, from the origin of the planet to the present. For more detail on any specific aspect of earth history, the interested reader should see the several related articles. Specifically, see EARTH, STRUCTURE AND COMPOSITION OF; ATMOSPHERE, DEVELOPMENT OF; OCEANS, DEVELOPMENT OF; and PRECAMBRIAN TIME for the earliest aspects of earth history, including the probable physical, chemical, and biological conditions that prevailed. See ROCK DEFORMATION; MOUNTAIN-BUILDING PROCESSES; SEA-FLOOR SPREADING; and CONTINENTAL DRIFT for discussions of earth movements of several types, including those of relevance to continental formation and distribution. See SEDIMENTARY ROCKS; IGNEOUS ROCKS; and METAMORPHIC ROCKS for treatment of the nature and formation of these basic constituents of the earth's crust; see FOSSIL RECORD; STRATIGRAPHIC BOUNDARIES; GEOLOGICAL TIME SCALE; and DATING, RELATIVE AND ABSOLUTE for information on the means of interpreting the record of the rocks and the record of life through time and the means of subdividing geological time. See also articles on the eras, periods, and epochs of geological time—*e.g.*, PALEOZOIC ERA, LOWER; JURASSIC PERIOD; PLEISTOCENE EPOCH.

FACTORS INVOLVED IN EARTH HISTORY

Earth movements. Rocks that contain fossils do not occupy the position they had when the rocks first formed. Formations containing sea shells may be found on the summits of some of the highest mountains, yet the sea never was that high. The rocks containing the shells were deposited as sediments in the sea, and the sediments then became compacted and consolidated to form rock. Subsequently, earth movements uplifted the rocks to their present locations, sometimes thousands of feet or even miles above sea level.

The earth movements that accomplish such changes are slow. A foot of uplift in a thousand years is not exciting; many parts of the world are more active than that. Yet that movement, if continued for only 10,000,000 years, can raise a mountain range 10,000 feet above sea level.

Displacement and uplift by faulting

Geologic processes that build mountain ranges and later destroy them by erosion can be observed today. The earth movements that raised mountains generally progressed as repeated small movements of the same kind as those accompanying modern earthquakes. In 1959, at Hebgen Lake, west of Yellowstone Park in Wyoming, in a matter of moments the bed of the lake was tilted 20 feet (6 metres) northeast. Houses and highways on the northeast shore were submerged, whereas boats and docks on the opposite southwest shore were stranded 20 feet above the new shore. Half of this displacement was taken up by a fault (fracture of the Earth's crust along which movement occurs) with a 10-foot displacement. Highways crossing the fault were displaced, and a barn over the fault was broken in two parts, one 10 feet higher than the other. The geological record indicates that the world's great mountains—the Himalayas, Alps, Andes, and Rockies—were broken and their parts uplifted by small displacements like those at Hebgen Lake, repeated over long periods of time.

Former environments. An important aspect of earth history is, of course, the environment that prevailed during the deposition or erosion of rocks in former times. Reconstructing the kinds, magnitude, and intensity of former geological processes from ancient deposits is the subject of paleogeography (*q.v.*). It is generally accomplished by recourse to the dictum that the present is the key to the past, meaning that past processes and their effects were similar to those processes and effects that are observable today. Modern coasts provide an example of the environmental factors involved.

Coastal facies and environments

A succession of different kinds of environments and sedimentary deposits occurs along most coasts. Inland are stream valleys with floodplain deposits of stream-transported gravel, sand, and silt. These sediments, referred to as alluvial or fluvial, contain the remains of land animals and plants or freshwater shellfish and finfish. Approaching the coast, this freshwater environment gives way to an estuarine environment where the water may be brackish. No longer is there much gravel, for streams entering the estuary are too sluggish to transport coarse sediment. The channels are of sand, silt, and mud. In marshes and swamps, the wet ground prevents oxidation and decomposition of the plant matter, which then collects as peat. Here the species of shellfish and finfish are those conditioned to brackish water and to tidal fluctuations. Plants around the shores of the estuary are those that can tolerate 1 or 2 percent salts in the water.

Deposits that represent such different environments are referred to as facies; the stream deposits constitute one facies, and this grades laterally into the estuarine facies. The estuarine facies, in turn, grades to the shore or littoral facies, which may include shore deposits, beach dunes, and offshore sediments (marine muds). Each facies has a distinctive composition and texture (lithology), and each has a distinctive fauna and flora.

Record of the rocks. Ancient rock formations record successions of facies along the shores of ancient seas that

once extended onto various parts of the continent. The gravel, sand, and silt of the fluviatile environment becomes rock, known as conglomerate, sandstone, and siltstone, respectively. The mud, sandy mud, and peat in the

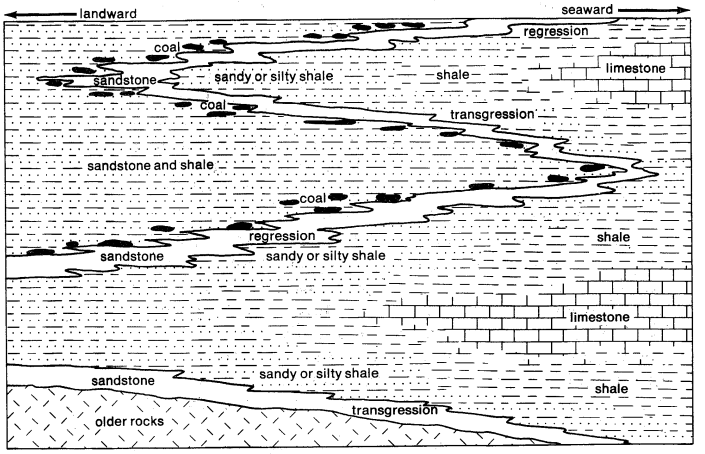


Figure 1: Environmental changes (facies) across a shifting, ancient shoreline. Floodplain deposits of sandstone and shale (left) containing fossil land animals and plants identify the old coastal plain. These grade seaward to estuarine deposits containing coal beds, and these grade to sandstone representing the ancient beaches. The diagram shows two transgressions and two regressions.

estuarine environment is transformed to shale, sandy shale, and coal. Sandstone representing former beach deposits grades into shale and sandy shale containing marine fossils; the latter represent former offshore marine muds. Where the water is warm, as in tropical latitudes, calcium carbonate may be precipitated, and this forms limestone, another indication of former environments.

Where earth movements cause coasts to subside and enable the sea to advance onto what had been land, the different environments become shifted landward, and the facies characteristic of seaward environments overlap earlier deposits characteristic of landward environments. Where coasts are stable and the influx of sediments from the land is considerable, the facies characteristic of landward environments extend seaward and overlap earlier deposits representing the seaward environments (Figure 1).

Transgression, regression, and stratigraphic succession

An advance of sea over the land is referred to as marine transgression; a retreat of the sea caused by uplift of the land or by accelerated influx of sediments is referred to as regression. Figure 1 illustrates two transgressions and two regressions. The continuity of the sedimentation may be interrupted by uplift accompanied by erosion. Such breaks in the geologic record are referred to as unconformities (Figure 2).

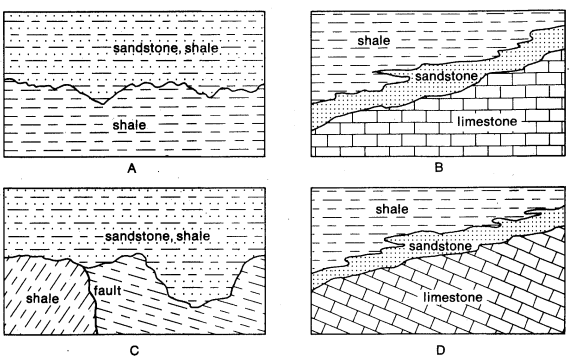


Figure 2: *Examples of unconformities.*
(A) Change from shale to sandstone and shale, slight erosion on top of shale. (B) Limestone surface eroded, buried by sandstone and shale. (C) Shale folded and faulted buried by sandstone and shale. (D) Limestone tilted and eroded before marine sandstone and shale deposited.

Younger sedimentary deposits must initially overlie older ones in accord with the principle of superposition. This provides the basis for studying the sequence of sedimentary rocks. The physical changes recorded by a succession of strata are accompanied by biological changes recorded by the fossils in those sediments. These biological changes are of two kinds; one relates to changes in environments (facies), and the other records the evolution of life forms through time.

Whereas stratigraphic succession is the basis for chronology, the subdivision of the sedimentary formations into units of geological time is defined by the fossil record for each of the eras and periods. The succession of paleontologic changes is the basis for correlating the rock formations and the episodes of geological history they record.

Igneous and metamorphic rocks

In addition to sedimentary rocks, the earth's crust consists of igneous rocks, which once were molten (*e.g.*, lava), and metamorphic rocks, which are produced by the alteration of pre-existing rocks by great heat or pressure.

Temperatures and pressures in the earth increase greatly with depth. The melting point of rocks such as granite, at atmospheric pressure, is around 1,200° C (2,200° F), and such temperatures are reached a few miles down. The rocks are not generally melted because of the great confining presssures, but earth movements can relieve the pressure locally and cause the rock to melt. When this happens, the molten mass becomes squeezed upward as an intrusion into the overlying rocks. The intrusions cut discordantly through the overlying rocks to form what are called stocks, or if an intrusion continues to the surface it may form a volcano and erupt. Stocks clustered closely together (such as the extensive granites of the Sierra Nevada and Coast ranges of Canada and Alaska) are called batholiths.

Igneous and metamorphic rock facies are somewhat like those in sedimentary rocks. An igneous mass, for example, may change greatly between two locations because of different reactions with the host (intruded) rocks, different rates of cooling that control the kinds and sizes of minerals that crystallize from the melt, differences in fluidity or rates of injection that affect the shapes of the intrusive masses, and differences in fluidity that may allow early-formed heavy minerals to settle and leave behind a melt with a composition different from that of the parent mass.

Metamorphic rocks have different facies depending on the degree of alteration and commonly exhibit changes with distance from a site of intensive deformation.

Geological dating. Correlations based on superposition and paleontology have been greatly extended by dating rocks by their radioactivity. The principle of radiometric dating is disarmingly simple.

A radioactive element decays to a daughter element (isotope) at a constant and known rate that is referred to as the half-life, the time required for one-half of the originally present (parent) element to decay to the daughter element. Knowing the decay rate, measurement of the ratios of the parent and daughter elements provides the age. The chief difficulty in applying the methods to rocks and minerals or to surface deposits and soils is the necessary assumption that there has been no depletion or enrichment of either isotope.

Radiometric-dating methods

Three methods widely used for determining radiometric ages involve uranium (and thorium)–lead, potassium–argon, and radiocarbon.

The method utilizing the ratio of uranium to lead and thorium to lead involves three pairs of isotopes occurring together and having different decay rates, as follows:
uranium-238–lead-206,
half-life 4,500,000,000 years;
uranium-235–lead-207,
half-life 700,000,000 years; and
thorium-232–lead-208,
half-life 13,500,000,000 years ± 500,000,000 years.
This method has the advantage of utilizing three different decay ratios, and when the three rates indicate comparable ages the probabilities are enormous that a

satisfactory age has been determined. The method is largely restricted to igneous and metamorphic rocks and is particularly useful in dating Precambrian events.

The potassium–argon method involves the transformation of potassium-40 to argon-40 and to calcium-40 with a half-life of 1,310,000,000 years. A great advantage of this method is the abundance and wide extent of potassium-bearing minerals in the earth's crust, including some sedimentary rocks. This enables absolute dating of paleontologic and stratigraphic successions.

The radiocarbon method involves the transformation of carbon-14 to nitrogen-14 with a half-life of about 5,500 years. With this half-life the method is restricted to dating deposits younger than about 50,000 years, those of the Holocene Epoch and latest Pleistocene time.

Other methods have been used for estimating ages of certain geologic events, such as retreat of waterfalls, salinity of the oceans, rates of weathering, rates of deposition of sediments, and rates of soil formation. These are not dependable, however, because the assumed rates probably have changed over the period of time being estimated.

The earth's interior. Geophysical data, including the velocity of earthquake waves and density or gravity measurements, indicate that the earth consists of concentric shells that decrease in density outward from its centre. At the centre is a core with a radius of about 3,400 kilometres (2,000 miles). The inner part is judged to be solid and the outer part has the properties of a highly viscous liquid. This core has a density of about 12 and is thought to be a mixture of iron and nickel. It is 15 percent by volume and 32 percent by weight of the earth.

Mantle, crust, and Mohorovičić Discontinuity

A sharp structural discontinuity separates the core from the next outer shell, the mantle, and the density drops abruptly to less than 5.5, which is less than the density of iron (7.9). The density of the mantle decreases outward to about 4.5 at a depth of about 900 kilometres (550 miles) and to about 3.5 at the Mohorovičić Discontinuity, which marks the base of the outermost shell, the crust. The mantle comprises 84 percent of the earth's volume and 67 percent of its weight. The crust has an average density of less than 3.0 and is about 5 kilometres (3 miles) thick under the oceans; it thickens to 30 to 50 kilometres (20–35 miles) under the continents. The crust has a lower layer termed sima (a contraction for silica–magnesium) with a density about that of basalt (2.8–2.9); the upper layer termed sial (for silica–aluminum) has a density about that of granitic rock (2.65–2.75). The crustal layer under the oceans is mostly basaltic; under the continents it consists of both basaltic and granitic layers. The crust makes up less than 1 percent of the earth by volume and about 0.4 percent by weight.

Structure of the continents. The continents are rafts of light rock floating on the denser mantle. Each continent consists of a core of highly metamorphosed sialic rocks, which outcrop as nuclei of the continents, referred to as continental shields. Cores and shields are composed of the oldest pre-Paleozoic rocks forming the continents. Around each shield is a stable platform of such pre-Paleozoic rocks overlapped by thin Paleozoic or younger sedimentary rocks (Figure 3) largely of marine origin and showing evidence of former submergence. Along the outer edges of these platforms are unstable mobile belts where the crust has been subject to broad downwarping to produce structural troughs known as geosynclines. These troughs were invaded by the sea and great thicknesses of marine sediments were deposited in them. The sediments commonly aggregate 35,000 to 45,000 feet (10,600–13,600 metres) thick, whereas those deposited on the stable platforms surrounding the shields are mostly less than 5,000 feet (1,500 metres) thick.

Role of geosynclines in earth history

In general, these marginal geosynclinal troughs consist of two belts. The outer belt seemingly was very mobile and was bordered by arcs of volcanoes; in it were deposited coarse clastic sediments (sediments made up from fragments of pre-existing rocks), lavas, and sediments derived from volcanic ash and lavas. This belt is referred to as the eugeosynclinal belt. Parallel to it and adjoining the stable platform is another geosyncline containing little volcanic material and much less coarse sedimentary deposits; thick deposits of limestone and other carbonate rocks are included in this miogeosynclinal belt, however.

The geosynclines sank as the sediments were deposited in them, as if the weight of the sediments contributed to pushing these mobile parts of the crust downward. Although the sediments in the geosynclines may be as thick as 45,000 feet (13,600 metres or 9 miles), the troughs never were that deep. The facies represented by the sediments mostly indicate moderately shallow water. Evidently the sinking progressed as the sediments accumulated.

During the history of a particular geosyncline, the sides of the trough at times would become compressed, and the rock strata in the trough would be folded. Where folding progressed beyond the limit of strain of the rocks, they broke by faulting. Most geosynclines record several episodes of such folding and faulting, especially in the eugeosynclinal belt. These episodes commonly but not invariably are accompanied by volcanism and intrusions of various kinds of igneous rocks.

After eight to ten miles of sediment had been deposited in most geosynclines, the sinking ended, and the sediments became strongly folded, faulted, and uplifted to form mountains. What had gone down now came up. These kinds of structural movements terminating geosynclines or interrupting their development have been termed orogenies; they are mountain-building episodes in the Earth's history.

Orogenies

BEGINNINGS OF EARTH HISTORY

The primitive earth. Most theories about the origin of the earth assume it began in a gaseous state, evolved through a liquid state, and finally became partly solid. Some theories assume that the solar system and the earth began as a dust cloud with the composition of interstellar gas. Such a cloud would have contained only 1 percent of most of the terrestrial elements, with a remainder of hydrogen and helium in equal amounts. As the dust-cloud particles aggregated, perhaps electrostatically, gravity would have caused further aggregation.

Some theories assume that the earth began in a hot state and has been steadily cooling; other theories assume that the earth began cold and grew warmer as a result of compression and subsequent radioactive decay of its interior constituent elements. It is generally supposed that the dust particles of the heavy terrestrial elements would gravitate toward the centre of the mass and that the light elements, especially hydrogen and helium, could escape again to outer space. As aggregation progressed the heavier terrestrial elements gravitated toward the centre and the lighter ones toward the top. The original gaseous mass differentiated into liquid and gaseous phases, and these in turn differentiated into solid and liquid phases. Some calculations suggest that these changes could have

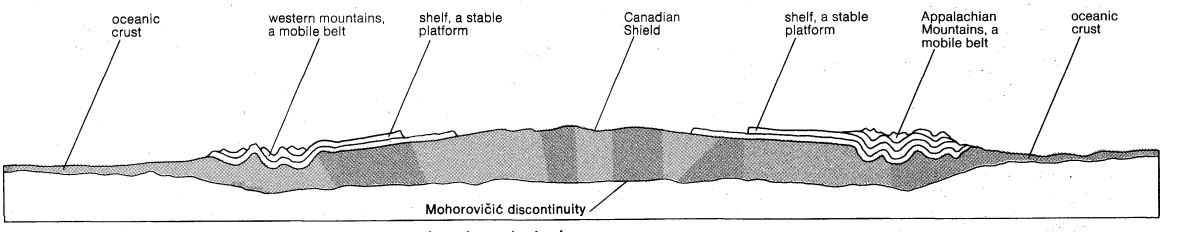


Figure 3: Cross section of North America illustrating the principal structural elements of a continent.

occurred quickly, in as little as a few thousand years. In any event, the formation of a solid crust can be taken as the beginning or zero time in geological history. Judging by the age of the oldest meteorites, of moon rocks, of terrestrial lead, and by calculations about the rate of retreat of galaxies, this occurred about 4,600,000,000 years ago. The earlier history is referred to as the astronomical period of earth development.

Differentiation of core, mantle, and crust

During the astronomical period of earth history the processes of differentiation into core, mantle, and crust began. The iron–nickel core is thought to have begun forming by the gravitating of liquid iron with nickel toward the centre of the mass. Outside of this, a shell of metallic silicates and sulfides collected, later to become the mantle. The lightest constituents, the silicate phase, rose to the top and formed the crust. Some of the hydrogen in the proto-atmosphere dissolved in the molten iron and was retained in the earth, but much of the hydrogen and most of the inert gases were lost to outer space. When the crustal layer first solidified, there is thought to have been little or no atmosphere. The atmosphere developed later by the escape of various gases from the earth's interior.

Primitive atmosphere, organisms, and soils. Perhaps the first crust on earth was like that on the moon and had a composition similar to stony meteorites. As this crust cooled and solidified it further differentiated into a comparatively dense sima layer and an overlying, lighter sialic layer. Such differentiation may have contributed water vapour to the atmosphere, but the early atmosphere must have been cloudy; the water could not fall as rain until the temperatures had dropped below 100° C (212° F). This stage can be referred to as the proto-atmosphere.

When the rains came, weathering and erosion began. Water running off the high ground formed the first streams, and where it collected in basins there formed the first lakes and seas. The abundance of chlorine, bromine, iodine, boron, and sulfur in the oceans and the comparative scarcity of these elements in the sialic crust has been interpreted as a relic of the proto-atmosphere. There still was no living matter, and weathering was entirely by physical-chemical processes.

There must have been some free oxygen in the primitive atmosphere, but most of the oxygen in today's atmosphere is thought to be the product of photosynthesis by plants that have lived since life on earth began. When life arrived on the land, the biochemical processes of soil development could begin, but the soils must have been primitive compared to the well-layered ones that develop under higher types of vegetation. This was the period of protosoils.

Assuming that the solid crust formed 4,600,000,000 years ago, the proto-atmosphere may have lasted for 250,000,000 years and protosoils may have lasted for a similar time span. These periods of the primitive earth must have lasted less than 1,000,000,000 years because granitic rocks as old as 3,500,000,000 years have been dated from the African and European shield areas.

PRECAMBRIAN HISTORY

Continental shields. Each of the continents has a core of Precambrian rocks extensively exposed over a large part of the continent and reduced by erosion to low relief. These are the continental shields, in which the rocks are complexly folded, faulted, and metamorphosed. The rocks composing them are mostly granite and gneiss; infolded and infaulted with them are metamorphosed volcanic and sedimentary rocks in belts that represent remnants of Precambrian geosynclines. The earliest geosynclines probably formed when the earth was less than 1,000,000,000 years old. There was sufficient water to cause rains and to form seas, and the runoff on the land was sufficient to cause erosion.

The Canadian Shield

The core of North America outcrops in the Canadian Shield, which extends over nearly 3,000,000 square miles. Topographically and structurally the Canadian Shield is a vast saucer that rises outward from an area beneath Hudson Bay and Foxe Basin. The shield was submerged by shallow water at times during the Paleozoic, but the flat-lying beds show no subsequent folding.

The Precambrian rocks are mostly below 1,000 to 1,500 feet (300–450 metres) in altitude but exceed 2,000 feet (600 metres) in some places. Their surface slopes away from the shield and forms a platform to the south and west under a cover of thin and nearly flat-lying Paleozoic and Mesozoic formations. To the southeast the shield ends abruptly at a deep trough, a geosyncline, in the Precambrian formations; this trough contains thick Paleozoic formations that are intensely folded and faulted.

The outstanding feature of the shield is its tremendously long, complex, and only partly understood geological history. Some of the rocks are dated at more than 2,000,000,000 years old, which is nearly four times longer than all of Phanerozoic time. The shield was eroded to low relief before the Paleozoic and was inundated by shallow seas intermittently during the Paleozoic. The shield probably was emergent during the Mesozoic and Cenozoic. It was covered by a continental ice sheet during the Pleistocene Epoch.

The rocks in the Canadian Shield are practically without fossils, and the stratigraphic succession must be based largely on radiometric dating of intrusive igneous rocks. These intrusions accompanied intense folding and faulting, and the resulting metamorphism caused sufficient recrystallization in the folded rocks to cause them to yield the same radiometric age as the intrusions that cut them. Age relations in the Canadian shield therefore are grouped on the basis of radiometric ages of granitic intrusions.

On this basis three major periods of intrusion and associated folding and other deformation have been distinguished; their names and ages are Grenville 1,000,000,000 years; Hudsonian 1,800,000,000 years; Kenoran 2,500,000,000 years. Hudsonian foldbelts occur in the Churchill, Bear, and Southern Superior Provinces; Kenoran foldbelts in the Superior and Slave Provinces (Figure 4). There must have been other periods of folding and igneous activity in addition to those named because it is unlikely that the intervals, which are as long as the entire Phanerozoic, were without folding and faulting. Present information permits dividing the Canadian Shield into the structural provinces shown in Figure 4.

By courtesy of the United States Geological Survey

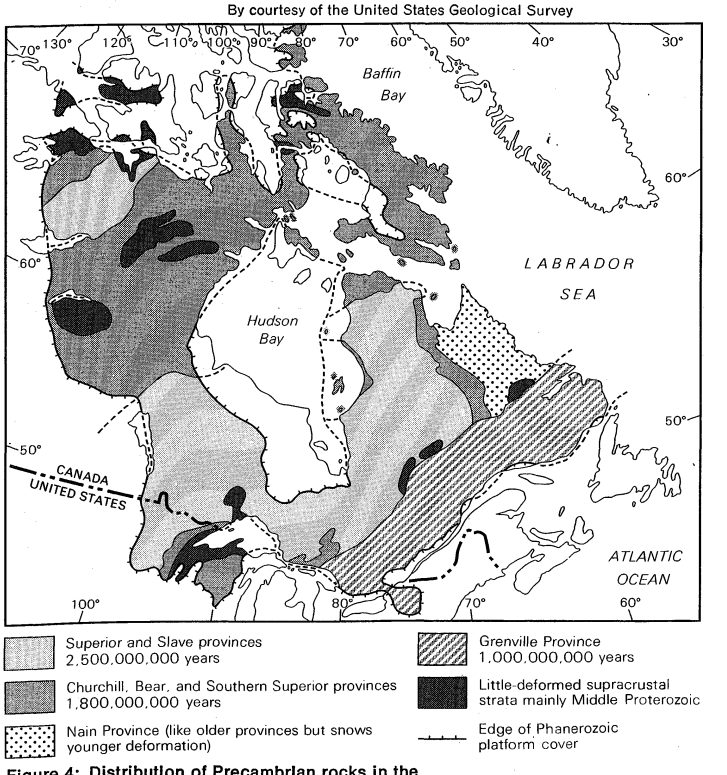


Figure 4: Distribution of Precambrian rocks in the Canadian Shield.

The rocks in the Superior and Slave provinces record major development of granite about 2,500,000,000 years ago. Antedating the granite, and largely engulfed by it are thick, tightly folded beds of basaltic lava, volcanic ash, banded iron formations, and graywackes (a poorly sorted sandstone that is characterized by rock fragments in an abundant, clay-rich matrix). The oldest graywacke contains fragments of granite and clearly indicates existence of a granitic terrain older than the one dated at 2,500,000,000 years ago. These dates suggest that the continental crust in North America 3,500,000,000 years ago was transitional between the primitive basaltic crust and the thick, extensive granitic crust that exists today. Before the end of Precambrian time the continents were well formed.

Precambrian rocks also are exposed in the eastern United States in the Piedmont Province and Blue Ridge and in the central and western United States in uplifts at the Ozark Plateau, central Texas, Rocky Mountains, southern Basin and Range Province, and some California coast ranges. Some of these rocks have been assigned radiometric ages as great as 2,500,000,000.

The Baltic Shield

Rocks of the Baltic Shield record at least four episodes of deposition of coarse clastic sediments, lavas, and tuffs, alternating with episodes of folding, faulting, and intrusion of granite. The early sedimentary rocks are like those in the ancient basins of Canada; basaltic lavas occur together with sandstone, shale, and graywacke from granitic and basaltic terrain and banded iron formations. The younger rocks include dolomite. Rocks as old as 1,800,000,000 to 2,000,000,000 years contain graphitic schist and traces of bitumen; this carbon suggests organic matter and living organisms.

The Precambrian rocks extend southward and southeastward under the plains of northern Europe and form a platform, capped by thin Paleozoic formations, very similar to the platform under the central United States. The eastern part of the Baltic Platform ends at the Urals; southward it is overlapped by Mesozoic formations.

Angaran, African, and Indian shields

A third continental core in the Northern Hemisphere is exposed in the Angaran Shield in central Siberia between the Yenisey and Lena rivers. Precambrian rocks forming this shield extend westward under the platform of the west Siberian Lowland and end at a mobile belt marked by the Ural Mountains. This shield is a high plateau, higher on the average than the Canadian or Baltic shields.

The Southern Hemisphere has shields in Africa, India, Australia, South America, and Antarctica, and all are plateaus. The African Shield, by far the largest, includes the entire continent except for narrow belts of folded rocks at the northwest and southern margins. Over much of Africa the shield rocks are overlain by thin, flat-lying Phanerozoic sedimentary formations, like those on the platform in the central United States. Africa, like the Canadian Shield and its platform, has remained comparatively stable since the beginning of the Paleozoic. This shield not only is largest but is the highest today; extensive areas are above 1,000 metres (3,280 feet) altitude. Perhaps the core of this continent is the thickest. In the Transvaal, granite dated at 3,200,000,000 years intrudes rocks that obviously are older and undated.

Of special interest among the rocks of the African Shield are those of the Witwatersrand formations, which consist of more than 24,000 feet (7,400 metres) of conglomerate, feldspathic sandstone, quartzite, slate, and lavas. These formations contain the world's largest known deposits of gold.

The Indian Shield, comprising nearly all of peninsular India, consists of a complex of gneisses, metamorphosed volcanic rocks, and granite. At least three Precambrian orogenies occurred, but this shield, like the others, has been stable during the Phanerozoic. The northern margin of the shield ends at Paleozoic and younger geosynclines marked by the Himalayas and the plain of the Ganges. The western part is capped by one of the world's great basalt fields, the Deccan Traps.

Precambrian glaciations. Among the conglomeratic formations of the shields are some tillites, heterogeneous rocks derived from glacial till. The boulders, faceted and striated, are in a fine-grained matrix. Associated with some of the deposits is banded slate that may have originated from varved (banded by annual layers) lake deposits. Tillite has been found in all of the shields, but only a few deposits are well dated. At least three stages of glaciation are indicated by deposits as shown in Table 1.

Table 1: Location and Age of Precambrian Glaciations

location	approximate age (years)
Uppermost Precambrian (Beltian Series), Wasatch Mountains, Utah	750,000,000
Uppermost Precambrian (sparagmite sandstone) on Baltic Shield in northern Norway	750,000,000
African Shield in Katanga, Angola, lower Congo, and South Africa	750,000,000
Australian Shield has banded slate thought to be seasonal (varved) banding	750,000,000
Canadian Shield (cobalt formation in the Huronian [Animikie] Series)	2,000,000,000
African Shield (Transvaal System)	2,000,000,000
Australian Shield (Broken Hill area)	2,000,000,000
African Shield (Witwatersrand Series)	2,600,000,000

Precambrian life. The first organisms probably appeared at the time of the first seas. Oxygen, still deficient, increased as photosynthesis developed. The first organisms releasing oxygen would be oxidized (burned up) by it unless the oxygen could combine with mineral matter. Supposedly this was ferrous iron, and its oxidation could have produced the banded iron formations that are widespread in Lower and Middle Precambrian rocks. These unique rocks are not found in younger formations. They consist of alternating layers of iron-rich and iron-poor silica and represent rhythmically banded, chemically precipitated sediment.

Banded iron formations and bluegreen algae

The last episode of banded iron formation was 1,800,000,000 to 2,000,000,000 years ago, and signs of bluegreen algae are associated with it. Thereafter, with much or most of the iron oxidized, oxygen was added to the atmosphere. Younger iron formations consist of red beds, the sediments of which are individual grains coated with iron oxide. The oldest thick and extensive red beds in the Precambrian are about 2,000,000,000 years old.

With the first appearance of red beds, calcium and magnesium carbonate (dolomite) occurred in abundance for the first time in marine sediments. The seas may then have developed a salinity comparable to that of today.

The younger Precambrian sedimentary deposits, notably the calcareous ones, contain concentric limy structures thought to have been formed by colonies of algae. Sponge spicules and the impressions of a jellyfish have been found, and some beds have burrows looking very much like worm burrows. The animals of those times did not have shell or skeletal material that would survive burial, and consequently there are almost no fossils.

This virtual absence of fossils, though, should not be construed to mean the absence of living organisms. The abrupt appearance of abundant fossils of animals with hard parts at the beginning of the Paleozoic requires considerable and diverse ancestry. An indirect record of the ancestral floras and faunas is provided by the carbon content of black shale and slate and by bedded deposits of graphite in some of the formations.

Eocambrian formations and the first trilobites and shelled invertebrates

The beginning of the Paleozoic has been defined by the first appearance of trilobites and associated fauna of shellfish, including linguloid brachiopods and ostracods. In many parts of the world, however, a thick series of essentially nonfossiliferous sedimentary formations conformably underlie the earliest Paleozoic beds that contain the fossils. These formations are physically part of the Paleozoic, but they are Precambrian on paleontological grounds because of the absence or primitive nature of contained fossils.

In the southern part of the Appalachian Geosyncline these Late Precambrian or Eocambrian formations consist of 7,500 feet (2,300 metres) of shale, sandy shale,

feldspathic sandstone, and conglomerate. Northward and westward the formations thin and are only about 2,500 feet (760 metres) thick in Maryland and southern Pennsylvania; they are absent in New York.

Eocambrian formations also are represented in the Paleozoic geosyncline west of Grand Canyon. In the Death Valley region they comprise 8,500 feet (2,580 metres) of nonfossiliferous dolomite, claystone, and quartzite. Eastward, the formations thin by overlap and are not represented under the Paleozoic formations on the continental platform at Grand Canyon.

Mineral resources. Some of the world's most important mineral deposits are in Precambrian rocks. Among them are the world's largest iron deposits (Canadian, Baltic, Amazonian, and Guianan shields, and in the Ukraine, U.S.S.R.); the world's largest gold deposits on the African Shield and other important gold deposits on the Canadian Shield and in Western Australia; and the world's largest nickel and cobalt deposits on the Canadian Shield.

Other major deposits include: copper on the Canadian Shield and in England; tin in England; uranium on the Canadian and African shields; platinum in Canada and the U.S.S.R.; chromium in South Africa and the U.S.S.R.; manganese in India, West Africa, and Brazil; lead and zinc in New Jersey, Idaho, British Columbia, and New South Wales; graphite in Korea and Siberia; mica in the U.S.S.R. and India; and talc in northeastern and southwestern United States.

Absence of fossil fuels

Equally notable are the mineral deposits that are lacking, especially the fossil fuels—coal, oil, and gas. There may have been such deposits in Late Precambrian time, but the intense metamorphism that has altered those old rocks has destroyed any fossil fuels that may have existed, except traces represented by deposits of graphite and the carbonates.

PALEOZOIC HISTORY

Stability of shields and mobility of geosynclines. During the Precambrian the shields included mobile geosynclines containing seaways that were bordered by mountainous uplifts. The mobile belts were subject to folding, faulting, volcanic activity, and intrusion of granite. By the beginning of Paleozoic time, the shields had become stable parts of the granitic crust forming the continents, and their complicated rock structures were eroded to low relief. The stable areas included the shields and the structural platforms bordering them. The platforms, and to a lesser extent the shields, were overlapped by shallow seas during the Paleozoic; both areas were subject to warping, but there was no more strong folding, faulting, or igneous activity.

Along one or more sides of each of the shields, unstable mobile belts formed deep troughs during the Paleozoic. In each such geosyncline, sediment accumulated to thicknesses of six to ten miles as the troughs sank, and these sediments thin against and overlap the platforms bordering the shields (Figure 5). Curiously, most of the sediment in the geosynclines was derived not from the shields but from mountains in the unstable mobile crust outside the geosynclinal belt; the sediments were mostly from mountains that no longer exist and were transported toward the shields.

This poses a double problem. Not only have the mountains been lost but if they were approximately symmetrical, there has also been lost that half of the sediments that must have been eroded from the far side. The problem can be relieved, perhaps, by assuming that the lost mountains were asymmetrical and represented a turned up edge of the continent. It is also possible that island arcs (*q.v.*) were present, having formed along the zone where sea-floor spreading caused underthrusting of the continental margins.

Early Paleozoic geosynclines and shelves. Just before the beginning of Cambrian time, a geosynclinal trough had formed along the west, south, and southeast sides of the Canadian Shield and along the northwest side of the Baltic Shield. The geosyncline in Europe is named the Caledonian. With the beginning of Cambrian time the geosynclines deepened. Similarity of fossils in the Caledonian and northeast Appalachian geosynclines suggests a connection across what now is the Atlantic Ocean.

The Caledonian and Tethys geosynclines

In North America and northwestern Europe, the Late Precambrian and earliest Cambrian deposits are mostly clastics—sandstone and shale—whereas the later Cambrian deposits are mostly limestone.

In the Anti-Atlas Mountains of northwest Africa the base of the Cambrian is within a thick limestone sequence. The faunas in these formations are related to those in the Cambrian of Asia Minor and eastern Asia. This marks the beginning of a long-lasting east–west trending geosyncline known as Tethys; the Mediterranean Sea is a descendant.

Ordovician formations in the Caledonian Geosyncline in Wales are of the eugeosynclinal types, but calcareous deposits of shallow seas overlapping the platform extend toward the Baltic Shield. Around the core of North America, Ordovician formations include shallow-water carbonate formations on the platform, miogeosynclinal formations bordering the platform, and eugeosynclinal facies away from it. The miogeosynclinal formations are largely marine carbonate and shale; the eugeosynclinal facies are mostly clastic, consisting of sandstone and black shale (Normanskill and Deepkill shales) with lavas, volcanic ash, and other volcanic derived sediments.

During the Silurian Period subsidence continued. In North America the seas on the platform divided into basins where evaporation increased the salinity and led to the precipitation of salts. In the British Isles the Early Paleozoic geosyncline filled with sediment, with shale accumulating in the inner basin and calcareous and sandy deposits along the margin. The volcanism that had occurred during the Ordovician almost ceased.

The Caledonian Geosyncline ended with an episode of strong folding in Devonian time. Tethys and the North American geosynclines also were subjected to folding during the Devonian, but those geosynclines persisted.

Early Paleozoic folding and igneous activity. Both in North America and along the Caledonian Geosyncline, rather steady subsidence during the Cambrian caused the sea to transgress on the platforms. There is little evidence of folding or of igneous activity in those mobile belts during the Cambrian. In the middle European (Czechoslovakia) and Mediterranean (Morocco) geosynclines, however, during the early stages of Tethys, there was volcanism and deposition of graywackes, and there was folding in Central Asia and Australia.

Throughout the world during the Ordovician and Silurian, there were repeated episodes of folding and volcanism. This deformation and igneous activity attained a

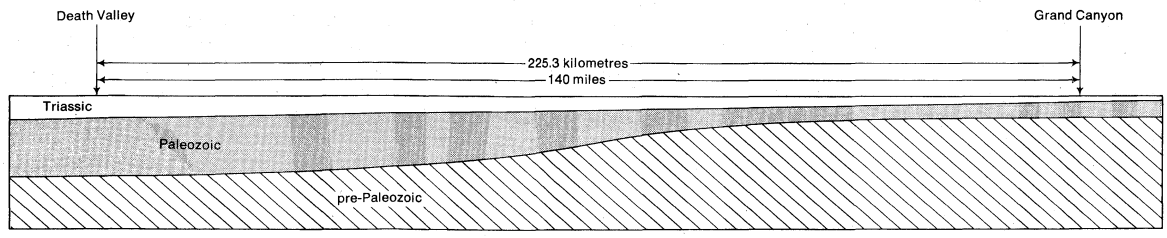


Figure 5: Thinning of sediments upon a stable shield area, illustrated by a cross section from Death Valley to the Grand Canyon. The vertical and horizontal scales are the same.

climax during the Devonian. Much of the granite in eastern North America formed at this time.

Late Paleozoic geosynclines and shelves. In North America, after the Devonian, the eugeosynclinal part of the Appalachian Geosyncline ended. The north part of the geosyncline divided into three deep troughs, each of which received distinctive sediments during the Mississippian and Pennsylvanian. In the central Appalachians the Mississippian is represented chiefly by clastic sediments derived from erosion of Devonian formations that had been uplifted to the east. Farther south and west these grade to limestone. During the Pennsylvanian (Upper Carboniferous) and Permian, deltaic or coastal plain deposits including coal beds spread westward across what had been the geosyncline and onto the platform to the west. In western North America the Cordilleran Geosyncline continued to deepen and received marine sediments as late as Permian time.

The Variscan and Uralian geosynclines

The European geosyncline is called the Variscan. In England, Mississippian (Lower Carboniferous) formations are largely limestone; in Scotland they are limy sandstone. Major folding at the end of the Mississippian produced Late Paleozoic mountains, and the great European coal deposits formed in the basins around them. Central Europe was unstable at this time, and episodes of folding are recorded by numerous unconformities. A geosyncline at the site of the Ural Mountains had begun deepening by Silurian time; in it accumulated 10,000 feet (3,000 metres) of Mississippian deposits, 2,000 feet (600 metres) of Pennsylvanian, and 7,000 feet (2,200 metres) of Permian.

The Late Paleozoic seas reached their greatest extent during the Mississippian. Thereafter, as sediments were deposited faster than the troughs sank, coastal plain deposits completed the filling of the geosynclines. These deposits and their coal beds are the outstanding feature of the Late Paleozoic; indeed, this part of earth history is referred to as the Carboniferous Period.

Late Paleozoic deposition on the platforms in central Europe and southwestern United States ended in the Permian when the salt beds were deposited. In Europe, these beds (Zechstein) include the famous Stassfurt Salts. The formations, 3,000 feet (900 metres) thick, extend across most of the north German plain. Potassium salts at the centre of the basin are surrounded by a belt of sodium chloride which in turn grades outward to salt-free clastic deposits. Important deposits of potassium and other salts formed also in the Permian basins in western Texas and southeastern New Mexico and in Pennsylvanian basins in western Colorado and eastern Utah.

Late Paleozoic folding and igneous activity. At least five major episodes of folding occurred during the Late Paleozoic. The first was at the end of the Devonian. After Middle Mississippian time regression became general, and the seas became more and more restricted. In Europe, folding occurred at the end of the Mississippian. Folding was renewed during and after the Permian. Similar folding terminated the Appalachian and Cordilleran geosynclines in North America.

Late Paleozoic tillites

Paleoclimates. Tillites are known in Devonian and Permian formations. The Devonian occurrences are in South America and South Africa. By contrast, the marine Devonian in North America, Europe, Southeast Asia, and Australia contains coral limestone, which suggests a relatively warm environment. Possibly the continents have shifted so that what is now the South Atlantic Ocean near 20° west longitude and 40° south latitude was at the South Pole. The north part of the Pacific Ocean would have been at the North Pole. The Equator, in this reconstruction, would have been near the south coast of Australia, extending through the Arabian Peninsula, central Europe, and eastern North America, and diagonally back across the Pacific to Australia.

Permian tillite is widely distributed in southern South America, southern Africa, India, and Australia. The deposits are associated with a distinctive flora (*Glossopteris* flora) that is indicative of a cool climate. The Pennsylvanian and Permian fern forests of Europe and North America, source of the rich coal deposits there, are suggestive of a warm climate. Permian climatic conditions were therefore somewhat like those of the Devonian.

Paleozoic life. One of the most remarkable events of geological history is the abrupt appearance, at the beginning of the Cambrian, of animals representing all of the major phyla. The suddenness with which they appear in the record is especially impressive because, as already noted, in many geosynclines there was continuous deposition of sediments from the Late Precambrian into Paleozoic time. In the midst of this continuous sedimentation a striking change in population occurred. The large numbers of animals with complex morphology that suddenly appear in the geologic record must have been preceded by a long period of evolution of the living forms, but the Precambrian rocks contain no such record.

To be preserved as fossils, animals must have hard parts —shell, bone, or tests. The seas that advanced into the Late Precambrian geosynclines must have contained an abundant and varied fauna of morphologically complex forms without hard parts, and something happened during an early stage of those seas to enable the animals suddenly to develop shells or tests. Most of the first hard parts seem to have been of chitin, but by Middle Cambrian time most animals were utilizing calcium carbonate. Perhaps it was increasing CO_2 (carbon dioxide) in the atmosphere passing some critical threshold that made certain complex carbohydrates available for animal use in manufacturing chitin at first and calcium carbonate later.

Known Cambrian fossils are from marine deposits. Freshwater and terrestrial animals and plants appear later in the Paleozoic, but this may be due to the fact that the Cambrian record is not complete. The Cambrian fauna is dominated by trilobites, arthropods that shed their chitinous exoskeletons during growth. Brachiopods with simple phosphatic shells also appear in the Cambrian; later in the Cambrian shells of calcium carbonate developed. Another Cambrian animal, the archaeocyathid, may have been related to the corals or sponges.

Diversification of life after Cambrian time

During the Ordovician, trilobites and brachiopods became more numerous and more diversified; cephalopods and gastropods also became abundant. Ordovician black shale contains planktonic (floating and swimming forms) animals known as graptolites. These also are present in the Silurian but then decline. The first vertebrates, in the form of fishes, appear in the Ordovician but do not become abundant until the Devonian.

Devonian marine formations are dominated by brachiopods, bryozoa, and corals that are numerous and varied; trilobites and graptolites greatly decline. The first known forests are Devonian, although some land plants are known from the Silurian. Amphibians first appear in the Devonian.

During the Mississippian, stalked echinoderms, crinoids, are abundant, as are bryozoa and thick-shelled brachiopods. During the Pennsylvanian, great forests of fern trees growing in swamps gave rise to extensive coal deposits. Amphibians further evolved, and by Permian time there were reptiles. During the Permian, forests changed from the fernlike forms to plants related to conifers, and these became dominant during the Mesozoic. Among the important marine shells of the Pennsylvanian and Permian are the calcareous protozoa, fusulinids.

Extinctions, or near extinctions, at the end of the Paleozoic include the colonial, rugose, and tabulate corals and the chief groups of Paleozoic plants, the seed ferns and scale trees.

Mineral resources. The Paleozoic rocks and the most important mineral deposits in them are largely of sedimentary origin. Most notable are the fossil fuels—coal, oil, and gas. The principal coal deposits of the world are of Late Paleozoic (Carboniferous) age. Oil and gas resources constitute about 15 percent of the world's total; this is considerably less than the known occurrences in Mesozoic or Cenozoic rocks.

Evaporites yielding potash, sodium, and other salts are major resources in Silurian and Permian formations. Sedimentary iron formations are important in the Appalachian Geosyncline and in European platform deposits.

Metalliferous deposits in Paleozoic rocks locally are important, but many or most of these are due to post-Paleozoic emplacement or enrichment.

MESOZOIC HISTORY

During the Mesozoic, the shields continued to be stable cores of the continents. They had been submerged by shallow seas during the Paleozoic, but when they emerged they remained so until the second half of the Mesozoic. During the Cretaceous the platforms bordering the shields again were shallowly submerged.

Some parts of the continental margins that had been mobile during the Paleozoic became stable platforms during the Mesozoic. Along the Atlantic and Gulf coasts of North America, for example, Cretaceous and younger formations rest on a smoothly bevelled platform of folded and metamorphosed Paleozoic formations.

Triassic stratigraphy and paleogeography. Geosynclines generally were more restricted during the Triassic than during the Permian, but Tethys persisted from the western Mediterranean to eastern Asia, as did the geosyncline along the west side of North and South America.

On the platform Permian evaporites became buried by fluviatile sandstones and conglomerates around the edges of the depositional basins and by marine siltstone with beds of carbonate, anhydrite (calcium sulfate), and rock salt in the inner basin. As the basins became connected with the open sea, limestone was deposited. The Triassic ended with deposition of deltaic beds, in part coal-bearing, followed by brightly coloured, shallow water, marine clay, and sandstone grading upward into the Jurassic.

Alpine Triassic formations

At the present site of the Alps, Triassic formations are 10,000 feet (3,000 metres) thick. They are almost entirely marine, in contrast to the generally land-derived deposits on the platform to the north. The western Tethys (Alps) Geosyncline grew from south to north. Marine Triassic deposits appear first in the south; correlative deposits in the north Alps are marginal continental deposits. As the marine deposits transgressed northward across these continental deposits, reefs began forming in the southern part of the geosyncline. During the Middle Triassic the sea floor in the geosyncline became covered by carbonate deposits, partly of coral and partly of calcareous algae.

Renewed uplift and folding again restricted the sea and led to weathering of the exposed formations, and these became buried under conglomerate, sandstone, and shale as the sea again advanced. The sea reached its maximum extent during the Late Triassic, when thick bedded limestone and dolomite were deposited.

In North America, a geosyncline developed along the western edge of the continent, and more than 25,000 feet (7,700 metres) of eugeosynclinal sediments were deposited in it, derived by erosion of mountains that lay to the west, in what now is the Pacific Ocean. To the east were discontinuous miogeosynclinal basins. Farther east, in what is now the area of the Rocky Mountains and Great Plains, was a continental platform where less than 1,000 feet (300 metres) of sediments accumulated, largely of continental origin. Beyond, to the northeast, was the Canadian Shield, much of it still covered by Paleozoic formations. Triassic geosynclinal deposits also occur along the length of the Andes in South America.

Triassic deposits in the Appalachian and Himalayan belts

In eastern North America, the Triassic is represented by continental red beds with associated dikes, sills, and lava flows of diabase. The red beds, consisting mostly of conglomerate, sandstone, and shale a few thousand feet thick, were deposited as alluvial-fan and river-floodplain deposits in basins formed by faulting. The sills and lavas in the red beds form some prominent topographic features in the eastern United States, including the Holyoke Ridges in Massachusetts; East and West Rock at New Haven, Connecticut; the Palisades of the Hudson; Watchung Ridges in northern New Jersey; and Cemetery Ridge at Gettysburg, Pennsylvania.

In Asia, peninsular India continued to be stable, but continental Triassic sandstone and shale were deposited in linear, downfaulted basins. Marine Triassic in the mountains of Baluchistan and the Himalayas are generally conformable with the underlying Permian. The Lower and Middle Triassic formations are mostly shale, whereas the Upper Triassic is mostly limestone. The deposits are many thousands of feet thick and represent the eastern continuation of Tethys, which continues eastward to Burma and Laos, where the structural trend turns southward.

In peninsular India and on the shields of Africa and South America, continental Triassic and underlying Late Paleozoic formations are much alike and contain closely related floras, the *Glossopteris* flora. Plant species are so closely related as to suggest that the three widely separated areas were connected, probably by being juxtaposed and subsequently separated as the continents drifted apart.

Jurassic stratigraphy and paleogeography. The record of the Jurassic Period, like the Triassic, is best known in Europe, where it was named for the Jura Mountains. The Tethys Geosyncline was extended westward in the region of the western Alps. To the north and northwest, in Germany, France, and England, partly marine and partly continental platform deposits were formed. The Jurassic sea overlapped beyond the limits of the Triassic sea, and the sediments deposited on the platform north of the geosyncline extended across the eroded, folded mountains of Paleozoic formations.

Early Jurassic formations in Europe are mostly dark shale. The Middle Jurassic formations are mostly brown sandstone and oolitic limestone, and the Late Jurassic are mostly limestone.

Formation of salt domes and diapirs

The Jurassic generally was a period of structural quiet in Europe, but two important structural developments occurred. On the north German plain, the Permian salt deposits became squeezed upward in salt plugs. Thick deposits of salt are plastic, and under the load of Triassic and Jurassic deposits, the light density salt was squeezed upward in masses shaped like teardrops. As the salt rose, the beds at the sides were dragged upward. The rising masses of salt, known as salt domes (*q.v.*) and diapirs, are analogous to air bubbles rising in dough. The upward rise probably began in the Triassic; it continued during the Cretaceous.

The other important Jurassic structural development in Europe was in the Alpine Geosyncline, where the broad downwarp became interrupted by submarine, anticlinal ridges. Uplift of the ridges caused submarine sliding; mountainous masses, known as nappes, slid as much as five kilometres into the synclinal basins. The anticlinal ridges divided the geosyncline into northern and southern troughs, which extend eastward as the Carpathian and Dinaride geosynclines. They in turn rejoin in the geosyncline extending through Asia Minor and along the Caucasus Mountains.

Tethys continued eastward along the site of the Himalayas, where the Jurassic formations are largely carbonates conformable on those of the Triassic, although the upper part of the section is shale grading upward to sandstone in the Cretaceous. Farther east the geosyncline divides irregularly.

In North and South America, a Jurassic geosyncline extending the length of the Pacific coast of both continents consisted of a eugeosyncline on the west and miogeosyncline on the east. The sediments were derived from mountains to the west in what now is the Pacific Ocean. In California, 30,000 feet (9,200 metres) of Lower and Middle Jurassic sediments were deposited. A second trough was in western Utah. Eastward the formations thin to 1,000 feet (300 metres) or less on the platform at the site of the Colorado Plateau and Rocky Mountains, and they thin out against the shield.

Cretaceous stratigraphy and paleogeography. The Cretaceous Period was named for the chalk beds that outcrop in northwest Europe, but chalk is by no means the characteristic Cretaceous rock.

Tethys continued to develop and was extended northward along the site of the Pyrenees. Shallow seas lay to the north and wrapped around the Baltic Shield. In the Alps, Cretaceous formations aggregate about 9,000 feet (2,700 metres) thick; on the platform to the north they are up to 3,000 feet (920 metres) thick but thin against

the shield. Toward the west in the geosyncline the deposits are calcareous; northeastward they are shaly and sandy. In the Bavarian–Austrian foothills are deposits consisting of sandstone, breccia, siliceous limestone, and chalk derived by erosion of ridges within the geosyncline and the accumulation of sediments in rapidly subsiding troughs north of the ridges. These deposits, termed flysch, are sparsely fossiliferous but are believed to range from Early Cretaceous to older Tertiary in age.

Cretaceous transgressions and regressions of the seas

In Europe, the Early Cretaceous began with a stage of regression followed by major transgression that ended when there was folding in the eastern Alps and northwest Germany. This folding was followed by renewed transgression at the beginning of Late Cretaceous time, and again the transgression was interrupted by folding in the Alps, Dinarides, and northwest Germany. Following this, the Cretaceous sea reached its maximum extent in latest Cretaceous time, and this stage was terminated by folding of the Alps and Apennines at the end of the Cretaceous and beginning of the Tertiary.

In North America, a Lower Cretaceous geosyncline in northern Mexico and southwestern United States received as much as 15,000 feet (4,600 metres) of sediments. A similar but smaller geosyncline in the Great Plains in Canada received 5,000 feet (1,500 metres) of sediments. A third one, in northwest California and westernmost Oregon, Washington, and British Columbia, received a maximum of 25,000 feet (7,700 metres) of sediments.

Marine transgression continued during the Upper Cretaceous, and the Gulf of Mexico became connected with the Arctic Ocean. The trough of this geosyncline extended northwestward across Texas and the transcontinental arch to the site of the Rocky Mountains and then northward along the Rockies to the Arctic. A maximum of about 15,000 feet (4,600 metres) of sediments was deposited in this trough. A second trough along the Pacific Coast continued to develop and received up to 10,000 feet (3,000 metres) of sediments.

Mesozoic folding and igneous activity. Early Mesozoic deformation occurred during the Triassic in eastern North America. During the Middle Jurassic, granite batholiths were emplaced along the Pacific Coast of North and South America. Folding is not necessarily accompanied by granitic intrusion, but granitic intrusions are accompanied by folding. Radiometric age determinations of granitic rocks in the Sierra Nevada suggest the occurrence of at least five intrusive episodes there: Early and Middle Jurassic, Middle and Late Jurassic, Late Jurassic, Early Cretaceous, Late Cretaceous. If age determinations were available for rocks along the length of the Pacific Coast, continuous Mesozoic igneous and structural activity along the Pacific edge of the continents probably would be revealed.

Deformation of Tethyan deposits; continental drift

Tethys, which was well formed by the Late Paleozoic, persisted without much deformation until the end of the Turonian Stage of the Cretaceous, when the geosynclinal sediments, together with part of the crystalline basement, were folded and uplifted to form the core of the western Alps. Similar uplift and folding in the northeastern Alps produced the troughs that received the marine flysch deposits.

At the end of the Mesozoic Tethys was folded and uplifted, as was the Upper Cretaceous geosyncline that extended along western North America from the Gulf of Mexico to the Arctic Ocean.

In the Paraná and Uruguay River valleys in South America are vast quantities of basalt that were erupted during the Jurassic. In South Africa extensive dikes and sills of basalt have been interpreted as accompanying the breakup of the equatorial continent (Gondwanaland) that included lands of the Southern Hemisphere. The breakup occurred during the Cretaceous; this dating is suggested by the occurrence of Cretaceous sediments (overlain by Tertiary and Quaternary) resting on basalt in the South Atlantic Ocean.

Mesozoic life. Mesozoic plants and animals differ greatly from those of the Paleozoic. Whereas Paleozoic trees are giant ferns, Triassic plants include groups still extant, such as conifers. Broadleaf trees became abundant during the Cretaceous and probably contributed greatly to the evolution of those mammals and birds that depend on nuts and fruits of such plants.

Mesozoic life: the age of reptiles

The Mesozoic commonly is known as the age of reptiles because of the dinosaurs, which survived only to the end of the Mesozoic. Most Triassic dinosaurs were small; Jurassic dinosaurs included some weighing many tons. In Cretaceous time the horned dinosaurs appeared.

Birds first appear in the Jurassic. Mammals appear in the Triassic but were small and remained so until the end of the Mesozoic.

As early as Triassic time modern forms of corals, clams, and gastropods become common. The most distinctive Mesozoic marine forms are the ammonites, whose rapid evolution and wide extent have provided the basis for the finest chronologic subdivision of the geologic column. They underwent three great extinctions, first in the Late Permian, from which only one stock survived; again in the Late Jurassic; and finally at the end of the Mesozoic, when total extinction occurred. Subdivisions of the Mesozoic are based on evolutionary changes in the ammonites. The end of the Mesozoic may be further characterized as having no species of invertebrates or vertebrate animals surviving to the present.

Of equal or perhaps greater importance in earth history is the massive extinction of microscopic forms of marine plankton at the end of the Mesozoic. Of 44 taxa of calcareous microscopic plankton, 34 did not survive from the Mesozoic to the Cenozoic. Another large group that became extinct were the rudistids, a group of benthonic (bottom-dwelling) pelecypods thought to have lived in shallow water.

Why these massive extinctions during so brief a span of geologic time? An explanation based on predators seems inadequate. Excessive radiation caused by some cosmic event does not seem to fit with the survival of the land flora or the extermination of some benthonic forms that would have been shielded from radiation by water. A change in ocean chemistry may have exterminated part of the microscopic marine fauna and flora and so reduced the rest that the loss of this population deprived other populations of their food supplies. The problem resembles that of the seeming population explosion at the beginning of the Cambrian; an understanding of one might lead to better understanding of the other.

Mineral resources. Principal mineral resources in Mesozoic rocks include oolitic iron in Jurassic formations in England, France, and Germany, which makes the Jurassic deposits the most important Phanerozoic source of iron. Fossil fuels are major. Coal in the Cretaceous in western North America is second only to the Pennsylvanian. Oil and gas from the Cretaceous in Texas, California, the Great Plains, the Wyoming Basin, and Venezuela is second in value only to production from Cenozoic fields. Major deposits of uranium are in Triassic and Jurassic platform deposits on the Colorado Plateau.

Copper, tin, silver, and gold deposits are associated with the Late Mesozoic granite intrusions, including the Mother Lode deposits in California that caused the 1849 Gold Rush. In Africa, diamonds occur in volcanic pipes of Jurassic or Cretaceous age.

CENOZOIC HISTORY

The Cenozoic Era, comprising the last 65,000,000 years of earth history, is divided into the periods and epochs listed in Table 2.

Massive extinction of biota at the end of the Cretaceous led to the rise and dominance of mammals, modern shellfish, deciduous trees and other flowering plants. Geosynclines that had formed during the Mesozoic were ended by folding and uplift. The high mountain chains of the world are largely Cenozoic for this reason. Also, the continents and oceans assumed their modern shapes during Cenozoic time. The last 2,500,000 to 3,000,000 years witnessed climatic changes sufficient to cause several major glaciations; produce lakes in the deserts of the southwestern United States; led to substantial shifts in the bio-

Table 2: Division of the Cenozoic Era by Period and Epoch

period		epoch	approximate age of time boundary
Quaternary		Holocene (Recent)	began about 10,000 years ago
		Pleistocene	began about 2,500,000 years ago
Tertiary	Neogene	Pliocene	began 7,000,000 years ago
		Miocene	began 26,000,000 years ago
	Paleogene	Oligocene	began 38,000,000 years ago
		Eocene	began 54,000,000 years ago
		Paleocene	began 65,000,000 years ago

ta, such as the southward migration of North Sea species of shellfish into the Mediterranean; and other changes in soils and ground conditions.

End of Tethys in Europe and growth of the Alps. Uplift of the Alps had begun during the Cretaceous and it continued into the Paleocene and Eocene. A depression north of the uplift received flysch deposits and these overlapped northward as the trough and Alpine thrust sheet advanced northward onto what had been platform.

Flysch and molasse deposits and subsequent deformation

This Early Tertiary deformation greatly restricted the Tethys Sea, and the low areas became buried by as much as 12,000 feet (3,700 metres) of Oligocene and Miocene clastic deposits derived by erosion of the newly formed mountains by northward flowing streams. The deposits, known as molasse, include conglomerate and sandstone in alluvial fans that grade northward to alluvium, which in turn grades northward to lake or brackish water deposits. The molasse overlaps and extends northward beyond the marine flysch. The flysch was involved in this folding and so were the more southerly and early formations of the molasse. South of the uplift the molasse also overlaps and extends beyond the flysch.

Finally, during the Pliocene and Quaternary, the Alps were uplifted by broad arching, and the sea was expelled. The Alps owe their great altitude to this very late uplift, and they owe their spectacular landforms to later sculpturing by Pleistocene glaciers.

The platform north and northwest of the Alps was largely land at the beginning of the Tertiary. Marine transgression reached a maximum in middle and late Eocene time and again during the early and middle Oligocene. The Paris Basin was a bay. After Middle Tertiary time regression occurred, and the platform emerged.

Tethys has not completely ended. It still exists as the Mediterranean Sea, and the earth movements that reduced Tethys continue in the form of regional earthquakes and the active volcanism in Sicily and Italy.

End of Tethys in Africa and Asia. On the south flank of the Tethys Geosyncline in northwest Africa, marine transgression was followed by folding and uplift during the Eocene and at the end of the Oligocene. Two straits connected the geosyncline with the Atlantic Ocean, one in southern Spain and one in the Atlas Tellien (Tell Atlas) of northern Africa.

East African rift system and Himalayan development

Most of the rest of Africa remained stable during the Cenozoic, but East Africa developed a unique, extensive system of rift valleys (*q.v.*) resulting from block faulting. The fault valleys extend from Lake Nyasa to the Red Sea, the Gulf of Aqaba, and the Dead Sea and join the south flank of the Tethys Geosyncline at the head of the Jordan River Valley. The rifting may have started as early as Cretaceous time, but most of the displacement is Cenozoic, and some is as young as Pleistocene. It is today a belt of moderately active seismicity.

In India, at the eastern end of Tethys, uplift and folding of the Himalaya Mountains began toward the close of the Eocene and reached a climax during the middle Miocene, when there was much folding, thrust faulting, and intrusion of granite in the mountain belt. South of the Himalayas a trough separated the mobile geosynclinal belt from the stable shield of peninsular India, and in it was deposited 15,000 to 20,000 feet (4,600–6,200 metres) of upper Miocene to lower Pleistocene fluviatile and lacustrine sediments (Siwalik Series). These formations are folded in the Siwalik Hills, for which they were named, at the south foot of the Himalayas and probably extend southward under the Indo-Ganges Plain.

On the shield of peninsular India, during latest Cretaceous and earliest Tertiary time, the plateau basalts (Deccan Traps) were erupted. The lavas cover 200,000 square miles and are up to 5,000 feet (1,500 metres) thick.

The belt of Tertiary Himalayan folding turns southward in Burma and continues southeastward through the islands of Indonesia. There was folding during the Miocene, and the continued activity in the mobile belt is reflected in the seismically active belt of volcanoes in western Sumatra, southern Java, and the lesser islands to the east. Southwest of this belt is a structural trough, and south of that there was Miocene folding, but volcanoes were absent. North of the volcanic belt, in Malaysia and parts of Borneo, a platform existed on rocks that had been folded during the Paleozoic and Mesozoic.

Volcanism and earthquakes around the Pacific. The belt of active earth movement and volcanism noted in Indonesia turns northward in the western part of the Pacific Ocean and forms long island arcs that are convex toward the ocean basin at the Philippine Islands, Japan, the Kurile Islands, the Kamchatka Peninsula, and the Aleutian Islands. A second earthquake belt with active volcanoes extends southwestward through the Solomon Islands and New Hebrides. New Zealand is on the projection of this arc. On the opposite side of the ocean, along the coast of North and South America, earthquakes are frequent and volcanoes are active or have been recently. Around all of the Pacific the depth of the earthquake centres is least toward the ocean basin and greatest toward the bordering continents, as if the continents were drifting toward the ocean basin.

The island arcs that rim the Pacific

Volcanoes surrounding the Pacific Ocean have erupted some basalt but greater amounts of more siliceous rocks, and this has been attributed to melting of continental sialic rocks as well as basaltic material from the lower part of the crust or even from below it. By contrast, volcanoes within the ocean basin, like those in Hawaii, are entirely basaltic. Gravity, seismic, and other evidence suggests that the Pacific is underlain almost entirely by basalt. The Paleozoic and Mesozoic geosynclines along the Pacific side of North and South America, however, received sediments derived from mountainous landmasses of sialic composition that were located in what is now the Pacific Ocean. One theory holds that North and South America have been drifting westward, and the lost mountains have foundered and have been overridden by the westward drifting continents. Whatever the cause, the opposite sides of the Pacific Ocean, both youthful, are very different structurally.

Structural development of North and South America. Like the other continents, North and South America rose above sea level during the Cenozoic, and only the edges were transgressed by the sea. Many of the Cenozoic formations on these continents are terrestrial deposits derived by erosion of the mountains that were uplifted at the end of Cretaceous and beginning of Tertiary time.

Development of the Andes and Rocky Mountains

In South America, Early Tertiary marine sediments of Eocene (?) and Oligocene age were deposited in the upper basin of the Amazon River. According to one theory, the embayment extended eastward from the Pacific coast of Peru; the lower Amazon at that time would have been an elevated plain draining westward. When the Andes were uplifted in Middle Tertiary time the outlet to the Pacific was closed, and the basin, filled with Tertiary sediments, spilled eastward to the Atlantic. Uplift of the Andes in Peru (peaks above 20,000 feet [6,200 metres]) continued to Late Pleistocene time. The mountains are probably still rising, as indicated by frequent strong earthquakes, fresh fault scarps, and volcanic activity.

During the Cenozoic, the Isthmus of Panama was submerged intermittently, and at such times the two Americas were separated. When the continents were connected the vertebrate faunas mixed; they developed independently when the two continents were isolated.

In North America, at the end of the Cretaceous and be-

ginning of the Tertiary, the Rocky Mountains rose from what had been the Upper Cretaceous geosyncline. In Early Tertiary time the Great Plains and foot of the Rocky Mountains were near sea level. The drainage was disintegrated and collected in basins such as the Green River Basin, Powder River Basin, San Juan Basin, and others. In Oligocene time the basins overflowed, and as the region became uplifted during the Middle Tertiary, drainage eastward off the Rocky Mountains built broad alluvial plains that extended to the central United States. By Late Tertiary time the drainage, collected in a few valleys, could join the Mississippi River system. Since then, as uplift continued, the drainage system has become incised into the older Tertiary fans. Drainage westward off the mountains reached Grand Canyon, and the Colorado River was initiated in Miocene time.

Along the Pacific Coast of North America small deep basins formed during the Cenozoic (*e.g.*, the Los Angeles and Ventura basins) and became filled with sediments several miles thick. During the Cenozoic, the Mesozoic granite batholith of the Sierra Nevada was faulted on the east, uplifted, and tilted west. To the north, the Cascades began in the Early Tertiary as a series of shield volcanoes, and in the Late Tertiary and Quaternary on top of this volcanic ridge were built the magnificent volcanic cones of the High Cascades.

East of the Cascades is one of the world's great plateau basalts, the Columbia Plateau. Most of the lavas were erupted during the Middle Tertiary, but there were sporadic earlier eruptions, and the volcanism has continued to the Quaternary. At the east end of the volcanic field, at Yellowstone National Park, the volcanic activity is recent enough to maintain the celebrated hot springs and geysers.

The eastern edge of North America had been submerged during the Cretaceous and very gradually emerged during the Cenozoic. The Atlantic coastal plain formations rest on a platform of folded and faulted Paleozoic and Triassic rocks.

The Americas are thought to be drifting westward into the Pacific and away from Europe and Africa. The rate of drift may be as great as two centimetres a year, which would be enough to span the Atlantic since Late Mesozoic time.

Pleistocene glaciations. The Pleistocene Epoch lasted about 2,500,000 years and has been called the Glacial Period or Ice Age because at several times ice caps formed in polar regions. The resulting ice sheets formed continental glaciers that spread into temperate latitudes. In North America the ice covered all of Canada and advanced southward into the United States as far as Long Island and the Ohio and Missouri rivers. In Eurasia, ice caps formed on the Baltic Shield, the west Siberian lowland, and the Angara Shield and spread southward across the northern European and western Siberian plains. Local ice caps formed also in the mountains beyond the continental glaciers.

Extent of ice sheets

To develop glaciers requires not merely low temperatures but also adequate precipitation. Mountains in Alaska and eastern Siberia had local alpine glaciers but no continental ice cap because the precipitation was not sufficient. Parts of the cold north slope of Alaska, for example, today receive less than five inches of precipitation yearly.

The last glaciation reached its maximum extent some 20,000 to 25,000 years ago and since then has melted back to the position of the present ice caps on Greenland, Antarctica, and some high mountains. The last glaciation was recent enough to have profoundly affected the landforms both directly and indirectly. Drainage systems were disrupted, sea level rose and fell, lakes expanded and contracted, and relics of such changes exist today. The permafrost, or permanently frozen ground of the far north, is one example. See PLEISTOCENE EPOCH for a detailed treatment of this ice age in earth history.

Development of modern fauna and flora. During the Cenozoic the big animals have been mammals. Marsupials and insectivores appeared during the Cretaceous; most of the other orders appeared during the Early Tertiary. Elephants are known in the earliest Tertiary deposits in Italy. The early forms were small; they reached maximum size during the Pleistocene, when mammoths and mastodons were abundant.

The development of the grasses and the seed- and nut-bearing trees and shrubs during the Cenozoic probably was both a cause and an effect of the proliferation of the mammals. Grasses appear in the Middle Tertiary, and about that time horses and camels changed from browsing leaves to grazing (shown by the evolution of their teeth).

At the beginning of the Cenozoic the vegetation covering the land appeared much like that of the present, but the geography of the plants was very different and grasses did not become abundant until the Miocene.

Changes in the marine invertebrate faunas during the Cenozoic can be expressed in terms of percentages of species that have survived to the present. Practically no Mesozoic species survive. Less than 5 percent of the Eocene species survive, and about 15 percent of the Miocene species, 35 percent of Pliocene species, and 95 percent or more of Pleistocene species still survive. Moreover, with changes in temperature of the ocean water caused by the glaciations, cold-water forms have alternately migrated southward and retreated northward.

Marine invertebrate faunas

Earliest man in the world dates to the beginning of the Pleistocene. The earliest remains are from the Lower Pleistocene in Africa where the Australopithecinae (*Australopithecus, Paranthropus, Zinjanthropus*) have been found. In the Middle Pleistocene appears Java man (*Homo* or *Pithecanthropus erectus*), Peking man (*Homo* or *Pithecanthropus pekingensis*), and Heidelberg man (*Homo heidelbergensis*). Neanderthal man (*Homo neanderthalensis*) is found in Upper Pleistocene deposits. Modern man (*Homo sapiens*) dates from the last glaciation.

Mineral resources. Cenozoic formations, especially in the Near East, Caribbean region, and United States, contain more than half of the world's oil and gas. Notable deposits in the United States are the fields along the Gulf Coast and in California, and the oil shale deposits in the Eocene Green River Formation on the Colorado Plateau. Other fossil fuels include coal in Japan and vast deposits of lignite on the Great Plains in Canada and the United States. Other deposits in sedimentary rocks include bromine, borax, and other salines in lake beds in the Wyoming Basin and southwestern United States, phosphate in Late Tertiary formations in Florida, and bauxite in Arkansas. Practically all sand and gravel in the world is from Cenozoic deposits, and in addition there are widespread deposits of diatomaceous earth, perlite, pumice, clay, and peat.

Principal Cenozoic metallic ore deposits are associated with igneous intrusions that accompanied the folding and uplift of the Cenozoic mountain ranges. They include tin in Bolivia and copper and silver in the Andes; chromite in Asia Minor; mercury in the Paleozoic rocks at Almadén, Spain; porphyry copper ores in the Basin and Range Province of western North America (including Butte, Montana); molybdenum at Climax, Colorado; gold and silver at numerous districts in western North America; and gold placers, especially those at the west front of the Sierra Nevada in California, in Alaska, and in Australia.

BIBLIOGRAPHY. Summaries of world geologic history include: M. GIGNOUX, *Géologie stratigraphique*, 4th ed. (1950; Eng. trans., *Stratigraphic Geology*, 1955); E. KRENKEL (ed.), *Geologie der Erde* (1939); B. KUMMEL, *History of the Earth* (1961); R.C. MOORE, *Introduction to Historical Geology*, 2nd ed. (1958); H. and G. TERMIER, *Histoire géologique de la biosphère* (1952); J.H.F. UMBGROVE, *The Pulse of the Earth*, 2nd ed. (1947); A.O. WOODFORD, *Historical Geology* (1965); and R.J.W. DOUGLAS (ed.), *Geology and Economic Minerals of Canada* (1970).

Summaries of the geologic history of some continents and countries include: R. BRINCKMANN, *Geologic Evolution of Europe*, 2nd rev. ed. (1969; pub. orig. in German, 1956); T.H. CLARK and C.W. STEARN, *Geological Evolution of North America*, 2nd ed. (1968); A.L. DUTOIT, *The Geology of South Africa*, 3rd rev. ed. by S.H. HAUGHTON (1954); A.J. EARDLEY, *Struc-*

tural Geology of North America, 2nd ed. (1962); W.F. JENKS (ed.), *Handbook of South American Geology* (1956); P.B. KING, "The Tectonics of North America: A Discussion to Accompany the Tectonic Map of North America, Scale 1:5,000,000," *Prof. Pap. U.S. Geol. Surv. 628* (1969); J.S. LEE, *The Geology of China* (1939); D.V. NALIVKIN, *The Geology of the U.S.S.R.* (1960; pub. orig. in Russian, 1957); R. SAID, *The Geology of Egypt* (1962); F. TAKAI *et al.* (eds.), *Geology of Japan* (1963); and D.N. WADIA, *Geology of India*, 3rd ed. (1953). See also E. BEDERKE and H.G. WUNDERLICH, *Atlas zur Geologie* (1968), an atlas of world geology.

(C.B.H.)

Earth, Gravitational Field of

The gravitational field, or gravity, of the Earth is manifested as the force acting on a free, unsupported body, causing it to move in the general direction of the Earth's centre. The speed of the body is not uniform; it increases with time and distance, the rate of change in speed being called acceleration. Acceleration is a measure of the Earth's gravitational field and is denoted in the literature by the symbol *g;* it represents one of Earth's fundamental mechanical properties (see EARTH, MECHANICAL PROPERTIES OF). The Earth's gravity is not fixed but varies from place to place on the surface, as well as in the Earth's interior. It is expressed as an increase of velocity per second; and, because velocity is measured in units of centimetres per second or feet per second, acceleration is stated in centimetres per second per second (denoted as cm/sec^2) or feet per second per second (ft/sec^2).

The average gravitational field at the Earth's surface is 979.76 cm/sec^2, or 32.144 ft/sec^2. The main variation of this acceleration is with latitude, or distance from the Equator. At the Equator it averages about 978.03 cm/sec^2; at the poles, about 983.22 cm/sec^2. There are other variations of about ±.04 cm/sec^2, which are also of interest: they indicate variations of density within the Earth that are of scientific importance, and they affect measurement systems designed to obtain accurate locations, such as geodetic control surveys and artificial-satellite-tracking networks. The expression "study of the Earth's gravitational field" is conventionally used to mean the study of the properties of the Earth's interior as inferred from the gravitational acceleration at the surface. This type of study is the subject of this article, which includes the definition of terms used to express the field, measurement of the Earth's field, and the geological and geophysical interpretations that can be made.

The study of the Earth's gravitational acceleration to infer its effect on measuring systems and its relation to other modes of expressing the field is physical geodesy (see EARTH, FIGURE OF). Other related articles include GRAVITATION; EARTH, STRUCTURE AND COMPOSITION OF; RIFT VALLEYS; OCEANIC RIDGES; MOUNTAIN-BUILDING PROCESSES; ROCKS, PHYSICAL PROPERTIES OF; EARTH, MECHANICAL PROPERTIES OF; and EARTH AS A PLANET.

DEFINITION OF GRAVITY

According to Newton's second law of motion, the force between two bodies is directly proportionate to the product of their masses and inversely proportionate to the square of the distance between them. A body of mass (symbolized as m) at the surface of the Earth is attracted as though all the mass of the Earth (M) were concentrated at its centre. The force of attraction between the body and the Earth (F) will be equal to the product of a constant of proportionality (G) and the masses of the body (m) and Earth (M), divided by the distance between the body and the Earth (R) squared. This can be written in equation form as $F = GmM/R^2$, in which R is the separation distance (*i.e.*, the radius of the Earth), m and M are the masses of the body and the Earth, and G is the constant of proportionality (called the universal gravitational constant). A free body acted upon by this force is accelerated toward the centre by an amount (g) equal to the force divided by its mass—that is, $g = F/m$—or, substituting the value of force from the former equation, the acceleration of gravity is equal to the gravitational constant times the Earth's mass divided by its radius squared; *i.e.*, $g = GM/R^2$, which is seen to be independent of the mass of the body. As this equation shows, from the acceleration of any mass relative to the Earth, the Earth's total mass and something about its mass distribution can be determined.

Hydrostatic figure of the Earth. If the Earth were a perfectly homogeneous sphere, then the acceleration at every point on its surface would be GM/R^2, in which G is the gravitational constant (6.67×10^{-8} cm^{3}gm^{-1} sec^{-2}), M is the Earth's mass (5.97×10^{27} grams [6.6×10^{21} tons]), and R is its mean radius (6.37×10^8 centimetres [2.09×10^7 feet]). The major departure from this gravitation law arises from the rotation of the Earth. The rotation of the Earth on its axis adds a component to the acceleration at its surface through centrifugal acceleration (approximately 3.36 cm/sec^2 [0.11 ft/sec^2] at the Equator but zero at the poles). This centrifugal acceleration causes an additional effect on gravity by distorting the shape of the Earth, which is not rigid, thus changing slightly the distance of the surface from the centre of mass. A homogeneous fluid sphere would be distorted into an ellipsoid of revolution, with a slight flattening at the poles. The Earth has an appreciable increase in density toward the centre, so that its equivalent rotating fluid figure (called its hydrostatic figure) would differ slightly from an ellipsoid. The shape of an ellipsoid is conventionally expressed by its flattening (f) expressed as the (dimensionless) ratio of the difference between the equatorial radius (a) and the polar radius (b) to the equatorial radius, or $f = (a - b)/a$. The flattening (f) of the hydrostatic figure of the Earth is 1/299.8, which is a reference figure for geophysical purposes.

Effects of Earth's rotation

Reference ellipsoid and standard gravity. Because the hydrostatic figure differs by only a few metres from an ellipsoid, it is mathematically convenient to use one as a reference figure. The ellipsoid that best fits the observational data has a flattening of 1/298.25 (this difference from the hydrostatic figure implies stresses in the Earth); nevertheless, the ellipsoid that is used as the basis for referring gravity measurements has a flattening of 1/297, because this value was the best estimate when the International Union of Geodesy and Geophysics adopted a standard formula in 1924. This formula for standard gravity (denoted by the Greek letter gamma, γ) is equal, in centimetres per second per second, to 978.049 times a correction factor composed of sine terms of the Earth's latitude (symbolized by the Greek letter phi, ϕ):

$$\gamma = 978.049\,(1 + 0.0052884 \sin^2 \phi - 0.0000059 \sin^2 2\phi).$$

This formula applies to gravity at sea level. Besides giving an incorrect variation with latitude, the equatorial value ($\phi = 0$) of gravity of 978.049 cm/sec^2 is too high and 978.032 cm/sec^2 is now believed to be a better estimate. Because many data have been expressed in terms of the 1924 standard and because the difference has no significance for most applications of gravity data, the standard will probably never be changed.

Gravitational attraction decreases with altitude, and so gravity values at an elevation above sea level are customarily corrected to apply at sea level. These corrected observations are called free-air gravity values, because the mass of Earth material between the observation station and sea level directly below is neglected. A free-air gravity value thus represents the gravity at sea level (g_0) and is equal to the gravity at the observation station (g_h) plus a numerical factor, 0.3086×10^{-3} times the elevation (h) in metres of the station above sea level; *i.e.*, $g_0 = g_h + 0.3086 \times 10^{-3}\, h$ cm/sec^2.

Gravity above sea level

When the crustal Earth matter between the observation station and sea level is taken into account, a correction for its gravitational attraction must be considered. The effect of such matter can be approximated by assuming that it is an infinitely broad slab having the same thickness as the elevation. The gravity value corrected for sea level in this manner is called the Bouguer gravity (see below *Bouguer anomaly*) value, after Pierre Bouguer, a mathematician and hydrographer of 18th-century France. The Bouguer gravity value (g_B) is equal to gravity at the elevation of the observation station (g_h) minus a factor 0.04815 times density (symbolized by the Greek

letter rho, ρ) of the material slab times the elevation (h) in metres: $g_B = g_h - 0.04815\rho\, h$ cm/sec^2.

Gravity anomalies. *Free-air anomaly.* The quantity symbolized by Δg represents the difference between the free-air gravity value (g_0) and the standard gravity value (γ) cited above, or $\Delta g = g_0 - \gamma$, and is known as the gravity anomaly. It is measured in units of the gal, a conventional term for the acceleration unit of one centimetre per second per second, named in honour of the Italian Galileo. Because the gravity anomaly Δg is always less than 0.5 cm/sec^2 and usually less than 0.1 cm/sec^2, the unit customarily used is the milligal, 0.001 cm/sec^2.

The sea-level surface to which the gravity anomaly refers is known as the geoid (Figure 1). The geoid represents a surface that is everywhere at right angles to the direction of gravity. Over the oceans the geoid coincides with the mean (a statistical average, taken midway between two extreme values) sea level. Hence, its continuation on the land is inside the Earth.

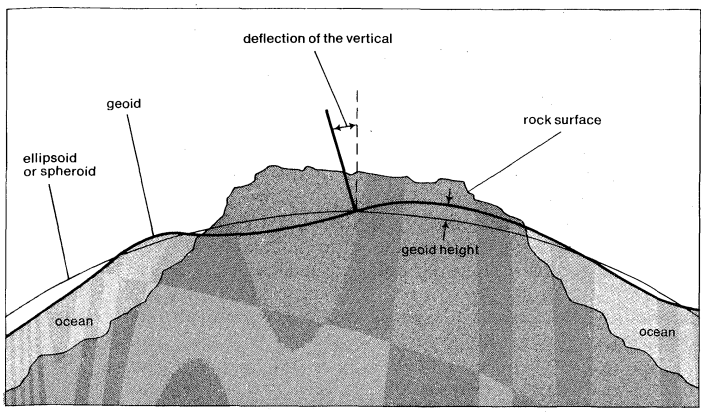


Figure 1: Deflection of the vertical from the geoid to the spheroid (see text).

Bouguer anomaly. When the mass of the Earth material between the observation station and the geoid is taken into account the anomaly is referred to as the Bouguer gravity anomaly, Δg_B. It is equal to the Bouguer gravity value minus the standard gravity value (γ), or $\Delta g_B = g_B - \gamma$; that is, $\Delta g_B = g_h - .04815\rho\, h - \gamma$, in gal units. The error caused by the flat-slab assumption in the Bouguer anomaly is usually less than one milligal; in mountainous terrain, however, it may amount to more than 50 milligals. A topographic correction must be applied, therefore, in which the differential attractions of mountains higher and valleys lower than the gravity-observing station are computed from a topographic map.

The free-air gravity anomalies (Δg) normally have a positive correction with height above sea level in a given region, because of the attraction of the extra matter. The Bouguer anomalies (Δg_B), however, have a much more pronounced negative correlation with elevation, as though the rocks had a density significantly less than what they are observed to have (the value of ρ used in the equations), 2.5 to 2.8 grams per cubic centimetre. The hypothesis therefore arose that in the topography beneath sea level there is a compensation of equal mass per unit area but of opposite sign.

Isostatic anomalies. Gravity anomalies corrected for a compensation in a manner similar to the topographic correction are known as isostatic anomalies. Isostasy literally means "equal standing." There are two principal types of isostatic correction, the Pratt–Hayford and the Airy–Heiskanen. As shown in Figure 2, the Pratt–Hayford correction assumes that the compensation is accomplished by horizontal variations in density (symbolized by ρ) down to the level of equal pressure, whereas the Airy–Heiskanen correction assumes that it is accomplished by variations in the thickness of a layer of constant density. The Pratt–Hayford correction was originally suggested in the 19th century by John Henry Pratt, an English archdeacon and amateur scientist, and later developed in detail by a United States civil engineer, John Fillmore Hayford; and the Airy–Heiskanen correction was first proposed in the 19th century by an English astronomer, George Biddell Airy, and later developed by the Finnish geodesist Weikko Aleksanteri Heiskanen.

The two types of isostatic correction

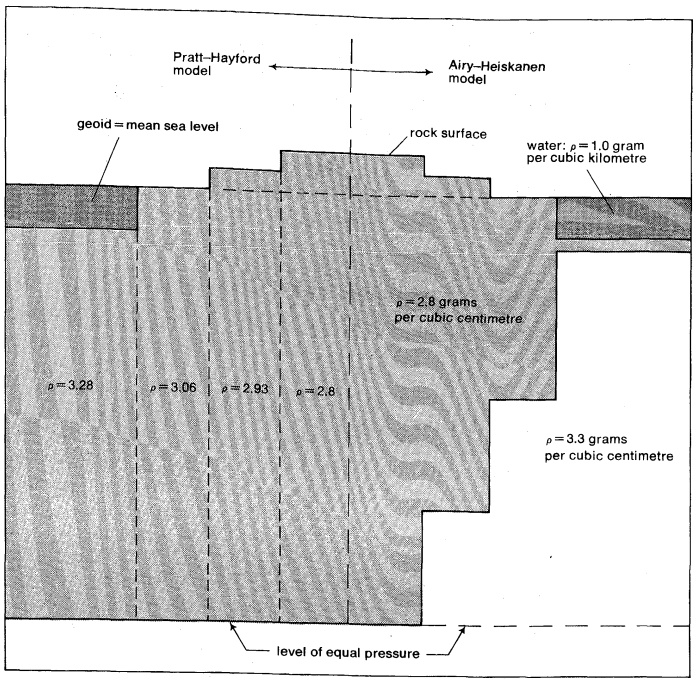


Figure 2: Pratt–Hayford model and Airy–Heiskanen model of isostasy (see text).

By adjusting the depth of compensation, the systematic correlation of isostatic gravity anomalies with elevation can be removed. For Pratt–Hayford anomalies, the depth to the level of equal pressure is usually 60 to 120 kilometres (37.5 to 75 miles). For Airy–Heiskanen anomalies, the crustal thickness corresponding to zero elevation above sea level is usually 20 to 40 kilometres.

Because the crust does have some strength, in the real Earth the isostatic compensation is regional. A regional gravity anomaly reduction correction was devised by a Dutch geophysicist, F.A. Vening Meinesz. In this system the compensation of each topographic feature is distributed over a specified radius. Other, more elaborate types of gravity anomaly have been proposed, but the free-air and Bouguer anomalies are those most commonly used.

Harmonic analysis of gravity. Any variation of a quantity over the surface of a spherical or nearly spherical body, such as the Earth, can be expressed mathematically as the sum of terms known as spherical harmonics. The expression is particularly convenient for theoretical manipulation, in which the gravity anomaly is the radial gradient of the gravitational potential (change of potential with radial distance; potential is related to gravity in that it represents the work required to move a unit mass to infinity against the force of the Earth's gravitational attraction). When spherical harmonics are used, the conventional expression for gravity field is the potential, which varies with latitude and longitude. The dimensionless coefficients C_{lm} and S_{lm}, in which the indices l and m take successive integer values until the potential reaches the approximation desired, are conventionally used to represent gravity as an alternative to the spatial representation by anomaly Δg or geoid height (displacement distance of geoid from reference ellipsoid). The most important coefficient is C_{20} ($l = 2$, $m = 0$), known as the oblateness. The oblateness is the consequence of the distortion of the Earth by its rotation, the gravitational equivalent of the flattening (f). The order of magnitude of coefficient C_{20} is 10^{-3}; of all other coefficients, 10^{-6} or less.

Oblateness

MEASUREMENT OF GRAVITY

There are three principal systems for measuring the gravitational acceleration of the Earth, namely, gravimetry,

astrogeodesy, and artificial-satellite orbit perturbations. Because each system measures the gravitational field differently, a different expression of the field is appropriate for each.

Gravimetry. The gravimetry system measures the gravitational acceleration (g_h) directly. The instruments used are of three distinct types: (1) the timing of the free fall of an object in a vacuum, (2) the timing of the period of a pendulum, which is inversely proportional to the square root of gravitational acceleration, and (3) the measurement of the force exerted on a mass by the tension on a spring required to hold it at a null point (point at which the operative forces balance and hence cancel). Of these instruments, those timing the free fall can measure the absolute value of the gravitational acceleration most accurately; but it is a difficult experiment, performed in only a few laboratories around the world. Type (3), the spring balance (called a gravimeter), is the most sensitive for measuring small differences in gravity from place to place, less than about one cm/sec^2, or 1,000 milligals. Gravimeters are, however, subject to systematic errors such that they require calibration over a greater range of values. Type (2), the pendulums, are less sensitive than gravimeters but are more stable; they are less accurate than free-fall devices but are more portable. Because of these various instrumental characteristics, the entire gravimetric system is a combination of absolute measurements at a few laboratories that are connected to north–south calibration lines by pendulums that measure differences up to about five gals. East–west reference ties of less than one gal difference by gravimeter make up the remainder of the network that serves as a control for field measurements by gravimeter.

Gravimeters have been developed that are extremely portable on land, and the effort spent in using one is actually less than that expended on a survey to obtain the elevation h, a preliminary to the free-air reduction to sea level. Pendulums were taken to sea in a submarine for the first time in 1923. Since then, considerable development in stabilizing devices has made possible measurements in surface vessels and in aircraft, the principal source of inaccuracy often being the navigation to obtain latitude and longitude.

Pendulum measurements at sea

From gravimetry, the free-air anomaly over a large portion of the world is known with varying detail. There still remain, however, large areas without measurements, particularly in the oceans of the Southern Hemisphere (see GRAVITATION).

Astrogeodesy. The vertical is the direction of the force of gravity, perpendicular to the geoid, as indicated by the solid line in Figure 1. The direction of the vertical with respect to the coordinate axes defined by the fixed stars (called inertial coordinates) can be measured by theodolites; *i.e.*, in effect, telescopes that have attached sensitive level bubbles and accurately divided circles. These inertially referred directions can be corrected to directions referred to Earth-fixed coordinates by radio time signals that in effect give the Greenwich sidereal time, the angle between Earth-fixed and inertial coordinate axes.

The normal is the direction of the perpendicular to the reference ellipsoid, as indicated by the dashed line in Figure 1. The normal can be considered as the horizontal coordinates in a system referred to the ellipsoid. Differences in horizontal coordinates from place to place are measured by methods, called triangulation and traverse, that use theodolites in conjunction with such distance-measuring instruments as tapes or electronic devices.

The difference in angle between the vertical and the normal is the deflection of the vertical, which is the same as the slope of the geoid with respect to the ellipsoid. The displacement of the geoid from the reference ellipsoid is known as the geoid height (N). It and the gravity anomaly (Δg) are both expressible in terms of a disturbing potential; that is, the difference in potential between the actual Earth and the reference ellipsoid. The geoid height may then be obtained by integrating the slope variations in the geoid over the anomaly wherever astronomic and geodetic measurements are available. Such measurements have been made over much of the major continents but were not in the early 1970s possible over the oceans. (See EARTH, FIGURE OF; SURVEYING.)

Artificial-satellite orbit perturbations. If the Earth were a homogeneous sphere and there were no atmosphere, Sun, Moon, or planets, then an artificial satellite would move in a perfect ellipse in an unvarying plane. The main difference of an actual satellite orbit from the hypothetical ellipse is represented by the oblateness factor C_{20}, the major effects of which are to cause rotations of a few degrees per day of (1) the plane of motion about the polar axis, called the precession of the node, a node being the point at which the plane of motion intersects the Equator; and (2) the ellipse within the plane, called the precession of the perigee, the point of closest approach. In addition, the satellite oscillates about these steady precessions at rates that are various combinations of the rates of rotation of the node, the perigee, the satellite itself, and the Earth. These oscillations, or perturbations, are of less than five kilometres (three miles) in amplitude and are caused by the other variations in the Earth's gravitational field. Because their effects depend very much on the wavelength of the oscillation and the satellite's direction of motion, for satellite-orbit analysis these variations are best expressed by the spherical harmonic coefficients C_{lm}, S_{lm}.

Effect of oblateness on artificial satellites

Photographic and tracking networks around the world enable the determinations to be made within about two metres (seven feet) of the amplitudes of systematic oscillations in certain selected satellite orbits. This accuracy is sufficient to determine 100 to 200 of the gravitational coefficients C_{lm}, S_{lm} (see EARTH, FIGURE OF; MECHANICS, CELESTIAL).

Results of gravity measurements. Figure 3 shows the Earth's gravitational field expressed in terms of the free-air anomaly. This field is a determination made in 1970 by E.M. Gaposhkin and K. Lambeck of the Smithsonian Astrophysical Observatory from a combination of satellite-orbit perturbations and gravimetry.

Figure 4 is an example of a more limited regional survey made by gravimeters in northern California and northeast Nevada. Here the Earth's gravitational acceleration is expressed in terms of the Bouguer anomaly, as is customary. It is evident from Figure 4 that there is a considerable amount of local variation of the gravitational field as well.

INTERPRETATION OF GRAVITY DATA

Isostasy. The dominant fact about the variations of the Earth's gravitational field is that they are much smaller than would exist if the topography were rigidly supported on a homogeneous interior. This fact was first noticed in the Andes and was later confirmed in the Himalayas: the mountains always showed less attraction than calculated. The explanation for this mildness of gravity variations is assumed to be some sort of isostatic compensation for the topography, such as the alternatives sketched in Figure 2.

Evidence of isostasy appears in Figure 3 because there is no apparent correlation of the free-air gravity anomalies with the topography of the continents and oceans. By statistical analysis, the correlation of free-air anomalies with topography is negligibly small for features of more than 4,000 kilometres (2,500 miles) in extent. For smaller features, the correlation is positive but still much less than would prevail if the topography were rigidly supported.

Evidence of isostasy also appears in Figure 4; the Bouguer anomalies are most negative at places where the topography is highest, along the mountains of the Sierra Nevada Range between Sacramento, California, and Reno, Nevada.

Isostasy is apparently a consequence of the evolution of the Earth's crust coupled with the mechanical weakness of rocks. During the 4,600,000,000 years of Earth history, those topographic features that chanced to be associated with some sort of density compensation have persisted, whereas those that constituted an excess or deficiency were reduced by creep of the rocks in response to the stresses caused by these loads. The rate of this creep

Origin of isostasy

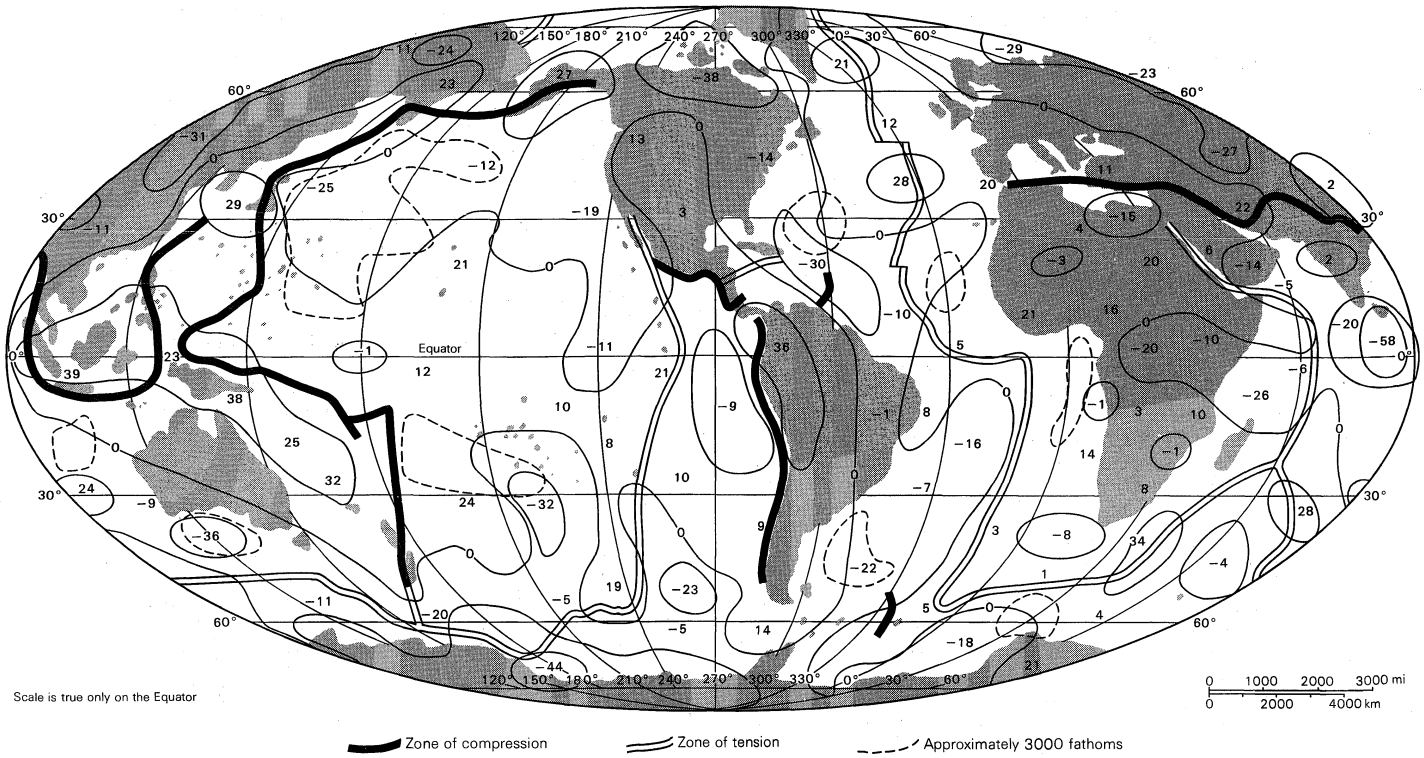


Figure 3: World gravity distribution, showing free-air anomalies in milligals.
After E.M. Gaposhkin and K. Lambeck in W.M. Kaula, *Science*, vol. 169, p. 983 (September 4, 1970); American Association for the Advancement of Science

is directly proportionate to the extent of the load: large features are rapidly obliterated, but small ones of less than 100 kilometres (60 miles) extent can remain for several 10^8 years.

The finds of seismology indicate that the Airy–Heiskanen hypothesis of isostasy probably prevails. The depth to the Mohorovičić Discontinuity (named after a Croatian geologist, Andrija Mohorovičić), the rather sharp interface between crustal rocks (density about 2.8 gm/cm^3) and mantle rocks (density about 3.3 gm/cm^3), is positively correlated with topography. However, there are also many areas in which variations in rock type indicate that a form of Pratt–Hayford compensation exists.

Because isostasy is mainly a consequence of the passive adjustment of rocks in the Earth to stresses, the concept tells little about the primordial forces that have shaped the Earth's crust. Though isostasy is a general characteristic that must be taken into account, it is far short of a full explanation of the variations in the Earth's gravitational field.

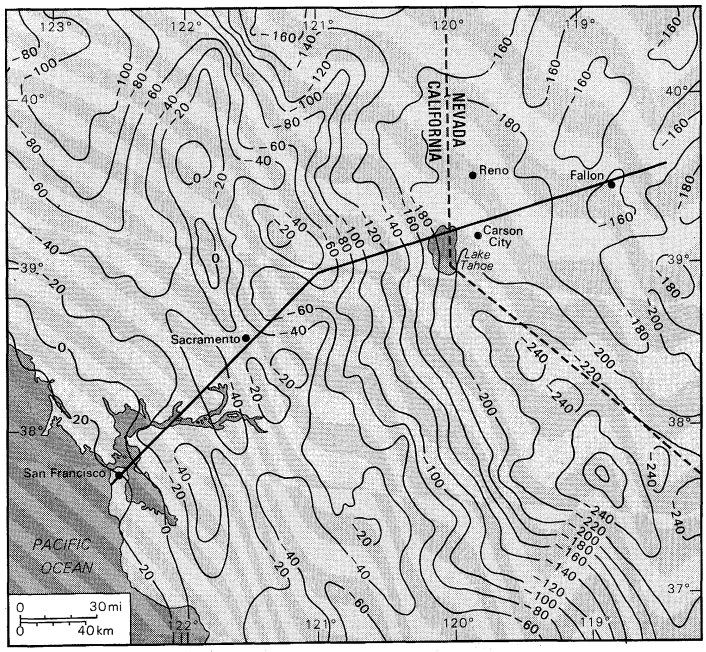


Figure 4: Gravity distribution in the region including the San Francisco to Fallon, Nevada, transect.

Local irregularities: geophysical prospecting. Variations in the gravitational field on a scale of kilometres are necessarily associated with densities and structures of rather shallow rock inferred from the surface geology.

Sometimes erosion and sedimentation will result in the concealment of geologic features uncorrelated with surface geology. In these circumstances, accurate gravimetric surveys have aided in the finding or evaluation of mineral resources. The indications of a gravimetric survey may be direct, as when an ore body consists of a mineral distinctly denser than its surroundings, or indirect, as when certain geologic structures trap economically exploitable deposits.

The use of gravimetry for prospecting requires measurement within plus or minus 0.05 milligal. The interpretation is usually complicated by the need to remove regional trends resulting from deeper seated variations. The ambiguity of the results is normally such that in promising cases further investigation is undertaken by seismological techniques that reveal greater detail. Hence, gravimetry performs mainly a reconnaissance function.

Regional irregularities: crustal structure. Most scientific use of gravimetry has been to investigate structures on a scale of tens to hundreds of kilometres. At the shorter end of this scale, the gravimetry can be most readily correlated with variations in rock type. Toward the longer end of the scale, the gravimetry can be correlated with variations in seismic velocities associated with both natural and artificial sources. These variations in seismic velocities are caused by variations in rock density, which is positively correlated with velocity, or by variations in the depths of interfaces between rocks of different densities.

Correlation of gravimetry with seismic velocities

As an example, Figure 5 is an interpretation of the Bouguer gravity anomalies along the section cut by the line from San Francisco to Fallon, Nevada, in Figure 4. This section crosses five distinct geologic provinces.

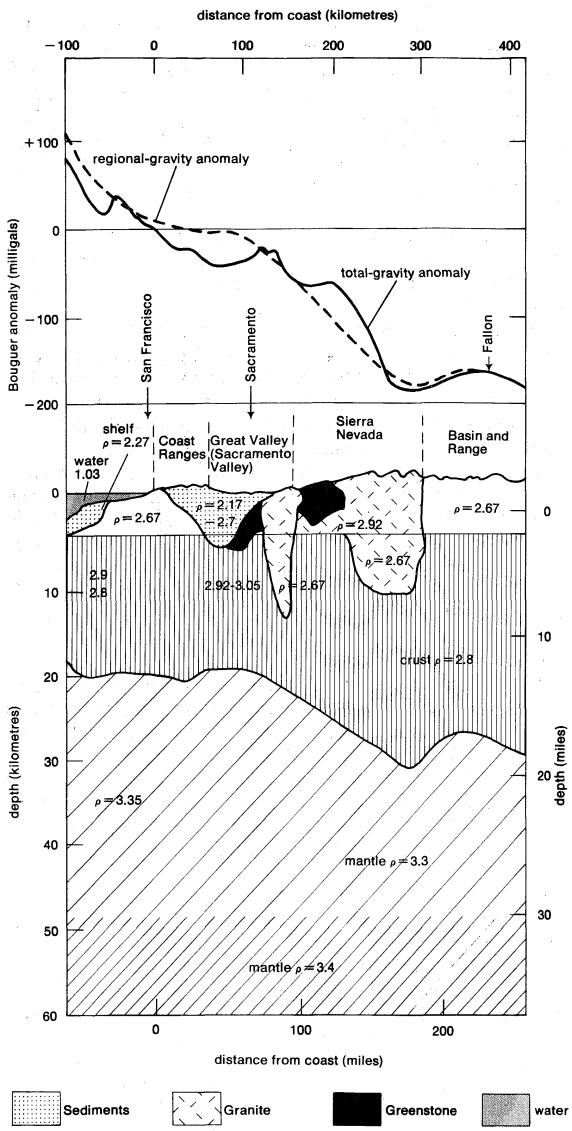


Figure 5: Gravity anomalies and inferred subsurface configuration and density distribution, along the San Francisco to Fallon, Nevada, transect.

From G.A. Thompson and M. Talwani, *Science*, vol. 146, pp. 1539–49. Copyright 1964 by the American Association for the Advancement of Science

From left to right, they are: (1) the Pacific continental slope of North America; (2) the California Coast Ranges, a complicated young, active mountain belt at an angle to the coast; (3) the Great Valley, a heavily sedimented region; (4) the Sierra Nevada, a great block tilted about 2° westward; and (5) the Basin and Range Province, a region of recent faulting, warping, and extension.

The dominant feature of Figure 5 is the inverse correlation of the Bouguer anomaly with topography, mainly a manifestation of isostasy. Variations about this trend of tens of kilometres in extent are explicable by geologic data extrapolated to less than five kilometres' (three miles') depth. Variations below the regional trend are associated with the light sediments of the continental slope and the Great Valley; anomalies above the trend are associated with belts of greenstone, metamorphosed volcanic rocks of high density. Larger variations are explicable by seismological data that reveal low-density rocks of the granitic Sierra Nevada and oscillations of the Mohorovičić interface between crust and mantle. Finally, a portion of the overall trend must be ascribed to a variation in density in the upper mantle itself. Such a variation, together with the shallower than usual Mohorovičić Discontinuity, is characteristic of a region of recent mountain-building activity.

Other types of regional irregularities have been explained by gravity anomalies, always in combination with geologic and seismic data, sometimes in combination with heat-flow or magnetic-anomaly data. Examples include (1) trench and island arcs, which have a sharp negative gravity anomaly over the trench and a more moderate but widespread positive anomaly over the islands; (2) oceanic rises, which have sharp local variations in gravity correlated with a rugged sea bottom but only a mildly positive regional character; and (3) areas of postglacial uplift, which are predominantly negative, as well as regions of sedimented basins (negative), granitic batholiths (negative), and volcanic belts (positive) similar to those occurring in the above-described example. (See RIFT VALLEYS; OCEANIC RIDGES; and MOUNTAIN-BUILDING PROCESSES.)

Large-scale irregularities: global tectonics. Not until 1970 had the determination of the major variations in the gravitational field become accurate enough to show distinct correlations with other global tectonic phenomena (those associated with crustal deformations), for which determinations from paleomagnetic, seismological, and other data have evolved rapidly since 1963. The main features of this global tectonic pattern are shown in Figure 3 by belts of tension, along which crustal material comes up and spreads out, mainly ocean rises; and belts of compression, along which crustal material is thrust back into the mantle, mainly trench and island arcs and mountain belts.

Features correlated with gravity anomalies

The most important correlation of the global gravity field is that large regions of positive gravity anomaly are regions of current geologic change, as measured by the amount of volcanism, earthquakes, heat flow, and mountain building. The one great exception to this rule is the East Asian portion of the Alpide Belt, the compressive belt extending from the Mediterranean to Indonesia in Figure 3. This region is unique in that it is the only case of compression resulting from abutment of two continental masses, in contrast to the usual case of a continent and an ocean basin; apparently, the great thicknesses of low-density crust are too buoyant to be thrust down into the mantle.

The principal regions of negative gravity anomaly are either areas of postglacial uplift, in which the great glaciers of the Ice Age melted away about 10,000 years ago, or basins: topographic lows characterized by thick sediments, no seismic activity, and low temperatures, as inferred from heat-flow measurements and seismic velocities. Ocean basins are marked on Figure 3 by dashed lines corresponding to the 3,000-fathom- (5,500-metre-[18,000-foot-]) depth line.

Correlations of the gravity field with other global tectonic phenomena are enhanced if the variations expressed by fifth and lower degree (corresponding to the maximum value of l that is used) spherical harmonics (*i.e.*, of more than 4,000-kilometre [2,500-mile] extent) are subtracted from the gravity. This fact suggests that part of the field may be caused by the lower mantle, perhaps the core–mantle interface.

The increasing indications that the major tectonic activity in the Earth moves at rates of one to ten centimetres (0.4 to four inches) per year and that the upper mantle contains a zone of relatively low viscosity, or resistance to flow, require that explanations of broad variations in the gravity field be in terms of a dynamic flow system. Probably the greatest effects of this system on the gravity field arise from the behaviour of the lithosphere, the layer of mechanical strength 50–80 kilometres (30–50 miles) thick that includes the crust, in response to flows in the asthenosphere, the low-viscosity zone 80–400 kilometres (50–250 miles) deep.

BIBLIOGRAPHY. The Earth's gravitational field is part of the study of the solid Earth; modern undergraduate texts on this topic include ARTHUR N. STRAHLER, *The Earth Sciences*, 2nd ed. (1971); and JOHN VERHOOGEN *et al.*, *The Earth: An Introduction to Physical Geology* (1970). Texts specifically treating the gravitational field are G.D. GARLAND, *The Earth's Shape and Gravity* (1965); and A.H. COOK, *Gravity and the Earth* (1969). More advanced treatises are by W.A. HEISKANEN and F.A. VENING MEINESZ, *The Earth and Its Gravity Field*

(1958); and HAROLD JEFFREYS, *The Earth*, 5th ed. (1970). Excellent collections of articles on recent investigations of the solid Earth are P.J. HART (ed.), *The Earth's Crust and Upper Mantle* (1969); and E.C. ROBERTSON (ed.), *The Nature of the Solid Earth* (1971). More technical books on specific aspects are W.M. KAULA, *Theory of Satellite Geodesy* (1966); MILTON B. DOBRIN, *Introduction to Geophysical Prospecting*, 2nd ed. (1960); and GUY BOMFORD, *Geodesy*, 2nd ed. (1961).

(W.M.K.)

Earth, Heat Flow in

Heat flow in the Earth is generally described as the rate of heat conduction per unit time and per unit area to the Earth's surface. Units in which heat flow is measured are small in terms of everyday experience: the heat flow unit (HFU) is 10^{-6} calories per square centimetre per second. The average value for the Earth's crust is about 1.5 HFU and is roughly the same for the continents and the ocean floors. Heat flowing at this rate would take about a year to melt a layer of ice only six millimetres thick. The heat received from solar radiation is 2,000 to 3,000 times as great. Yet heat derived from within the Earth during geological time has played an important role in forming the surface features of the Earth.

Early observations of the Earth's heat

The origin of the Earth's interior heat has long been a subject of inquiry. Volcanic eruptions led ancient scholars and alchemists to many and varied speculations on the mystical sources of subterranean fire. In Britain, Robert Boyle in the 17th century made one of the first significant inquiries into the subject of heat flow, however, and, when the first mines subsequently were dug deep into the Earth, it became apparent that a thermal gradient exists—*i.e.*, that the temperature of the Earth increases with depth. In the deepest modern mines, the temperature is so high that air-conditioning equipment must be installed to prevent miners from being roasted.

Truly accurate underwater measurements of heat flow were first made in the 1950s, and only in the 1960s, with the construction of more advanced electronic instrumentation and greater availability of suitable oceangoing vessels, were heat flow measurements made routinely on the ocean floor. Geothermal heat-flow measurements for the Earth's surface are by no means complete.

The thermal history of the Earth is intimately connected with the problem of the origin of the Earth. Similarly, information about the Moon's thermal history should prove valuable in explaining the origin of that satellite. Apollo heat-flow measurements, for example, have indicated a heat-flow value for the Moon about one-half that of the average heat flow for the Earth. Such a heat-flow value implies that most of the present heat emanation from the Moon results from the decay of radioactive isotopes with long half-lives. (The half-life of an element is the amount of time required for half of a given amount of that element to be changed into other elements or isotopes through radioactive decay.) It seems likely that very little of the present lunar heat flow stems from retention of initial temperatures during the period of the Moon's formation.

This article treats the origin of the Earth's heat, its transmission from the interior to the Earth's surface, and certain geological aspects of heat flow. For further information on the Earth's origin and internal constitution, see EARTH, GEOLOGICAL HISTORY OF and EARTH, STRUCTURE AND COMPOSITION OF. For coverage of the elements within the Earth and their radioactive decay, see ELEMENTS, GEOCHEMICAL DISTRIBUTION OF and DATING, RELATIVE AND ABSOLUTE. Articles treating the principal geological aspects of heat flow include SEA-FLOOR SPREADING; CONTINENTAL DRIFT; ORE DEPOSITS; MOUNTAIN-BUILDING PROCESSES; GEYSERS AND FUMAROLES; VOLCANOES; IGNEOUS ROCKS; and ROCK METAMORPHISM, PRINCIPLES OF. For additional information on the heat transmissibility of rocks of several kinds, see also ROCKS, PHYSICAL PROPERTIES OF.

SOURCES OF THE EARTH'S HEAT

There are several possible sources of heat flow in the Earth. These include the original heat of the matter that formed the Earth; the heat produced by the decay of radioactive isotopes within the Earth; gravitational contraction of the Earth as it solidified; tidal heat of friction produced by movement of materials inside the Earth under the influence of gravitational attraction of other heavenly bodies; chemical reactions within the Earth; and electromagnetic interaction with the solar wind, which is a stream of electrons and protons discharged from the Sun.

Of these possibilities, the most significant are the heat involved in the formation of the Earth and the heat produced by the decay of radioactive isotopes within the Earth.

Theories of Earth formation

Original heat of the Earth. One theory of Earth formation holds that the Sun was formed by condensation of interstellar matter. After its formation, a far-reaching cloud of gases and dust particles remained outside it. Gradually, as small particles in the cloud collided with larger ones, they coalesced to form still larger particles. As these bodies grew larger, they acquired more smaller particles through gravitational attraction. Tremendous amounts of heat were generated because the kinetic energy (energy of motion) of the smaller particles striking the large ones was converted into heat. As the surrounding cloud was exhausted of its matter, fewer collisions could occur and less heating took place. Heat was radiated from the Earth's surface, thus leading to the formation of a solid crust by cooling.

Another theory, which is actually a variant of that cited above, involves first the condensation of the gaseous cloud surrounding the Sun, forming bodies (protoplanets) that possess gaseous atmospheres. As the Sun formed, its heat dissipated most of the gases in the atmospheres of these protoplanets, leaving the planets in somewhat their present forms. Again, gravitational compaction and friction from particle collisions are held to have heated the Earth.

Any theory of Earth formation necessarily involves the gradual loss of initial heat through time. This heat loss thus constitutes a possible and plausible component of the Earth's heat flow.

Heat produced by radioactive decay of elements. At some stage in its development, the Earth almost certainly contained large pressures and large quantities of heat, which caused much melting to occur in its interior. Because they crystallize at lower temperatures than iron, silicon, or magnesium compounds, molten radioactive elements were forced upward as the denser materials settled toward the core under the influence of gravity. Heat produced by the decay of radioactive isotopes was brought to the surface by conduction. If some radioactive rocks of the crust were wedged too far down to lose heat to the surface by conduction, the heat given off as they decayed would raise the temperature in their locality sufficiently to melt the surrounding material and so let them rise eventually to a point nearer the surface. Thus heat generated by radioactivity would flow away from its source quickly enough to prevent further melting. As molten materials solidified, crystals would form. Any fluid remaining during this process would carry heat upward by convection. The crystals then would solidify in such a way as to force upward any remaining fluid. In this way a sort of stability would be maintained; the mantle would then be at a temperature below the melting point of any of its components. The solid crust would extend downward just far enough to carry away heat of radioactivity.

The first mentions of radioactivity as a source of the Earth's heat flow were made in England by two physicists, Ernest Rutherford and Robert John Strutt, 4th Lord Rayleigh. In 1906 Rayleigh showed that the amount of heat produced by radioactive decay of radium in the Earth's crust is enough to account for the entire flow of heat from the Earth.

Recent observations and calculations indicate that approximately 80 percent of the Earth's heat flow is due to radioactive decay, whereas the remaining 20 percent is derived from the original heat of the Earth. Naturally occurring radioactive substances with a very long half-

life that are sufficiently concentrated to produce a significant amount of heat are uranium-235, uranium-238, thorium-232, and potassium-40. Uranium, thorium, and potassium are found in measurable amounts in all rocks.

TRANSMISSION OF THE EARTH'S HEAT

Thermal conductivity of rocks. In order to determine the heat flow from the Earth, the thermal conductivity of the rocks in an area of interest and the temperature gradient, or thermal gradient, in this area must be measured. The thermal conductivity of a substance is a measure of the rate at which the substance conducts heat, and the thermal gradient is the rate at which temperature increases with depth.

Rate of heat transmission

Heat is conducted through the Earth at an extremely slow rate. At a few feet below the ground, daily temperature variations are hardly felt. The effects of changing seasons may be felt after a lag of about a year at a depth of approximately 100 metres (330 feet). Thermal conduction, in fact, is such a slow process that heat energy from radioactive decay taking place below a depth of a few hundred kilometres would take more than 4,600,000,000 years—the presently accepted value for the age of the Earth—to reach the surface.

Thermal conductivity of solids, although not very dependent on pressure, decreases somewhat as temperature increases. This is not always true: at high temperatures the thermal conductivity of many materials that do not conduct electricity increases greatly with increasing temperature. That is, it increases proportionally as the third power of the absolute temperature; the dominant heat-transfer process in this case is radiation.

Thermal gradients in the Earth. Thermal gradients in the continental crust show a wide variation, even in areas far removed from volcanic activity. These variations are due largely to differences in the thermal conductivity of various types of rock. The thermal conductivity, however, tends to be low where thermal gradient is high, and vice versa, so that their product, the heat flow, remains constant.

Numerous heat-flow measurements have been made on the ocean floor, and large variations are known to exist. Some of these variations were probably caused by local heat sources such as intrusions of magma or other volcanic events in the vicinity of midoceanic ridges.

Heat transfer is generally a slow process, but it may occur more rapidly under appropriate conditions. The existence of high temperatures approaching melting points in the Earth's mantle (the zone beneath the crust) can reduce the thermal gradient required to transfer a given amount of heat, because radiative heat transfer exceeds conductive transfer under these conditions.

The Earth is thought to have a solid inner core surrounded by a molten outer core. If the material in the outer core is in a state of convective motion, then the thermal gradient in it must be higher than the adiabatic value (that is, the value if no exchange of heat is assumed). Also, because the inner core is solid, the thermal gradient must be lower than the gradient at the melting point; this results because the outer core temperature is higher than the melting point and because the inner core temperature is lower than the melting point temperature. The strong temperature dependence of radiative transfer tends to stabilize internal temperatures.

Heat flow in continental and oceanic regions. Although radioactive rocks are not distributed in the same proportions in the continental and oceanic crust, continental and oceanic heat-flow values are equal within errors of measurement. This surprising equality of the continental and oceanic values implies that radioactivity, the source of most heat, must be similar in magnitude under continents and oceans. This, in turn, implies that the upper mantle is significantly hotter under the oceans than under the continents.

The difficulty with the above implications lies in the requirement the argument imposes on the density of the oceanic mantle, namely, that it must be less dense than the continental mantle because of its greater temperature. The concept of a warmer oceanic mantle also implies that it contains some of the lighter elements that the continental mantle no longer holds and, again, is thereby less dense. If the oceanic mantle were indeed less dense, the shape of the geoid (the surface within or around the Earth that is everywhere perpendicular to the direction of gravity and coincides with mean sea level in the oceans) would be affected. It has been suggested that some type of dynamic balance in the convecting upper mantle may account for the equality of oceanic and continental heat flow.

Reliability of heat-flow data. The surface temperature of the Earth has an annual oscillation that can be neglected at depths greater than 20 metres, but it must be considered an important correction in calculations of properties near the surface. Aside from this theoretical consideration, heat-flow data are affected by instrumental error and local conditions. Heat-flow measurements generally lose no more than 10 percent of their accuracy through instrumental errors.

Effects of topography and water circulation

A topographically rough ocean floor may cause variable water movements near the ocean floor. These movements could cause a variation in the rate of heat exchange in the ocean floor and thus produce variations in the effective temperature of the ocean floor. Generally, however, the correlation of heat flow in the oceans with topography is low, and global analysis shows little correlation of heat flow with topography.

The values of thermal gradients, especially those in shallow holes, may be affected by subsurface water circulation. Water flows in tunnels or in water-bearing seams are easily recognized, but the effect of slow movements of water in aquifers (bodies of highly porous, permeable rock through which water may flow) is not as easily seen. Water movement caused by other mechanisms such as salinity convection may also affect heat transfer in water near the bottom of the sea.

Surface conditions will exhibit change at the edges of lakes or oceans. Also, the effects of glaciation on the Earth's surface temperature may be preserved at depths commonly reached by boreholes. In some deeper holes, effects of changes from as far back as 1,000,000 years ago may be of importance and must be properly accounted for in any calculations.

The correlation between gravity and heat-flow data is not yet well established, since not enough is known yet about the gravity field, the heat-flow field, and any relation between them. Also, gravity readings are determined by the present distribution of the Earth's mass, whereas heat-flow readings may lag gravity readings by millions of years because of the slowness of the thermal conduction process.

Relatively high values of heat flow are associated with oceanic ridges (*q.v.*), the values showing a gradual decrease with distance from the crest of the ridge. It is possible that upward-flowing mantle currents form vertical sheets, intersecting the Earth's surface along the crests of the ridges. These upwelling mantle currents would explain the high heat-flow values and the production of great quantities of basaltic magma that occurs along the loci of the ridges.

GEOLOGICAL ASPECTS OF HEAT FLOW

The age of the Earth: Kelvin's problem. In the 19th century, a physicist, William Thomson (Lord Kelvin), assumed that the Earth formed as a hot gas that condensed and cooled. Kelvin assumed an initially molten crust that solidified from the outside inward. The thermal gradient in the crust was determined by the temperature difference across it, decreasing with the thickness of the crust. Accordingly, the thermal conductivity of the crust fixed the heat flow and the rate of solidification. In the 1890s Kelvin calculated the amount of time necessary for the Earth to reach a state, after solidification, in which the heat flow would have its present value between a maximum of 400,000,000 and a minimum of 20,000,000 years. The discovery of radioactive heat sources in the Earth invalidated Lord Kelvin's calculations, however. The presently accepted value for the age of the Earth is, as noted above, about 4,600,000,000 years.

Convection currents in the Earth. The lack of any significant correlation between heat flow and gravity may be explained better by convection in the Earth's mantle than by conduction, which, as mentioned before, is an extremely slow process. It has been calculated that abnormal readings in the heat flow and gravitational fields are consistent with convection currents with velocities of several centimetres per year.

Sea-floor spreading and continental drift

Sea-floor spreading (*q.v.*), a process in which new sea floor is created at ocean ridges, also points to the possibility of thermal convection. The theory of continental drift (*q.v.*), which holds that all the present continents once formed one large continent, finds some support in the theory of sea-floor spreading, since it seems that the rate at which the continents are separating can be measured (see further ROCK MAGNETISM).

The mechanism of thermal convection is thought to be set in motion by sources of great heat in the mantle and the consequent difference in temperature between the mantle and the base of the crust. The heat at great depths causes some material to become less dense than that surrounding it; this material is thus forced upward. The convection currents thus may rise upward from the bottom of the mantle, cross under the crust, and then return downward, flowing within a huge cell with dimensions as great as the radius of the Earth.

Arguments against the existence of thermal convection in the upper mantle are based on continental structure and on mechanical properties of the mantle. If continental structure should extend to depths of hundreds of kilometres, that would pose a serious obstacle to any large-scale horizontal convection (and to continental drift). Also, convection might homogenize the upper mantle, and so differences between oceanic and continental structure might no longer exist. The mantle should behave like an ultradense liquid for convection to occur. The calculated viscosity necessary for such convective motion is so high, however, that heat flow about 10 times as great as that presently observed would be required to initiate convective currents. Therefore, convection covering the entire mantle seems doubtful. No satisfactory theory based on convection yet formulated explains the Earth's thermal constitution.

Metamorphism of rocks. Metamorphism is the structural and mineralogical change of solid rocks caused by pressure and temperature conditions differing greatly from those prevailing at the time of their formation. Rock texture may be changed either by recrystallization or by alteration of the rock's chemical composition through the formation of new minerals.

Several types of metamorphism exist; those in which heat plays a dominant role are contact metamorphism, regional metamorphism, and plutonic metamorphism. Contact metamorphism is caused when changes are induced in rocks by the thermal effect of intrusive igenous rocks (rocks formed deep within the Earth's crust by cooling molten material). Regional metamorphism is caused by a combination of the thermal effects and pressures involved in orogeny (the forming of mountains). This is the most important type of metamorphism and is effective over areas of hundreds of square kilometres. Plutonic metamorphism is associated with great depths, great hydrostatic (equal in all directions) pressures, and extremely high temperatures.

Mountain building and formation of ore deposits. According to the contraction theory of Earth formation, as the molten planet cooled and the crust solidified, heat loss from the interior caused contraction and compressional stresses in the crust. The stresses produced folding and belts of fold mountains in weak spots in the crust. According to the continental-drift theory, however, all the continents were once joined together in one large protocontinent. When the protocontinent was broken up and the continental blocks migrated toward their present positions, the friction generated by their motion caused them to buckle, forming fold mountains. Finally, according to the convection-current theory, convection currents are responsible for folding of the Earth's crust. The more recent concept of plate tectonics holds that the Earth's crust and upper mantle consist of six major segments of plates that abut and intersect on the surface. Mountain systems result from the collisions at plate boundaries, which are also associated with earthquakes, magmatism, and heat flow. Ore bodies are bodies of rock containing a concentration of one or more chemical elements sufficient to allow profitable mining. Development of ore bodies occurs near or in igneous rock intrusions, often during mountain building.

Plate tectonics and heat flow

Volcanism and geothermal power. Volcanism is the geological process involving the activities of magma and lava. The high heat-flow readings obtained in the mid-oceanic crests are probably related to recent volcanism. Generally, the energy discharge associated with volcanic eruptions is transported mostly in the form of the thermal energy of the ejected material. The energy transported by the motion of the ejected material is not of significantly large magnitude, however. An estimation of the size of the Earth's crust can be used to gauge the share of heat flow caused by volcanism; this amounts to only about 1 percent of the Earth's total heat flow. Useful energy can be extracted from naturally occurring hot water and steam associated with volcanism. Power is obtained by releasing the hot water or steam through drill holes, then channelling it through pipes to a turbine-generator unit. Several analyses have been made of the cost of geothermal power; each concluded that it is indeed competitive with conventional means of electric-power generation. The first large-scale modern use of natural heat sources began in Italy in 1904, and some 50,000 homes in Iceland are heated by natural sources. Several other countries (notably Japan, the Soviet Union, and New Zealand) have similar geothermal-power operations, and exploration for more sources of geothermal power continues throughout the world (see further GEYSERS AND FUMAROLES).

BIBLIOGRAPHY. The subject of the Earth's heat flow is in a constant state of flux due to the relative recency of many of the data. Consequently, most references are contained within a rather specialized literature (*e.g.*, the *Journal of Geophysical Research*) and only a few textbooks or general works can be cited here. The interested reader should consult P.J. HART (ed.), *The Earth's Crust and Upper Mantle* (1969); W.H.K. LEE (ed.), *Terrestrial Heat Flow* (1965); and JOHN VERHOOGEN *et al.*, *The Earth* (1970). The first two works cited contain many additional references of relevance; the last presents a general treatment of many of the topics presented in this article.

(A.L.La.)

Earth, Magnetic Field of

The Earth is surrounded by a magnetic field, within which it behaves as if it were a ball of magnetized iron with north and south magnetic poles. The field is bigger than those of Venus and Mars; Mercury possibly has none; and only Jupiter has been shown to possess, like Earth, an appreciably intense dipolar magnetic field.

The discovery of the magnetic field of the Earth was dependent in part on the fact that a freely suspended lodestone (the mineral magnetite or maghemite, an iron oxide), which is naturally magnetized, seeks to align itself in a northerly direction everywhere on the Earth. Thus, the lodestone was man's earliest magnetic compass. In the 13th century a French engineer, Petrus Peregrinus, suspended an iron needle over various parts of a lodestone; the needle responded in different directions, which, when plotted as lines, converged to points on opposite ends of the lodestone. These points he called poles. In his *Epistola de magnete* he wrote about the attraction of the poles of heaven upon the poles of lodestone. Peregrinus also described compasses in which a magnetized needle attached to a stick was floated on water or was mounted on a pivot. In the Middle Ages, these man-made magnetic compasses were widely used by mariners for navigation, and it was known in the first half of the 16th century that a needle of a magnetic compass does not point exactly toward the North Pole. An English scientist, William Gilbert, summarizing all existing knowledge about magnetism in 1600 when he dis-

Naming of the magnetic poles

pelled the idea that lodestone and compass needles are attracted by the Pole Star or some other object in the heavens. From an experiment with a spherically shaped piece of lodestone, he concluded that the entire Earth is a great magnet with its two magnetic poles situated in regions near the geographical poles (Figure 1).

In the sense of exact science, however, a full understanding of the distribution of the Earth's magnetic field was not achieved until a German physicist, Carl Friedrich Gauss, published "Allgemeine Theorie des Erdmagnetismus" in 1838, in which his mathematical analysis showed that more than 95 percent of the Earth's magnetic field originates within the Earth's interior, and only a small remaining portion comes from outside sources.

Theoretical studies on possible physical mechanisms of an inside source of the Earth's magnetic field have developed only recently. Scientists agree, however, that the inside source is not one or more large aggregates of magnetic material but an electric-current system that most likely exists within the Earth's core and is caused by the flow of electric charges. Modern theoretical treatments of the origin of the Earth's magnetic field are concerned with possible mechanisms of generating and maintaining electric currents within the core, where there is a

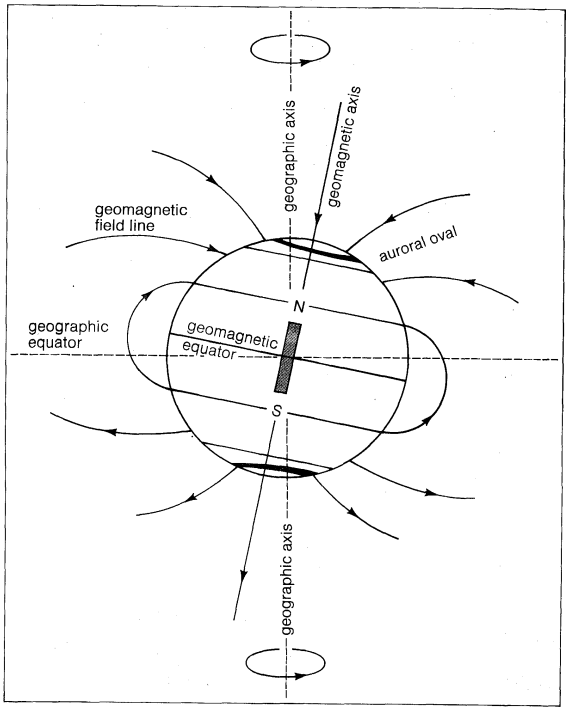


Figure 1: Distribution of the geomagnetic field in the space surrounding the Earth in terms of a hypothetical bar magnet, labelled N and S to correspond to the north and south geomagnetic poles, respectively.

circulation of material. It is well known that the Earth's magnetic field is gradually changing with time in its intensity as well as in its distribution pattern. Paleomagnetism, which deals with the behaviour of the Earth's magnetic field through geological time, has shown that the Earth's magnetic field has been frequently reversed. Any theory of an internal source, therefore, must be consistent with agelong variations of the Earth's magnetic field during historic, prehistoric, and geologic time.

Magnetism produced in ionosphere

On the other hand, the fraction of the Earth's magnetic field produced by outside sources is now understood to be an important representation of the electromagnetic activities in the Earth's upper atmosphere. Balfour Stewart, a Scottish physicist, first proposed that the daily varying part of the Earth's magnetic field can be ascribed to electric currents flowing in the Earth's upper atmosphere. The currents flowing in the atmosphere were identified later as the ionosphere by Kristian Birkeland, a Norwegian physicist, who concluded that electrically charged particles impinging on the polar upper atmosphere give rise to displays of auroras and to simultaneous magnetic storms. Until recently, outer space beyond the Earth's atmosphere was believed to be an almost perfect vacuum. Recent measurements of the physical effects on vehicles in outer space, however, have clearly indicated that the solar corona extends beyond the Earth's orbit, the space between the Sun and the Earth being filled with plasma gas (an ionized gas that, because it consists of equal numbers of positive ions—nuclei of atoms, mostly hydrogen—and electrons, is electrically neutral), called the solar wind. Although the particle density of the solar-wind plasma is only five to ten particles per cubic centimetre (one cubic centimetre equals 0.061 cubic inch), the macroscopic effect of the plasma on the Earth's magnetic field is still significant because of the very large spatial scale of phenomena concerned.

During a period of low solar activity (quiet Sun), the solar wind flows radially outward from the Sun at speeds of 300–500 kilometres (186–311 miles) per second. Electrons and positive ions in the solar wind are repelled by the magnetic field of the Earth as they approach the Earth until they reach a region where the kinetic pressure of the solar wind and the pressure of the Earth's magnetic field are in equilibrium. Near the Earth the magnetic field is compressed and neutralized in an outward direction, and thus the magnetic field is confined within a limited space surrounding the Earth. The limited space and its outer boundary are called the magnetosphere and the magnetopause, respectively, of the Earth (Figure 2).

Because the solar wind blows outward from the Sun, the magnetosphere on the dark side of the Earth is elongated along the direction of solar wind, the elongated shape resembling the tail of a comet. The tail of the Earth's magnetosphere, called the geomagnetic tail, has been observed to extend over a distance greater than 2,000 times the Earth's radius.

Although the Earth's magnetic field prevents charged particles of the solar wind from penetrating into the magnetosphere, some of them do, nevertheless, diffuse into it from the magnetopause. In some cases, electrically neutral particles (such as the neutron component of cosmic rays), which can get into the magnetosphere because they are not influenced by magnetic fields, decay into electrons and positive ions within the magnetosphere. These charged particles (mostly electrons and protons) are trapped by the magnetic field and are confined to the magnetosphere. A circular belt of these trapped charged particles surrounding the Earth is called a radiation belt or a Van Allen belt.

Van Allen belts

Though the magnetic field of the Earth, generally called the geomagnetic field, has been long valued as a direction indicator, it is now realized that the magnetic field plays a much more important role for man as the protector of the Earth from a continuing lethal attack of solar charged particles.

This article treats of the Earth's main magnetic field, its

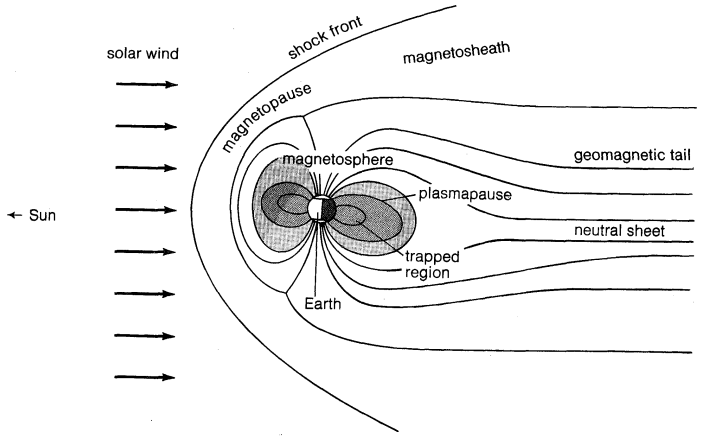


Figure 2: Distribution of the geomagnetic field in space surrounding Earth (see text).

transient and short-term fields, including telluric currents (those electric currents on or near the Earth's surface), and includes theories of the origin of the geomagnetic field. For further information on the Earth's interior, see EARTH, STRUCTURE AND COMPOSITION OF; for a discussion of paleomagnetism, see ROCK MAGNETISM; and for relevant information on the interaction of the Earth's magnetic field with the solar wind and allied atmospheric phenomena, see AURORAS; IONOSPHERE; VAN ALLEN RADIATION BELTS. See also COMPASS; MAGNETISM.

EARTH'S MAIN MAGNETIC FIELD

Elements of the geomagnetic field. The intensity, or strength, of a magnetic field at a given point is defined as the force exerted on a unit north pole positioned at that point. Fields are sometimes called field lines, because the number of lines per unit area and their direction pictorially represent the intensity and direction of the magnetic field. The geomagnetic-field strength, or intensity, at any point on the Earth's surface is customarily specified by means of the rectangular components, X, Y, Z, of the total intensity, F, defined as follows: X is the component along the horizontal northward direction, Y is the horizontal eastward component; Z is the downward component, all in the geographical coordinates, as illustrated in Figure 3. Another common way of specifying the geomagnetic intensity is by means of the magnetic elements, symbolized by H, D, and I, defined as follows: H is the magnitude of the horizontal component, considered to be positive whatever its direction and called the horizontal intensity; D is the azimuth of the horizontal intensity, reckoned positively from the geographic north toward the east and called the magnetic declination, or, by mariners, the magnetic variation, of the compass; I is the angle made by the direction of the total magnetic intensity with the horizontal, reckoned positive when the direction of the intensity inclines downward and called the magnetic dip, or the magnetic inclination. A horizontal line along the magnetic horizonal intensity, H, is called the magnetic meridian. The components of the intensity (X, Y, and Z) and the magnetic elements (D, I, and H) have a simple trigonometric relationship, as shown in Figure 3.

The same specification of magnetic elements can be used for describing the vector quantity (magnitude and direction) of the geomagnetic intensity at a position not too much above the Earth's surface that is observed from an aircraft or orbiting satellite at low altitude; say, 100–300 kilometres (60–190 miles). To specify the geomagnetic intensity observed at a point far from the Earth—for example, on a space vehicle just inside the magnetopause—another set of coordinates, referring to the ecliptic plane of the Earth (plane of the Earth's orbit), has been frequently used. The intensity components, designated as X', Y', Z', are defined as follows: X' is the Sunward (*i.e.*, directed toward the Sun) component, Y' is the component perpendicular to the direction of X' lying in the ecliptic plane, and Z' is the northward component.

From S. Chapman and J. Bartels, *Geomagnetism*, vol. 1, p. 2; Clarendon Press, Oxford

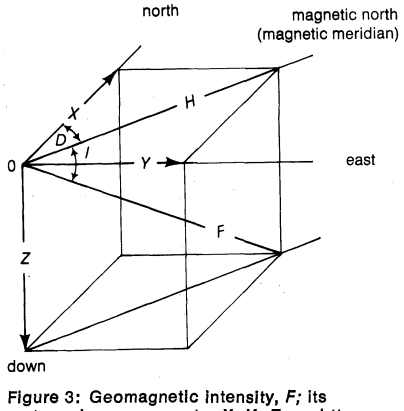


Figure 3: Geomagnetic intensity, *F*; its rectangular components, *X*, *Y*, *Z*; and the magnetic elements *H*, *D*, *I*. The magnetic meridian coincides with the direction *H* (see text).

World magnetic surveys and magnetic observations. A three-dimensional vector of the geomagnetic intensity at any point can be determined by measuring any three magnetic elements independently of one another. In the case of magnetic surveys on land or sea, a combination of measurements of the two angles, magnetic declination (D) and inclination (I), and one intensity component, such as the magnetic horizontal intensity (H), the total geomagnetic intensity (F), or the downward component (Z) of the geomagnetic intensity, has been generally adopted. The two angles are measured because an angle can be determined to a much higher accuracy than can a magnetic-intensity component. Recent progress in atomic and nuclear physics, however, has resulted in great improvement in the precise measurement of weak magnetic intensities, and, as a result, a combination of three rectangular magnetic-intensity components is now frequently used in determining the geomagnetic intensity; in particular, on board space vehicles.

Magnetic observations

At the present time, the geomagnetic intensity is being measured by magnetic surveys at a great number of points on land, over oceans, in the air, and in outer space. Over the whole surface of the Earth there are about 140 permanent magnetic observatories at which certain combinations of three magnetic elements are continuously recorded; and it is significant that the value of the geomagnetic field is observed to change appreciably with time. Magnetic fields are measured in gauss units (a field of one gauss is produced at a distance of one centimetre [0.39 inch] from a straight wire in which a current of five amperes—3×10^{19} electrons per second—is flowing; for comparison, the field in the vicinity of a magnetized needle is a fraction of a gauss, whereas the field between the poles of a horseshoe magnet may be hundreds or thousands of gauss). The total geomagnetic force at the geographical Equator and near the poles of the Earth amounts to approximately 0.3 and 0.7 gauss, respectively. The total geomagnetic force, however, changes even within a day's time, with a range of 0.0002 gauss at the Equator and 0.0005 gauss at the poles. One hundred-thousandth of one gauss, called a gamma ($1\gamma = 0.00001$ gauss), has been adopted as the unit for specifying the strength of geomagnetic forces. It may be stated, therefore, that the geomagnetic intensity has a daily variation of about 20 gammas and 50 gammas in amplitude, respectively, near the Equator and the poles, even during the period of a quiet Sun. In addition to the temporal variations, such as the daily variation, the seasonal variation, and sporadic disturbances, the annual mean of the geomagnetic force is subjected to a nonperiodic secular variation, amounting to several gammas per year. To represent the world distribution of any geomagnetic element on a chart, therefore, the temporal variations must be eliminated, and the value of the geomagnetic element concerned must be reduced to a certain common period of time (such as a year) by reference to the secular variation chart of the element. Figure 4 illustrates a world magnetic chart of the magnetic declination (D) on the Earth's surface in 1965.

At the magnetic poles, which are defined as the points on the Earth's surface where the geomagnetic intensity is vertically directed, there is no horizontal intensity, and consequently the magnetic declination is indeterminable. The north magnetic pole, where the geomagnetic intensity is exactly downward (*i.e.*, the magnetic inclination, or dip, $I = +90°$) is located in northern Canada; the south magnetic pole, in which $I = -90°$, is located almost diametrically opposite near the Antarctic coast at about 140° east in longitude. The horizontal component of intensity (H) reaches its maximum magnitude of about 0.3 gauss (3×10^4 gammas) at a magnetic inclination of $I = 0$ (roughly along the geographical Equator) and vanishes at the two magnetic poles. The vertical magnetic intensity (Z) takes maximum positive and negative values of about 0.7 gauss in magnitude at the north and south magnetic poles, respectively, whereas it becomes zero along the line of zero magnetic inclination ($I = 0$). The

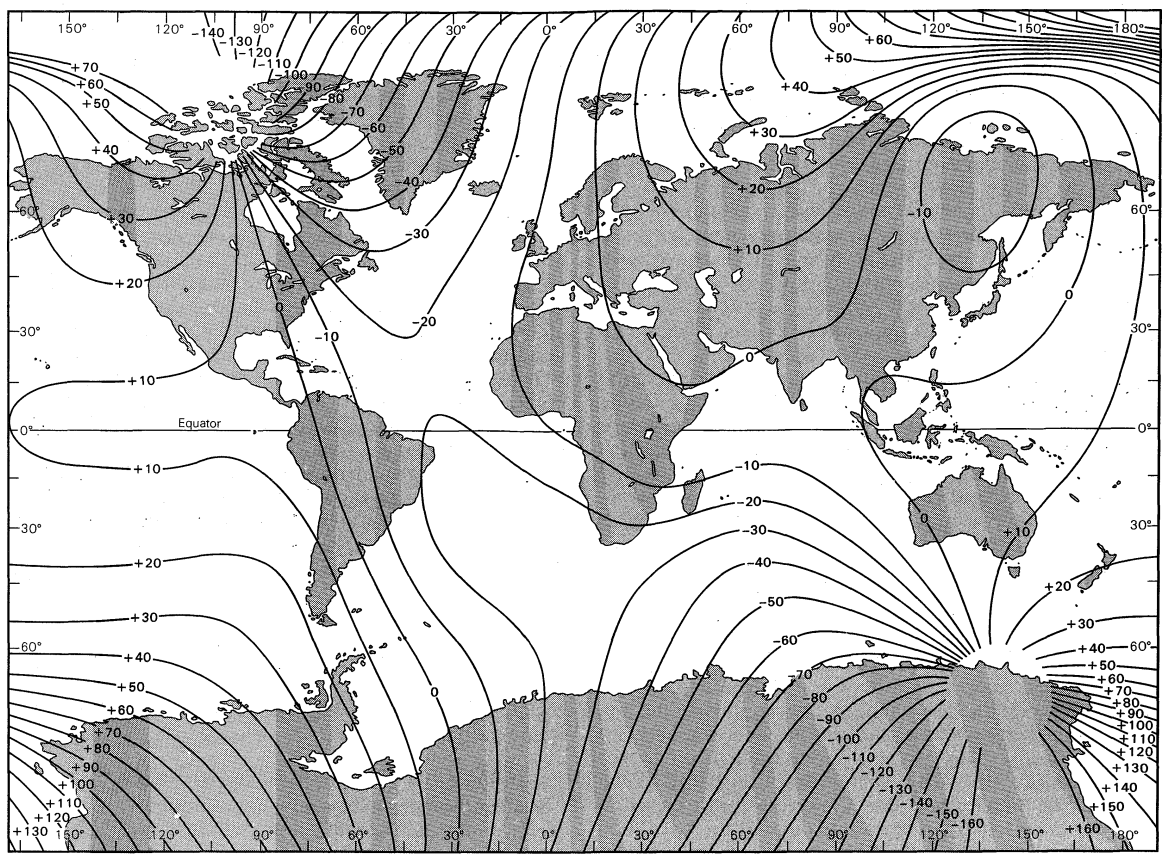


Figure 4: The geomagnetic declination in 1965. Numerals represent the geomagnetic declination in degrees.
By courtesy of National Aeronautics and Space Administration

line of $I = 0$ surrounding the Earth is called the magnetic equator. World charts showing the distribution of the magnetic elements D, H, Z, F, and I on the Earth's surface are compiled for every five-year epoch.

If it is assumed that there is a bar magnet at the Earth's centre as shown in Figure 1, in which the axis of the bar magnet is inclined about 11° of arc from the Earth's geographical axis so that a pole of the magnet is directed toward northern Canada, then the distribution of all magnetic elements of the magnetic field caused by the hypothetical magnet will resemble those of the real geomagnetic field, at least on and near the Earth's surface. In this case, the magnetic moment (a constant of any bar magnet is its magnetic moment, the product of distance between its north and south poles and the pole strength) of the hypothetical magnet corresponding to the geomagnetic field at the present time would be 8×10^{25} gauss-centimetres^3. Such a hypothetical bar magnet is called the centred geomagnetic dipole. The poles of the centred geomagnetic dipole are exactly opposite one another on the surface of the Earth and are called the geomagnetic poles. The location of the geomagnetic north pole in 1965 was 78.5° north in latitude and 69.1° west in longitude. A great circle along which the vertical (Z) intensity component of the hypothetical magnetic field is zero is called the geomagnetic equator. The great circle of the geomagnetic equator deviates somewhat from the magnetic equator, which itself is not a great circle.

Geomagnetic poles and equator

Although the geomagnetic field can be approximately represented by the magnetic field of the centred geomagnetic dipole, there still remain residual magnetic fields when the centred dipole field is subtracted from the real geomagnetic field. The residual fields are called the geomagnetic regional anomalies. It has been demonstrated that these regional anomalies can be accounted for by clockwise and counterclockwise vortices of electric currents, which are hypothetically located at the Earth's core beneath the centres of regional anomalies.

Surveys for the geomagnetic-field distribution within the magnetosphere by means of space-vehicle magnetometers have not yet been completed; nevertheless, the general feature of the configuration of lines of the geomagnetic field, summarized from a large number of observations, is illustrated in Figure 2. Attention must be paid, in the Figure, to the observed fact that the geomagnetic field in the sunlit side is markedly compressed by the solar wind, whereas that in the dark side is markedly elongated, and that such a difference or a change in the field-line configuration is dependent upon the local time of the field-line meridian and not upon the location of the field line on the Earth. In accordance with the rotation of the Earth, the outer lines of geomagnetic intensity that have their roots on the Earth's polar regions are compacted within the sunlit side of the magnetosphere but are dragged far away behind the Earth as they enter into the night-side space; these outer lines of geomagnetic intensity are, thus, periodically subjected to a vast daily change in configuration.

The inner lines of force, which have their roots in lower latitudes of the Earth, on the contrary, are always closed, regardless of whether they are on the sunlit side or the night side, although their shapes are changed to a certain extent: they can be considered as stable lines of intensity. It is evident that charged particles, such as electrons and protons, trapped by these stable inner lines of force, can survive much longer than those that interact with the changing outer lines. This trapping is regarded to be the main cause for the Van Allen radiation belts, which consist chiefly of a high population of electrons and protons.

Variations in the main magnetic field. It has been known for many years that the geomagnetic field is subjected to a secular variation, gradually varying with a rate of several gammas per year for the magnetic-intensity elements. During the last 150 years, since the measurements of the magnetic elements of the Earth were first initiated, the magnetic moment of the geomagnetic dipole has been continuously decreasing; for instance, it is now 8×10^{25} gauss-centimetres^3, but in 1838 it was 8.6×10^{25} gauss-centimetres^3. It also has been noticed that some centres of the regional anomalies in the geomag-

netic field are drifting westward with an average speed of about 0.2° of arc per year. This observed characteristic of the geomagnetic secular variation is known as the westward drift of the geomagnetic regional anomalies. An increase or a decrease of the strength is more markedly observed, however, for some other regional anomalies. Each regional anomaly seems to be drifting westward with either growth or decay of its strength.

Knowledge of the geomagnetic secular variation can go back further into prehistoric or geologic times with the aid of rock magnetism. The science that deals with the geomagnetic field in these distant times is called paleomagnetism; a similar science that deals specifically with geomagnetism in prehistoric times is called archaeomagnetism.

Rock magnetism and paleomagnetism

In archaeomagnetism, the direction and intensity of the geomagnetic field at a certain site at a certain past time can be determined by measuring the direction and intensity of magnetization of igneous rocks or baked earths, provided that these materials have maintained the initial magnetization (called remanent magnetizations) they acquired when they were formed in the presence of the old, but now changed, geomagnetic field. The chief results of archaeomagnetic studies for the past 8,000-year period are as follows: the distribution mode of the geomagnetic field during this period is substantially the same as its present mode; that is to say, the geomagnetic field during the period also can be represented approximately by that of the present geomagnetic dipole. The moment of the geomagnetic dipole has changed slowly during the period, having been 1.5 times the present value about 2,000 years ago and 0.8 times the present value about 4,000 years ago. The observed decrease in the intensity of the geomagnetic dipole for the last 150 years is considered to be a continuation of its decrease during the last 2,000 years. On the other hand, the secular variations of declination (D) and inclination (I) during the last 8,000 years seem to imply that the westward drift of regional anomalies has continued with a speed of 0.2°–0.4° per year during the period also.

The same method of determining archaeological geomagnetic fields by use of remanent magnetization of natural rocks can be applied to much older geologic times. The results of paleomagnetic research by means of a large number of rocks collected from various continents indicate that the geomagnetic field in geologic times also can be substantially approximated by the geomagnetic dipole field. In dealing with the geomagnetic dipole field in the last 100,000,000 years, it is found that the geomagnetic north pole was located within two circular areas of about 30° in radius either at the geographic North Pole or the geographic South Pole; seldom was the geomagnetic north pole located in the temperate latitudes between the two polar areas. The geomagnetic field in geologic times, in other words, was approximately the same as it is at present, or almost perfectly reversed with respect to the Equator, and the time interval for a reversal of the field must have been small compared with the time during which the geomagnetic-field polarity was either normal, as it is at present, or reversed.

The phenomena of the geomagnetic-field reversals for the last 6,000,000 years have been thoroughly studied. During this geologic period, the average lifetime of a geomagnetic-field polarity, either normal or reverse, was 230,000 years, but individual lifetimes varied appreciably and may be ascertained to a close approximation by probability theory.

The magnetic moment of the geomagnetic dipole for the same geologic period is found to be $(9.1 \pm 3.3) \times 10^{25}$ gauss-centimetres^3 (the figure after the plus-or-minus sign represents the standard deviation) for both the normal and the reverse polarities. It may be concluded, therefore, that the geomagnetic field took either the normal or the reverse polarity, which now appears to be in a quasi-steady state, having an average magnetic moment of about 9×10^{25} gauss-centimetres^3. In several instances, the time interval for transition between the two polarities has been estimated to be less than 10,000 years.

The geomagnetic field in remote geologic times is more complicated, and there is a hypothesis associated with it called polar wandering. The age of the oldest geomagnetic field ever estimated from terrestrial rocks goes back 2.5×10^9 years, and the geomagnetic field has, therefore, been subjected to repeated polarity reversals for at least 2.5×10^9 years. The average moment of the geomagnetic dipole throughout this period remained at the same approximate magnitude as at present. The position of the geomagnetic pole, however, was not located within the geographical polar regions 10^9 years ago and before but was near the geographical Equator. It seems likely that the geomagnetic-pole position has gradually moved from the equatorial zone to the polar region during the past 10^9 years, regardless of whether the geomagnetic field had a normal or a reverse polarity.

Polar wandering

There is a problem related to the polar-wandering hypothesis: Although the conclusion that the geomagnetic-pole position has moved from the equatorial zone to the polar region is supported by experimental data derived from rock samples of different continents, the routes of the polar wandering according to data from different continents are not alike. The route of the polar wandering derived from North American data, for instance, is systematically different from that derived from European data. If it be assumed, however, that there was no Atlantic Ocean between North America and Europe 100,000,000 years ago and that the Atlantic Ocean floor has continued to spread since then, the discrepancy between the two routes becomes understandable, and it is in part for this reason that the hypothesis of continental drift has been reintroduced.

TRANSIENT AND SHORT-TERM GEOMAGNETIC FIELDS

Solar and lunar daily magnetic variations. It has been observed that the geomagnetic field at any point on the Earth's surface is subjected to a quasi-regular daily variation of from ten to 30 gammas. The vertical intensity decreases around local midday in the Northern Hemisphere, whereas it increases around midday in the Southern Hemisphere. The declination takes the maximum and minimum values in the local mornings and afternoons, respectively, in the Northern Hemisphere, whereas daily variation of the declination is reversed in the Southern Hemisphere. Variations in the geomagnetic field are small in the dark time compared with those in the sunlit time. In the local summer, the daily variation is enhanced; it is reduced in the local winter. It thus becomes evident that the geomagnetic daily variation is primarily controlled by the position of the Sun at the observing site.

Applying Gauss's method of analysis (*i.e.*, dividing a magnetic field into outside and inside sources), it is found that two-thirds of the daily varying component of the geomagnetic field is caused by a source outside the Earth's surface, while the remaining one-third is caused by a source inside. The location of the outside origin has been determined by measuring the vertical distribution of the daily varying field with sounding rockets at various sites over the Earth. The results have indicated that the outside source is caused by horizontal electric currents flowing in a region between 105 kilometres and 130 kilometres (65 and 81 miles) in altitude; namely, through the lower part of the Earth's ionosphere. As expected, the intensity of the electric current in the lower ionosphere is observable only in the sunlit time, and the current intensity in the dark time is unobservably weak. The total electric current in the sunlit time is of barely sufficient magnitude to account for the outside-origin part of the daily geomagnetic variation observed on the ground.

To account for the inside source, which contributes one-third of the daily geomagnetic variation, it is assumed that the primary outside source produces a flow of electric current in the Earth's electrically conductive interior by a process of electromagnetic induction. This, in turn, generates a magnetic field that varies in intensity and frequency with the primary outside source and is consistent with Gauss's mathematical representation.

The horizontal wind in the upper atmosphere has been observed directly by the use of sounding rockets and indirectly by observing, for example, the flow of ionized

clouds (clouds consisting of positively and negatively charged particles) generated by incoming meteors. The results of such measurements have shown that a strong wind speed of from 50 to 100 metres (164 to 328 feet) per second is always blowing in an observed altitude range of 80–300 kilometres (50–186 miles) and the direction of wind is subjected to a fairly systematic daily variation. When an air mass containing charged particles moves in the presence of a magnetic field, an electric field is created in the air mass; it is perpendicular to the direction of motion and to the magnetic intensity and is represented by the vector product $E = v \times F$, in which E is the electric-field intensity, v is the velocity, and F is the magnetic-field intensity. If the air mass is electrically conductive, with electric conductivity (σ), then an electric current (j) will flow through the air. The flow of current will be equal to the product of the conductivity and the field intensity (E), namely, $j = \sigma E$. The electric current produces a magnetic field, and, in accordance with changes in the intensity and direction of the electric current, the resultant magnetic field also should change. The system consisting of the parameters (v, E, and σ) is usually called the ionospheric dynamo.

Ionospheric dynamo

In the sunlit time, the neutral atoms and molecules in the ionosphere are partly ionized into electrons and positive ions by the solar ultraviolet radiations and X-rays, which have high ionizing energy, and, as a result, the ionized part of the ionosphere becomes electrically conductive. During dark time, there is scarcely any ionization of the ionospheric particles; instead, the electrons and ions ionized during time of sunlight tend to recombine with each other to become neutral particles once more; thus, the electrical conductivity of the ionosphere in the dark time is much reduced compared with its sunlit-time magnitude.

Because the ionospheric wind is caused mostly by the tidal force of the Sun, there should not be an appreciable difference in its velocity between the day and night. The electrical conductivity of the ionosphere, however, is much smaller in the dark time than in the sunlit time, and this condition can be regarded as the principal reason why most of the daily variation of the geomagnetic field, caused by the product of conductivity and velocity, takes place in the daytime.

Effects of the Moon

The Moon's gravitational force, on the other hand, also gives rise to a tidal current in the ionosphere, just as it does in the ocean water, and consequently a lunar ionospheric dynamo also exists. If the time variation in the geomagnetic field is carefully analyzed, therefore, a periodic change during the lunar semidiurnal period is clearly observed. Evidently, the lunar geomagnetic variation is attributable to a lunar-tidal electric-current system in the ionosphere, which is caused by the lunar-tidal current, and the resultant electric current is the product of the electric conductivity of the air mass times the vector product of its velocity and the magnetic-field intensity, or $j_L = \sigma(v_L \times F)$, in which j_L is the electrical current due to the lunar-tidal current, σ is the conductivity, v_L is the velocity of the lunar-tidal current of air, and F is the magnetic intensity in the ionosphere. It is obvious that the lunar geomagnetic variation also is appreciable only on the sunlit hemisphere, where the ionospheric electric conductivity is high. In general, the lunar geomagnetic variation has a much smaller amplitude than the solar daily one, being only about one or two gammas in magnitude.

Magnetic storms and substorms. Occasionally geomagnetic disturbances of large amplitude take place, in particular when the Sun is active. The behaviour of such a geomagnetic disturbance observed in temperate latitudes is characterized by a sudden increase in the horizontal intensity, followed several hours later by its sharp depression of 50–200 gammas and a gradual recovery of the depression at a later stage. Such a geomagnetic disturbance is called a magnetic storm. In such a case, a satellite outside the magnetopause always records a marked increase in the speed of the solar wind or an increase of the number density of plasma particles in the solar wind. A satellite located near the magnetopause on the sunlit side will record that parts of the magnetosphere become compressed and contracted. The magnetopause is located at a distance of ten to 12 Earth radii from the Earth during an ordinary, quiet period but will become contracted on the Sunward side to a distance of five to seven Earth radii in case of a magnetic storm. It is evident, then, that a geomagnetic storm is primarily caused by a sudden increase of the solar-wind kinetic energy, which in turn is induced by solar activity. The increased kinetic (gaslike) pressure of the solar wind compresses the magnetosphere toward the Earth more so than ordinarily, resulting in an initial sudden increase of the geomagnetic field. Several hours after the initial increase in the geomagnetic field, it is observed that plasma particles at a distance of four to six Earth radii from the Earth that are within the magnetosphere begin to move so as to give rise to westward moving electric currents (composed of protons), whereas the electrons flow in an eastward direction. The magnetic field caused by their motion is mainly responsible for an observed marked depression of the geomagnetic field, called the main phase of a magnetic storm. The westward electric current of protons, called the equatorial-ring current, is maintained by the increased solar-wind energy, because the compressed magnetospheric plasma particles are continuously influenced in their motion by the geomagnetic field, which is increasingly more intensive toward the Earth. When the kinetic energy of the solar wind recovers to the ordinary state, the maintenance of the equatorial-ring current ceases, and the magnetic storm begins to go into its gradual recovery phase.

In regard to a detailed physical mechanism for the generation of equatorial-ring currents, no unanimous conclusion has yet been reached among specialists. Comprehensive experimental and theoretical studies on this problem are being carried out on the basis of numerous observations taken on satellites and on the ground.

When a magnetic storm is in progress, tremendous disturbances of various aspects are observed in the Earth's north and south polar regions. The most fantastic of these disturbances in the polar regions is the display of a phenomenon called an aurora (*q.v.*). Auroras are seen mostly in special zones surrounding the north and south geomagnetic poles, lying within an angular distance of 25°–30°. These zones are specifically called the auroral zones (see Figure 2). Recent studies of the distribution of auroral displays have indicated that auroral zones are not exactly circular in form with a geomagnetic pole at the centre but rather are of an oval shape, and, for this reason, the term auroral oval is sometimes used.

Auroral zones

Simultaneously with an auroral display, a sharp and substantial change in the geomagnetic field is observed in and near the auroral ovals, its magnitude often exceeding 1,000 gammas. The duration of such a sharp geomagnetic disturbance varies in time from 30 minutes to three hours and is called a polar magnetic substorm. Because the shape of a change in the geomagnetic horizontal force with time in the case of a polar magnetic substorm has a depression or bay shape, polar magnetic substorms are often called geomagnetic bays. There is a high degree of correlation between the auroral display and the magnetic substorm with respect to time of duration, place of appearance, and magnitude.

A polar magnetic substorm is accompanied not only by an auroral display but also by other, simultaneous electromagnetic phenomena, such as an unusual increase in the density of electrons in the ionosphere, an emission of X-rays from above, an emission of very low frequency radio waves, and so forth. All of these simultaneous phenomena indicate that a shower of electrons is impinging along the geomagnetic-intensity lines from outside toward the auroral-oval ionosphere. From these observed data, the energy of impinging electrons has been estimated to be between 10^8 and 10^{10} electron volts (some laboratory devices can accelerate electrons to these energies). One electron volt is the energy that a charged particle such as an electron acquires when it is accelerated by an electric potential of one volt, which means that electrons can penetrate into the Earth's atmosphere down to 80–120 kilometres (50–75 miles) in altitude. The

source of a polar magnetic substorm can thus be identified as strong electric currents that flow through the auroral-oval ionosphere and become anomalously conductive because of the ionization of upper-atmosphere gas by impinging electrons. There is a question as to why and how the electromotive force that drives the electric currents is produced in the auroral oval ionosphere. Some clarification has been achieved from satellite-borne observations. The following appear significant: The polar magnetic substorms occur mostly at night and frequently at times around local midnight. In the night side within the magnetosphere, electrons and protons tend to drift eastward and westward, respectively, in the outside of the plasmapause (see Figure 2), implying that plasma flows, moving inward from the geomagnetic tail into the stronger geomagnetic field, drift in accordance with a law that states that the force acting on a charged particle is in a direction at right angles to the direction of motion of the particle and the direction of the magnetic field within which it is moving. The force acts in one direction for a positive charge and directly opposite for a negative charge. In vector notation, one has the cross product of the velocity and field vectors, or $ev \times F$, in which e denotes a positive electric charge for a proton or a negative electric charge for an electron, v denotes the velocity vector, and F denotes the magnetic-field intensity vector.

Just before a breakup of a substorm accompanied by an auroral display, the number densities of electrons and protons in the neutral sheet (see Figure 2) of the geomagnetic tail suddenly decrease. It is likely that these particles are pushed forward in a direction toward the Earth, with the result that the population of electrons and protons in the space outside the plasmapause may suddenly increase.

One possible explanation for the occurrence of substorms is as follows: As the result of a certain trigger action, plasma gas in the night-side magnetosphere flows toward the plasmapause, where electrons and protons separately drift in opposite directions, producing two regions of electrostatic charge, negative in the dawn side and positive in the dusk side; these regions may then accelerate charged particles of the plasma toward the Earth, because the charged particles can move only along the lines of geomagnetic force. A closed electric circuit is thus formed, starting from the positively charged region in the magnetosphere, along geomagnetic-field lines and the auroral-oval ionosphere of high conductivity, and coming back to the negatively charged region along lines of the geomagnetic field. In brief, a breakup of a substorm accompanied by an auroral display and other, simultaneous phenomena can be regarded as an electric discharge along lines of the geomagnetic field between the magnetosphere and ionosphere. The discovery of the trigger phenomenon that produces such an electric discharge is one of the goals of the main research projects of magnetospheric physics. Studies so far suggest that the magnetic field concomitant with the solar wind plays a principal role in the trigger action. The observed magnetic field in the solar wind amounts to five to ten gammas in intensity, and its direction is sometimes approximately northward and at other times approximately southward. The solar-wind magnetic field is due to electric-current vortices within the wind mass, which were generated when the wind moved away from the Sun's magnetic field. When the solar-wind magnetic field is directed northward, it is repelled by the geomagnetic field of the magnetopause over the northern and southern polar regions of the Earth. But, on the contrary, when the solar-wind magnetic field is directed southward, a part of the field will be connected with the geomagnetic field of the same region, because the directions of both magnetic fields there are parallel and therefore connectable. The solar-wind magnetic field and the geomagnetic field are thus sometimes repelled away from each other and at some other times connected with each other. It is a statistical fact that a polar substorm takes place when the magnetic field of the solar wind turns southward.

The trigger phenomenon

If the foregoing theory of the physical mechanism of polar magnetic substorms is true, then a polar magnetic substorm accompanied by an auroral display and other related phenomena ought to occur simultaneously, both in the northern and southern polar areas, which has been corroborated so far by observation. The simultaneous occurrence of polar magnetic substorms is called the geomagnetic conjugate relationship.

Electromagnetic induction and telluric currents. All kinds of transient variations in the geomagnetic field observed on the Earth's surface are caused by the changing of the outside and inside sources.

As was pointed out earlier in relation to the daily geomagnetic variation, the inside source of field variations affects magnetic storms, substorms, and other transient magnetic variations, and they are attributable to the transient electric currents that are electromagnetically induced within the Earth's interior by the primary magnetic-field variations of the outside sources. The magnitude of the inside-source component is about half that of the outside one in most cases, and the variation of the former is systematically subjected to that of the latter in accordance with Maxwell's equations for electromagnetic induction. Experimentally, the potential difference between two separate points on the Earth's surface has been measured at various places. The potential differences thus measured, the so-called electric fields, consist of a steady component and a transient variation component. The steady component is caused by geological conditions that form a certain kind of electric cell owing to local differences in the electrochemical compositions of materials. The transient variation component, expressed in terms of an electric field, is approximately related to the geomagnetic field, as given by Maxwell's equation for electromagnetic induction.

In actual cases, electric currents are caused by the electric field within the Earth (even the uppermost part of the Earth has a finite electric conductivity). Thus, the relationship between the electric and geomagnetic fields is modified by an effect of the induced current. If the electrical conductivity is sufficiently small, as in the case of igneous rock, for example, then the electric current can be expressed as the product of the conductivity and electric field, and Maxwell's relation between the electric and geomagnetic fields holds approximately. The electric currents thus induced within the Earth are generally called telluric currents.

Telluric currents

With an increase in temperature, the electrical conductivity within the Earth increases sharply with increased depth. Analyses of transient geomagnetic variations of various periods in terms of the electromagnetic induction within the Earth have shown that the distribution of the Earth's electric conductivity is dependent on the geocentric distance. The results of such studies have shown that the average electric resistance is about 10^6 ohms per centimetre from the surface down to 400 kilometres (250 miles), where the resistance decreases sharply to about 10^3 ohms per centimetre; it also gradually decreases to 0.1 ohm per centimetre near the surface of the Earth's core.

The lateral distribution of electric conductivity near the Earth's surface, on the other hand, is not uniform at all but varies from place to place, even within a continent. This heterogeneous distribution of the conductivity results in anomalies of the induced currents in the uppermost part of the Earth, particularly for geomagnetic variations of short period.

ORIGIN OF THE GEOMAGNETIC FIELD

Main geomagnetic field and its secular variation. The origin of the main geomagnetic field, presumed to exist within the Earth, has long been a great mystery in the Earth sciences. It is evident that the Earth cannot be a permanent-type magnet because of the high temperature of the Earth's interior, and, therefore, electric currents that are responsible for the main geomagnetic field, with its secular variation, must be maintained somewhere inside. There is almost no possibility that these currents flow through the Earth's solid mantle, which extends between the Earth's crust and core, whose outer boundary is about 2,900 kilometres (1,800 miles) beneath the

Earth's surface. The electrical conductivity of the mantle is so low, being 10^{-3} to 10 (ohm-centimetres)$^{-1}$, that the maintenance of necessary electric currents would require an electrical-generator system of fantastically high power. It has long been believed, therefore, that a system equivalent to a powerful electric generator and consisting of a circulating electrically conductive metallic fluid may exist within the Earth's core.

Any theoretical model of the main geomagnetic field must be able to account for its frequent reversals in geologic times as well as its general behaviour at present. No theory has ever been proposed to account quantitatively for the observed various aspects. A number of hypotheses explain the origin of the geomagnetic field, but many of them are unable to account for even the present geomagnetic field. A U.S. geophysicist, Walter M. Elsasser, in 1939 proposed that a thermoelectric electromotive force in the core might give rise to the electric current responsible for the geomagnetic field. He assumed that some thermal convections are maintained within the core by the energy supply from radioactive substances and that the rising and sinking parts of the core's fluid caused by the convections may correspond to the hot and cold junctions of a hypothetical thermojunction system, provided that the core material changes with its radius. The thermoelectric force that might thus be produced could drive the electric currents within the core. Quantitative examination of this possibility, however, has indicated that the hypothesis is hardly acceptable because of the small magnitude of the electric current. Another possibility was suggested in 1954—namely, that a stray magnetic field accidently impressed within the Earth could be amplified by the Hall effect (in the Hall effect, a secondary electric current, and, hence, magnetic field, is produced when an electric current moves at right angles to the magnetic-field lines) of the lower mantle near the core's boundary. Although this Hall-effect hypothesis cannot be ruled out in principle, no favourable evidence has ever been given to support it.

Twisted-kink theory

In 1950 a possible mechanism to maintain the geomagnetic field, now called the twisted-kink theory, was proposed. It is based essentially on the magnetohydrodynamics of a highly conductive core. In this model, bundles of magnetic-field lines are assumed to be present as stray magnetic fields in the Earth's core, and the magnetic field corresponds, giving rise to the dipole-like magnetic field outside the core. Because the fluid within the core may be subjected to an appreciable differential rotation and the fluid is highly conductive, the bundles of field lines may be dragged and stretched by the fluid motion, and some of the bundles would become unstable under the strong twisting motion, eventually forming loops (kinkings) of field lines. Once such loops of magnetic-field lines are formed in the differentially rotating fluid, some of them could be separated from the original field lines, and the separated ones could be stretched out by the fluid motion to become approximately parallel to the original bundles of field lines. Thus, these magnetohydrodynamic processes could result in an enhancement of the original stray field by changing the fluid-motion energy into the magnetic-field energy. The instability of a twisted bundle of magnetic-field lines has been mathematically studied. It is not clear, however, whether these processes are sufficiently regenerative. The twisted-kink theory is acceptable, in a qualitative way, for explaining maintenance of the present geomagnetic field, but its quantitative aspects have not yet been worked out.

Another theoretical model to account for the origin of the geomagnetic field is the so-called dynamo theory, which is based essentially on electromagnetic induction processes within a moving and highly conductive fluid of the core, and is akin to the twisted-kink theory. Because the dynamo theory is pertinent to both the maintenance of the geomagnetic field and its secular variation behaviour, a discussion of it in some detail follows.

Dynamo theory of the geomagnetic field. The basic idea of the dynamo theory was advanced in a paper on the magnetic field produced by a rotating sphere. This notion of a self-exciting dynamo was later re-examined in terms of a disk dynamo, consisting of the following elements: a conducting disk rotating about its axle and a solenoid (a coil of wire in the shape of a cylinder; an electric current passing along the wire will produce a magnetic field parallel to the coil's axis) concentric with the axle in which one terminal contacts the rim, or periphery, of the disk and the other terminal contacts the axle. When the disk rotates about the axle in the presence of a magnetic field parallel to the axle, a radial electric field is induced in the disk between its centre and periphery, and the generated electromotive force between the periphery and centre will drive an electric current through the solenoid. The solenoid will produce a magnetic field, thus enhancing the initial magnetic field. The decrease in energy of the disk is thus transformed into an increase of magnetic-field intensity. An experimental demonstration of the disk dynamo may be impossible to achieve under laboratory conditions, inasmuch as a disk of ten centimetres (3.9 inches) in radius in a circuit of ten ohms electric resistance would have to rotate about 10,000,000 times per second to reach a steady state of no loss in energy and no gain in magnetic field. This situation does not mean, however, that the self-exciting dynamo mechanism is impossible for a rotating body on the scale of the Earth and the Sun. The self-exciting dynamo mechanism is based on the principle of repeated self-excitations of the magnetic field by the addition of a magnetic field generated by the electric current of the dynamo: an analogous mechanism of an electrically conductive rotating disk and solenoid could be present in the Earth's core, where the fluid core materials are perhaps circulated by heat sources or other mechanical forces.

Use of the disk dynamo

Of the various possible patterns of electric-current systems within the Earth's core, one may consider the possibility that two or more dynamos are coupled together. A model of such a two-disk dynamo was proposed by a Japanese geophysicist, Tsuneji Rikitake, in 1958, in which the electric current induced in a rotating disk flows through a solenoid to enhance the magnetic field affecting another rotating disk, and similarly the electric current induced in the second disk enhances the magnetic field for the first. In this two-disk dynamo system, the electric currents in the two solenoids, as well as the angular velocities of the disks, can oscillate, and the currents can even reverse for appropriate initial conditions, whereas, in the single-disk dynamo, the electric current can periodically change on an appropriate initial condition, but it can never reverse.

The models of the single disk or the coupled two disks demonstrate possible processes of electromagnetic induction within the moving conductive substances. In Earth's highly conductive core, on the other hand, there can exist two different types of electric-current systems and their associated magnetic fields. One type is the poloidal magnetic field, which extends from the inside of the core into the outside, its lines of magnetic force making a closed circuit through regions inside and outside the core. Such a poloidal magnetic field can be produced, for example, by a solenoidal-type electric current flowing around the core's rotational axis. The other type is the toroidal magnetic field, which makes a closed circuit of each line of the magnetic field only within the core's boundary. Such a toroidal magnetic field can be produced, for example, in a doughnut-shaped solenoid, within which all lines of magnetic field are necessarily confined. Even if a toroidal magnetic field were present within the Earth's core, it could not be directly observed from its outside because of electromagnetic laws. The geomagnetic dipole field must, therefore, be caused by a poloidal field if it originates within the Earth's core. It may be supposed that a dynamo process caused by fluid motion in the poloidal dipole field $\boldsymbol{B}_0$ generates a toroidal field $\boldsymbol{T}_1$, a similar dynamo process in $\boldsymbol{T}_1$, generating another toroidal field $\boldsymbol{T}_2$, and such a chain process of generating toroidal fields proceeds until a toroidal field $\boldsymbol{T}_i$ is generated. As has been said, these toroidal fields, $\boldsymbol{T}_1$, $\boldsymbol{T}_2$, etc., are not observable as such on the Earth's surface. If, however, a dynamo process of a fluid motion in a toroidal field $\boldsymbol{T}_i$ can produce a poloidal dipole field $\boldsymbol{B}_0$, then a sort of a closed

Poloidal and toroidal fields

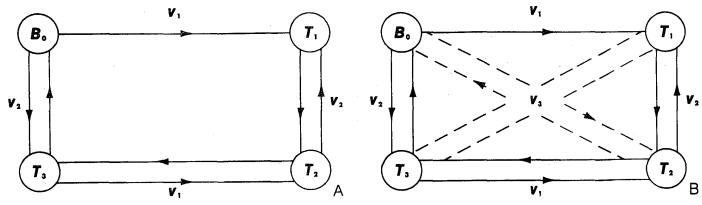


Figure 5: *Four-dynamo theory.*
(A) Chain reactions of dynamo model. (B) Modified dynamo model. See text for explanation of symbols.

chain of dynamo processes starting from B_0 and returning to B_0 again through toroidal field T_1, T_2,———T_1 can be set up.

In 1954 a consistent mathematical model of a limited number of involved dynamo processes was derived in which the fluid motion in the core is assumed to consist of two modes; that is, a zonal flow around the Earth's rotation axis (v_1) and a symmetric convective motion composed of four convection cells (v_2) whose longitudinal symmetricity is expressed as the cosine of 2λ, in which λ denotes the longitude, these motions being confined to the core's interior. The dynamo process between a poloidal dipole field B_0 and the zonal flow v_1 results in a toroidal magnetic field T_1, which is axially symmetric about the rotation axis. The dynamo process between T_1 and v_2 generates a toroidal field T_2 whose longitudinal symmetricity is expressed by the cosine of 2λ. The dynamo action between T_2 and v_1 generates a toroidal field T_3 whose longitudinal symmetricity is given by the sine of 2λ. Finally, the dynamo action between T_3 and v_2 results in a poloidal field of the dipole-like mode B_0, and thus a chain of dynamo processes, $B_0 \rightarrow T_1 \rightarrow T_2 \rightarrow T_3 \rightarrow B_0$, can be apparently completed as far as these four main modes of magnetic field are concerned. It must be noted that each dynamo process generates not only the main mode but also a number of higher harmonic modes of magnetic field having a higher degree of symmetricity as well.

The other possible dynamo processes, between B_0 and v_2, between T_2 and v_2 and between T_3 and v_1, generate, respectively, T_3, T_1, and T_2 types of magnetic field as their main results and associated higher harmonic mode fields in addition. The dynamo process between T_1 and v_1, however, generates only higher harmonic mode fields, and it does not generate a poloidal field of a dipole-like mode. If all the higher harmonic fields are truncated, therefore, the whole picture of the chain reactions of the four dynamo processes may be summarized as that shown in Figure 5A. It has been shown that a system consisting of four dynamos is very likely to have a steady state, insofar as the higher harmonic fields whose degree of symmetricity with respect to lambda are larger than two can be truncated (elimination of higher terms). Assuming that the electrical conductivity of the core is uniformly 3×10^2 (ohm-centimetres)$^{-1}$, the maximum speed of the zonal flow (v_1) has been estimated at about 0.04 centimetres per second. It has been pointed out that the maximum intensity of the toroidal field T_1 may be as large as 480 gauss, whereas the geomagnetic dipole field of 0.7 gauss at the magnetic pole should require only 4.5 gauss intensity at the magnetic pole of the core. Later studies on the possible effect of higher harmonic fields, which were truncated in the earlier model, have indicated that the contributions of these higher harmonic fields to the whole chain system is not negligible, and so it becomes doubtful whether or not the model really does simulate a steady dynamo. Quantitatively, it has been shown that the powers of the four main dynamos are reduced by taking into account the higher harmonic fields, whereas other higher harmonic fields actually increase when still higher harmonic fields are considered. In other words, the energy supplied by fluid motions is not spent mostly in maintaining the closed circuit of the four main dynamos but goes instead to support higher harmonic fields.

Four-dynamo theory

Although the model faces a serious difficulty, workers have shown that certain kinds of dynamo can be maintained in a spherical conductor. In 1958 it was shown that an axisymmetric toroidal magnetic field (T_1), which is generated from an axisymmetric poloidal field (B_0) with a certain appropriate flow in a conductive sphere, can enhance the initial poloidal field by coupling with another appropriate motion, in which all higher harmonic fields are suppressed as minor effects. In brief, it was proved that a dipole-type poloidal magnetic field can be maintained only if an appropriate pattern of fluid motions exists in the Earth's core.

In the same year, it was shown that a dynamo can consist of two rigid conductive spheres embedded in a rigid spherical conductor. It is certain that the self-exciting dynamo system can be maintained if appropriate conditions are satisfied within the earth's core; unfortunately, however, there is no evidence to justify the necessary conditions for the dynamo models in the actual makeup of the Earth's core.

In 1964 a new and important criterion was proposed in regard to the dynamo model, which is based on a realistic flow-pattern model in the core. In the original model, two simple modes of flow were assumed; namely, a zonal flow, which is attributable to a difference of the angular velocity between the fluid core and the solid mantle, and a simple convection pattern, which could be maintained by a possible radiogenic heat (heat due to radioactivity) source in the core in which the longitudinal symmetricity of the convection pattern is expressed as the cosine of two lambda. It was now suggested that a dynamo of this type can hold only when the convection pattern is not as simple as in the original model but consists of two modes whose longitudinal symmetricities are given by both cosine of two lambda and sine of two lambda.

In 1970 another worker, who had worked extensively on a similar model of the modified dynamo, showed that the modified dynamo can reach its steady state even though the higher harmonic fields are taken into account. In this model, a convection flow pattern (v_3) of sine of two lambda in longitudinal symmetricity is newly added to v_1 and v_2 in the original model, in which the v_3 convection is confined to a smaller spherical space whose radius is half that of the core. Thus, the modified four dynamo processes may be summarized as shown in Figure 5B.

In the extensive numerical calculation required to examine this dynamo system, higher harmonic fields up to cosine 6λ and sine 6λ in their longitudinal symmetricities are taken into account. It has been proved that this modified system can maintain the dynamo action and the whole system can reach its steady state. In the modified system also, the intensities of toroidal magnetic fields are large compared with those of poloidal fields of the same degree of symmetricity. In particular, the T_1-T_2-T_3-T_1 triangle of interactions through v_1, v_2, and v_3 plays the most important key role in the maintainance of the whole system. It may be concluded, therefore, that the dipole field B_0 corresponds to the geomagnetic dipole field, and other higher harmonic poloidal fields represent the geomagnetic regional anomalies.

Regional anomalies

Theory of secular variation. The most drastic characteristic of the geomagnetic secular variation may be the reversals of the geomagnetic field that have frequently occurred in past geologic times. The possible mechanism of this phenomenon has not yet been fully interpreted. One model of the geomagnetic field indicates that the dipole field B_0 is a minor field component compared with the major toroidal fields and that its direction can be easily reversed, just as can the perturbation effects of the

main interaction between the major toroidal components. It seems likely, therefore, that the geomagnetic field reversals are caused by such perturbations in the major toroidal field system. This presumption is in agreement with the observed fact that the lifetimes of individual geomagnetic polarities obey statistical laws. In examining the transient processes of the system, a worker in 1970 found that the time required to reach a final steady state from an initial state of no toroidal fields in the system is about 5,000 years. This conclusion seems to favour the estimated time interval for a polarity transition.

Another well-established characteristic of secular variation is the westward drift of geomagnetic regional anomalies. The westward drift phenomenon has been interpreted on the basis of the dynamo theory. If there are convection currents of matter in the core, the conservation of angular momentum requires that the deep interior of the core rotates faster than the outer layer. This means that the outer layer should move westward with respect to the inner part. Furthermore, if the deeper part of the mantle is of sufficient electrical conductivity to result in an electromagnetic coupling between the core and mantle, then the core as a whole and the deeper part of the mantle are tightly connected together. To reach a steady state under these conditions, the angular velocity of eastward rotation is the largest in the deep core, the smallest in the core's outer layer, and in the mantle the intermediate value between these two. Thus, the outermost part of the core apparently drifts westward with respect to the solid mantle and consequently with respect to the Earth's surface. It has been shown that an electric conductivity of one (ohm-centimetre)$^{-1}$ for the deeper mantle is sufficient to cause a westward drift of 0.2° per year of the uppermost layer of the core relative to the mantle. The magnitude of conductivity seems to be reasonable based on other evidence in regard to the Earth's interior.

Growth and decay of the geomagnetic regional anomalies can be interpreted as being caused simply by growth and decay of higher harmonic poloidal fields in the core in the course of transient dynamo processes. It may be concluded, therefore, that all significant aspects of the geomagnetic secular variation can be reasonably understood on the basis of the dynamo theory.

The origin of the geomagnetic field and its secular variation may be reasonably interpreted when they are assumed to be caused by an electromagnetic dynamo action going on in the Earth's core. It must be remarked, however, that the dynamo theory is not yet complete; the dynamo theory so far examined is the kinematic theory, in which the reaction of electromagnetic coupling upon the hydrodynamical fluid motions is ignored. In a complete magnetohydrodynamic (motion of fluids produced by electromagnetic forces) theory, both the electromagnetic coupling and the hydrodynamical motions and their mutual reactions must be simultaneously considered in the form of associated equations of motion. It is practically impossible at present, however, to perform theoretical researches along this line, because even the world's largest computer is barely sufficient to carry out the necessary computations for such work, and then only in a limited way.

BIBLIOGRAPHY. S. CHAPMAN and J. BARTELS, *Geomagnetism*, 2 vol. (1940), a comprehensive classical textbook of geomagnetism; S. MATSUSHITA and W.H. CAMPBELL (eds.), *Physics of Geomagnetic Phenomena*, 2 vol. (1967–68), a modern text dealing with almost all aspects of geomagnetic phenomena; G. FANSELAU (ed.), *Geomagnetismus und Aeronomie*, vol. 3 (1959), a text dealing with details of geomagnetic phenomena within the solid earth; E. IRVING, *Paleomagnetism and Its Application to Geological and Geophysical Problems* (1964), a comprehensive collection of paleomagnetic data up to 1963; T. RIKITAKE, *Electromagnetism and the Earth's Interior* (1966), a mathematical treatment of various geomagnetic phenomena within the solid earth.

(T.Na.)

Earth, Mechanical Properties of

Mechanics is concerned with the relationship between matter, the forces acting on it, both internally and externally, and the consequent motion of it in its parts or as a whole. Kinds of motion may be translational; that is, in a straight line, vibratory, and rotational. These interrelations are called mechanical properties, which may be invested in the smallest of particles, such as nuclei of atoms, to the largest aggregates of matter—the Earth, Sun, stars, and galaxies.

By knowing how the Earth can deform and recover throughout its volume and by being aware of such fundamental properties as density it is possible to derive useful information on gross phenomena—mountain building, earthquakes, continental drift, internal and external gravitational fields, sources of the Earth's internal heat, and properties of the Earth's core.

The five basic mechanical properties

There are five mechanical properties of the Earth's body that are particularly fundamental to a determination of its behaviour: namely, density, pressure, gravitational intensity, incompressibility, and rigidity. Density refers to mass per unit volume, which varies within the Earth because of the effects of pressure and temperature and because of variations of composition. Pressure refers to the force per unit area inside a body, and incompressibility indicates the extent to which a material resists pressure. Rigidity indicates the resistance of a material to the stresses that tend to distort it, and gravitational intensity is the force per unit mass arising from a gravitational field. These five properties go together as a group, the values of which are derived by an interrelated series of calculations from observational data, and they are by far the most reliably determined mechanical properties of the Earth's interior. Their values provide a firm basis for the investigation of other properties.

A large part of the observational evidence comes from a study of the Earth's response to various external and internal forces, and of the dynamical interplay between the Earth and outside bodies such as the Moon and artificial satellites. For example, the attractions of the Moon and Sun give rise to tides, including bodily tides of the solid Earth, that throw light on the Earth's internal rigidity; observations of the precession of the equinoxes (*i.e.*, a slow change in direction made by the Earth's axis), a consequence of Earth–Moon interaction, contribute toward determining the Earth's moment of inertia (see below for definition) and internal density distribution; observations of artificial satellites contribute to knowledge of the Earth's external gravity field and thence also to the density determination. Examples of internal forces include those associated with earthquakes (*q.v.*), which send mechanical waves throughout the Earth's interior and deliver important numerical information on three of the properties—density, incompressibility, and rigidity at internal points.

In interpreting the observations, the essential aim, as in all science, is to arrive at optimum mathematical model representations of the properties of interest. Scientific method requires models to be the simplest, among the infinity of those possible, that are compatible with the observational evidence on hand at any stage. As evidence grows, models are progressively refined. Models may take the form of mathematical relations, as in equation (1) below, or numerical tables.

Mathematical models

To a close approximation, the Earth is spherically symmetrical, so that the internal distributions of the five mechanical properties mentioned above are closely represented by Earth models, the properties of which vary only with the depth below the surface or the distance from the centre. The deviations from spherical symmetry are important in some problems, however. Reference is made later in this article to the Earth's mechanical behaviour over periods that are comparable with geological time. This is a topic that does not yet compare in precision with the determination of the basic mechanical properties.

The solid Earth (including the oceans) has a mean radius of 6,371 kilometres (3,960 miles) and consists of a crust with an average thickness of about 15 kilometres (about 10 miles), a mantle extending from the crust to about 2,900 kilometres (1,800 miles) in depth, and a core extending from there to the centre (Figure 1). Subdivisions

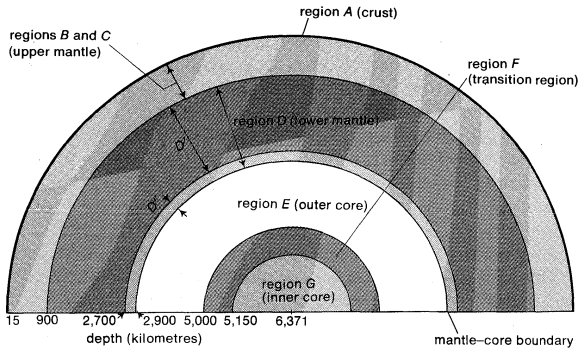


Figure 1: Cross-section of the Earth, showing the principal internal layers.

of the mantle and core will be introduced later. This article is chiefly concerned with mechanical properties below the crust and their large-scale variation throughout the Earth's interior.

PRELIMINARIES ON DEFORMABLE MEDIA

Stress and strain. The Earth consists of deformable matter capable of responding to internal and external applied forces. The forces generate internal stresses that are accompanied by deformation, or strain. Stress may be defined as the force per unit area across an indefinitely small area inside a body, and strain is the proportional change in shape brought about by a stress (see DEFORMATION AND FLOW; and ELASTICITY). In general, six independent components are needed to express the stress at any point. If orthogonal (Cartesian) axes Ox_1, Ox_2, Ox_3 are taken, the six components are commonly expressed in condensed notation as p_{ij}, in which the subscripts i and j take on values 1 through 3—*e.g.*, $ij = 11$, 22, 33, 23, 31, 12. The subscript i refers to the axis Ox_i that is perpendicular to the element of area acted on by the force, and the subscript j refers to the axis Ox_j, to which the component of force is parallel; p_{12}, for example, denotes the component parallel to Ox_2 of the stress, or force per unit area, across a small plane internal interface perpendicular to Ox_1.

The deformation is commonly represented in terms of six strain components e_{ij} (the subscripts i and j take values 1, 2, and 3, as with p_{ij}). Strain relates to the proportionate changes of distances between neighbouring particles when internal movement occurs in a body. The dilatation, denoted by the Greek letter θ (theta), is equal to $e_{11} + e_{22} + e_{33}$, is minus the compression, and indicates the proportionate increase in volume.

Simple stress–strain behaviour. A first task in studying quantitatively the response of a body to applied forces is to select a mathematical model employing stress–strain relations suited to the demands of the observational context. One commonly used set of relations is a generalization of Hooke's law (after an English scientist, Robert Hooke) of elasticity (a perfectly elastic medium is one in which the identical initial form will be assumed after the force of deformation is removed) stating that strain is proportional to stress. Hooke's law relates to the simple case of wires that are stretched under tension; the extension is proportional to the tension.

In the generalized law, the stress (p_{ij}) can be expressed as the sum of two terms, the first being the difference between the incompressibility, also called bulk modulus (k), and two-thirds the rigidity (symbolized by the Greek letter mu, μ), times the dilatation (θ) times a factor δ_{ij}, and the second being twice the rigidity times the strain component (e_{ij})—*i.e.*,

$$p_{ij} = (k - \tfrac{2}{3}\mu)\theta\delta_{ij} + 2\mu e_{ij}, \qquad (1)$$

in which the factor δ_{ij} is equal to 1 or 0 according as $i = j$ or $i \neq j$. The six relations resulting from (1) are the simplest serviceable stress–strain relations for a solid; they relate to the ideal case of a perfectly elastic, isotropic solid undergoing "infinitesimal" strain. The particular case for which rigidity equals zero ($\mu = 0$) corresponds to the ideal perfect fluid, for which p_{ij} always takes the form $-p\delta_{ij}$, in which p is the pressure. Inside solids, p is, strictly, minus one-third the sum of the principal stresses, $-\tfrac{1}{3}(p_{11} + p_{22} + p_{33})$. The stress inside a solid may take the form $-p\delta_{ij}$ as a particular case and is then called hydrostatic.

In contexts in which the model relations (1) are closely compatible with the observations, materials may be classified into "solid" and "fluid" without misunderstanding —solid when the ratio of rigidity to incompressibility (μ/k) is found to be significant (say, about 1 percent, or greater), fluid when the ratio is insignificant.

Deviatoric stress and strength. The stress components (p_{ij}) can be expressed as minus the product of the pressure (p) and the factor δ_{ij}, plus another term (P_{ij}) designated as the deviatoric stress components: $p_{ij} = -p\delta_{ij} + P_{ij}$ (2). When a material is subjected to increasing stress, a stage is reached at which the initial configuration is not even approximately recovered after removal of the stress (see ROCK DEFORMATION; ROCKS, PHYSICAL PROPERTIES OF). For ordinary short-term purposes, the strength of the material is suitably defined as the value of $\sqrt{(3\Sigma\Sigma P_{ij})^2}$ when this stage is just reached. Fracture or flow may subsequently occur. No strengths greater than 10^{10} dynes (a force producing an acceleration of one centimetre per second per second when acting on a mass of one gram) per square centimetre (abbreviated dyne/cm^2) are known, and this value is not likely to be exceeded in the Earth's interior. As seen from the Table, pressure (p) reaches 10^{10} dynes per square centimetre at about 35 kilometres (22 miles) depth and increases continuously with depth, exceeding 3×10^{12} dynes per square centimetre at the Earth's centre. (The pressure of water at the ocean's greatest depths is about 10^9 dynes per square centimetre). Hence, throughout most of the Earth, especially below the outermost few hundred kilometres, all the deviatoric stress components are small compared with the mean of the principal stresses so that $p_{ij} \approx -p\delta_{ij}$. Thus, the stress can be represented as hydrostatic to good approximation inside most of the Earth, even where the rigidity is substantial.

Seismic waves. Forces inside the Earth cause deviatoric stresses to build up to the stage at which the local strength is reached. When the stresses are relieved by fracture, an earthquake occurs (see EARTHQUAKES). Seismic waves are transmitted from the earthquake focus (origin).

Observational checking has shown that the model relations expressed by (1) can be used to good approximation in investigating these waves. The quantities p_{ij}, e_{ij}, and θ in (1) here relate to the comparatively small additional stresses and strains that are superposed, during the wave transmission, on the larger, approximately hydrostatic stresses and compressions already present in the undisturbed state.

***P* and *S* seismic waves**

Seismic waves include two bodily types that pass down into the Earth's interior: the *P* (primary) waves resemble sound waves and cause the material particles to vibrate in the direction of wave advance; the *S* (secondary) waves cause particles to vibrate transversely; that is, at right angles to the direction of wave advance (see WAVE MOTION). The speed α of primary waves is given to good accuracy by the relation, speed squared is equal to density (denoted by the Greek letter rho, ρ) divided into the sum of incompressibility and four-thirds the rigidity, and the square of the speed β for transverse waves is given by the ratio of rigidity to density; *i.e.*,

$$\alpha^2 = (k + \tfrac{4}{3}\mu)/\rho \qquad (3a)$$

$$\beta^2 = \mu/\rho. \qquad (3b)$$

Solid regions transmit both *P* and *S* waves; fluid regions transmit only *P* waves, the rigidity μ and hence the *S*-wave velocity being negligible. Seismic surface waves travel around the Earth's surface with amplitudes that die out more or less rapidly with depth. Earthquakes also excite free oscillations of the Earth in which large proportions of the whole Earth may be vibrating at one and the same time.

Observations of all these waves and oscillations have been applied in working out the Earth's internal mechanical structure.

Deviations from simple stress–strain behaviour. The model stress–strain relations given by (1) are not adequate for all Earth contexts. For example, they do not give a suitable description of deformations over geological time. When these relations are inadequate, the terms solid and fluid cease to be precisely defined terms, and it becomes better in general to classify material in other ways. Deep in the Earth, where compressions are large, the infinitesimal theory has for some purposes to be replaced by finite-strain theory, in which the stress–strain relations are not linear as in the relations in (1). When accumulated deviatoric stresses are relieved by flow or fracture, further theory again is required. The various complications, which are far from being fully resolved, bear especially on questions of the long-term mechanical behaviour of the Earth and are considered below in the section *Anelasticity in the Earth*.

THE BASIC INTERNAL MECHANICAL PROPERTIES

The key item in determining the main mechanical properties of the Earth in its present state is the determination of the distribution of the density (ρ). The distributions of the properties pressure (p), incompressibility (k), rigidity (μ), and gravitational intensity (g), are then readily derived.

Mean density and the moment of inertia of the Earth. The most precisely determined constraint on possible density distributions in the Earth is the mean or average density, estimated as 5.5 grams per cubic centimetre by the English physicist Henry Cavendish in 1798 (abbreviated g/cm^3) and now assessed as 5.517 g/cm^3.

A second powerful constraint on possible density distributions is the moment of inertia, an index of the degree of concentration of denser matter in a body toward its centre. For a sphere of uniform density and of mass M and radius R, the moment of inertia (I) is 0.4 MR^2. A value for the Earth's mean moment of inertia became available in the 19th century when it was calculated by French mathematicians, who, starting in 1743, combined a variety of observational evidence using ingenious mathematical steps. The evidence included numerical data on the universal constant of gravitation (G), the Earth's mean radius (R), mass (M), axial angular velocity, and on the precession of the equinoxes. The mathematical steps involved an analysis of the Earth's internal and external gravitational fields and theory on the Earth's ellipticity of figure (see EARTH, GRAVITATIONAL FIELD OF). Stresses inside the Earth were treated as hydrostatic. Initially, this was because geologists had thought the Earth to be wholly fluid below a thin crust. Although the Earth is now known to be solid (in the ordinary sense) for five-sixths of its volume, the hydrostatic theory can still be justified, up to a point, because of the relative smallness of the deviatoric stresses.

Moment of inertia

The 19th-century estimate of the moment of inertia was 0.3336MR^2. Because the coefficient 0.3336 is appreciably less than the value 0.4, which applies to a sphere of constant density, it followed that the deep interior of the Earth is much denser than the outer part. Along with studies of meteorites, this led to the suggestion that the Earth contains a fairly dense core. In 1963 the Earth's moment of inertia was calculated independently of the internal hydrostatic theory from artificial satellite observations. A reduced value of 0.3308MR^2 was obtained, implying that the deviatoric stresses inside the Earth's mantle are somewhat more significant than had earlier been supposed.

When the value of the moment of inertia first became fairly reliably known, provisional models of the Earth's density variation were formulated. Two such models gave the density increasing steadily from two to three grams per cubic centimetre at the surface to 10 to 12 grams per cubic centimetre at the centre. These values were accepted until the 1930s. Early in this century, simple models allowing for a core also began to appear; these models give approximately the same density range.

Data from seismic bodily waves. Analyses of records of bodily seismic waves have yielded tables giving the travel times of P and S waves along seismic rays in terms of the angular distances covered. Subsequent mathematical processes have yielded values of the P and S velocities α and β throughout the mantle and of α throughout the core. Through the equations (2) and (3), the values of α and β supply values of the physical quantities k/ρ and μ/ρ throughout the mantle and of $(k + \frac{4}{3}\mu)/\rho$ throughout the core. This is important information that must be fitted by the internal distributions of density, incompressibility, and rigidity.

Evidence from natural earthquakes is supplemented from analyses of seismic waves from artificial explosions. Chemical explosions supply information on structure near the Earth's surface and nuclear explosions on the whole interior. Although always smaller in energy than the larger natural earthquakes, explosions have a cardinal advantage in investigating the Earth's interior in that their locations and times of origin are knowable in advance and to high precision.

Explosions of nuclear devices have been important in another way; efforts to detect them have resulted in the establishment of seismological array stations and the application of the mathematical theory of communication in gathering and interpreting the records. An array station in Montana consists of 525 substations, all equipped with similar seismographs, spread over 25,000 square kilometres (10,000 square miles). This development is increasing the precision with which the primary and secondary wave, velocities α and β, and hence the ratios k/ρ and μ/ρ can be determined in the Earth's interior.

Seismograph array stations

In 1906 the British scientist R.D. Oldham showed from analysis of P-wave data that the Earth has a core. In 1909, Andrija Mohorovičić, a Croatian seismologist, gave the first evidence that the wave velocities α and β increase sharply across the boundary, now called the Mohorovičić discontinuity, which separates the crust from the mantle. The detection of S waves throughout the mantle about this time showed that the entire mantle behaves as a solid in its response to stresses produced in seismic waves.

A remarkable calculation by the German-born geophysicist, Beno Gutenberg, in 1914, placed the mantle–core boundary at 2,900 kilometres (1,800 miles) in depth. In 1936 the Danish seismologist Inge Lehmann showed that the P-wave velocity increases fairly sharply deep down in the core, requiring the core to be divided into a (so-called) outer core and an inner core (of radius about 1,200 kilometres [700 miles]).

By 1939, fairly reliable travel-time tables (here called the JB tables) of bodily seismic waves had been evolved by the geophysicists Sir Harold Jeffreys and K.E. Bullen, and by Beno Gutenberg and an American seismologist, C.F. Richter, leading to much further detail. Among other things, Jeffreys calculated distributions of the velocities α and β corresponding to the JB tables, estimated the core radius as 2,898 $\pm$ 3 kilometres (1,800 $\pm$ 2 miles), and gave evidence of a transition zone separating the outer core proper from the inner core. Over a period of years starting from 1926, Gutenberg put forward evidence of a low velocity layer (characterized by decreases of α and β with depth) over a limited range of depth inside the outermost 200 kilometres (150 miles) of the mantle.

Bodily wave data gathered since 1939 have suggested or required several modifications to the JB travel-time tables and the distributions of α and β. The modifications are mostly small and are taken into account in the accompanying Table. Recent bodily wave data have also contributed important evidence on lateral variations of structure—*i.e.*, differences in structure extending down to some 400 kilometres depth below different geographical regions.

Data from seismic surface waves. Starting from the application of an early theory of two English physicists, Lord Rayleigh (William Strutt) and A.E.H. Love, a large number of investigators (notably, Jeffreys, Robert Stone-

Representative Values of the Earth's Mechanical Properties

region	depth below Earth's surface (z) (kilometres)	pressure (p) (10^{12} dynes/cm^2)	density (ρ) (g/cm^3)	gravitational intensity (g) (cm/sec^2)	incompressibility (k) (10^{12} dynes/cm^2)	rigidity (μ) (10^{12} dynes/cm^2)	Young's modulus (E) (10^{12} dynes/cm^2)	Poisson's ratio (σ) (10^{12} dynes/cm^2)
A	0	0.000	...	982	(0.65)	(0.36)	(0.91)	0.267
	15	0.004	3.31	983	1.03	0.71	1.73	0.223
B	350	0.12	3.54	994	1.74	0.71	1.88	0.320
C	650	0.23	4.25	998	2.76	1.43	3.66	0.279
	1,000	0.39	4.52	994	3.48	1.83	4.67	0.276
	1,500	0.62	4.80	993	4.28	2.14	5.51	0.285
D′	2,000	0.87	5.06	1,000	5.04	2.43	6.28	0.292
D	2,500	1.13	5.32	1,030	5.82	2.73	7.09	0.298
	2,700	1.24	5.42	1,050	6.17	2.86	7.40	0.300
D″	2,890	1.35	5.64	1,070	6.53	3.00	7.80	0.300
	2,890	1.35	9.91	1,070	6.5	0	0	0.5
	3,000	1.47	10.08	1,050	6.9	0	0	0.5
E	3,500	1.99	10.79	930	8.5	0	0	0.5
	4,000	2.47	11.35	790	10.3	0	0	0.5
	4,500	2.88	11.79	650	11.8	0	0	0.5
F	5,000	3.23	12.56	490	13.2	0	0	0.5
	5,160	3.33	12.70	430	13.6	1.9	5.4	0.43
	5,500	3.49	12.85	280	14.2	1.5	4.4	0.45
G	6,000	3.64	12.98	130	14.9	1.2	3.5	0.46
	6,371	3.67	13.00	0	15.0	1.1	3.2	0.46

ly, Maurice Ewing, and Frank Press) have derived valuable additional information on the properties ρ, α, and β in the outer part of the Earth by analyzing records of seismic surface waves. The results assist in the fine delineation of upper mantle structure, including lateral variations.

Evidence from tidal observations. In 1862 Lord Kelvin (William Thompson) showed from an analysis of tidal observations (see TIDES), including bodily tides of the solid Earth, that the mean rigidity of the Earth (in response to stresses of tidal periods) is about twice that of ordinary steel. When estimates were made later of the rigidity in the mantle, it became possible to estimate the mean rigidity of the core. In 1926 Jeffreys inferred in this way that the core is mostly fluid, a result also indicated by the failure to detect S waves below the mantle. This result was substantiated and elaborated by the Japanese scientist H. Takeuchi in 1950. The calculations establish that the outer core is fluid but fall short of pronouncing on the fluidity or solidity of the small inner core. Thus, in effect, observational values of k/ρ and μ/ρ are separately available down to nearly 5,000 kilometres (3,000 miles) depth.

Fluidity of outer core

Density gradient in the Earth. Density gradients $d\rho/dz$ (rates of change of density with depth z) inside the Earth can be estimated with the help of the equations:

$$d\rho/dz = \eta G m\rho/r^2\phi, \quad (4)$$

$$\phi = \alpha^2 - 4\beta^2/3 = k/\rho, \quad (5)$$

$$\eta = dk/dp - g^{-1}d\phi/dz. \quad (6)$$

In these equations m denotes the mass inside a sphere of the Earth of radius r. In deriving equations (4) and (6), hydrostatic stress and spherical symmetry are assumed, and the theory of gravitation is used.

The special case of equation (4), in which the coefficient η (Greek letter eta) is equal to unity (one unit), innate in early geophysical theory of the French mathematicians Adrien-Maria Legendre and Pierre-Simon Laplace, was applied by the American geophysicists E.D. Williamson and L.H. Adams in 1923 using seismological data on α and β to determine the parameter ϕ. But this case (often called the Williamson–Adams equation) gives only that contribution to density gradient that arises from pure adiabatic compression—that is, compression without heat exchange—and is seriously inadequate in some parts of the Earth. The Legendre–Laplace theory was generalized by Bullen to the equations (4)–(6) as shown; the coefficient η takes account of deviations from adiabatic temperature gradients and of variations of chemical composition and phase inside the Earth. The utility of equation (6) lies in the fact that, for much of the Earth's interior, numerical bounds (limits) can now be set for the quantities dk/dp, g, and $d\phi/dz$, enabling η and hence $d\rho/dz$ to be assessed to useful precision.

In applying equation (4), independent evidence is required on the density ρ' just below the crust. The value of ρ' was assessed by Williamson and Adams as 3.3 grams per cubic centimetre on the basis of petrological evidence and comparisons of seismological data on α and β with experimental data of rocks. Any error in this value is unlikely to exceed 0.2 gram per cubic centimetre and is unimportant in the estimation of densities in the deeper interior.

Derivation of Earth models. During 1936–42, the foregoing principles were applied to construct what Bullen called A-type models, using the data on α, β, and the Earth's moment of inertia available at the time. Significant variations of chemical composition or phase (see EARTH, STRUCTURE AND COMPOSITION OF) were shown to exist inside the mantle, with corresponding departures of the coefficient η from unity. Down to about 5,000 kilometres (3,000 miles) depth, these models gave the density distribution within 5 percent or less and correspondingly reliable distributions of the properties k, μ, p, and g. Further down, only lower bounds could be set for the density, the minimum value at the centre being 12.2 grams per cubic centimetre. Improvements since 1942 have been mainly of the character of successive approximations starting from the A-type models.

A nomenclature, since widely used, was introduced during this work to delineate internal layers of the Earth (see Figure 1): A for the crust; B, C, and D for regions of the mantle; and E, F, and G for the core. The regions B and C together constitute the upper mantle, extending down to about 900 kilometres (550 miles) depth; D is the lower mantle; E and G relate to the outer and inner core proper and F to a possible transition zone between. Some modifications, chiefly in the form of subdivisions of layers, will be indicated below.

The A-type calculations revealed an interesting property of the incompressibility k. In contrast to ρ and μ, both k and its pressure gradient dk/dp showed at most quite small changes at the mantle–core boundary, leading to the idea that incompressibility varies fairly smoothly with pressure throughout the whole of the Earth's lower mantle and core. This hypothesis was fortified by later evidence from finite-strain theory and experiments on the behaviour of rocks at high pressure.

From the evidence on the compressibility, the inner core was inferred by Bullen in 1946 to be solid, the jump in P-wave velocity from outer to inner core being associated through equation (3a) with a jump in the rigidity. The presence of significant rigidity in the inner core is now supported by a variety of indirect evidence, including thermal and geomagnetic evidence. Until 1972 means had not been found to detect the rigidity directly (*i.e.*, seismically), but recent investigators now claim to have detected S waves in the core. Their results would indicate a rigidity of at least 1.0×10^{12} dynes/cm^2, a value greater than that of ordinary steel.

Attempts to measure core rigidity

The findings on the rate of change of rigidity with pressure (dk/dp) led to a subdivision of the region D into

D' and D'' (see Figure 1). The region D'', which occupies the lowest 200 kilometres (100 miles) of the mantle, is an abnormal region in which the seismic velocity gradients (rates of change of velocity with depth) $d\alpha/dz$ and $d\beta/dz$ become zero or negative. Through the relations (4)–(6), the evidence on dk/dp showed that either the density gradient is markedly greater in the region D'' than in D', or there is a marked fall in rigidity inside D'', or possibly both.

Various implications of the evidence on compressibility were incorporated in Bullen's Earth model B of 1950, including a preliminary estimate of the Earth's central density.

Evidence from free Earth oscillations. On May 22, 1960, there occurred in Chile a great earthquake that, indisputably for the first time, excited free oscillations of the whole Earth, with a period of up to an hour. Analyses of observatory records (following pioneering work of the American geophysicist Hugo Benioff on the instrumental side) have since yielded reliable estimates of more than 100 separate free Earth oscillation periods (fundamental and overtone). Each period estimate adds, in effect, one new constraint on the possible distributions of ρ, k, and μ in the Earth. Early analyses confirmed the reliability of the A-type and B models within the estimated accuracy. Subsequently, the oscillation data have been applied in an endeavour to refine the previously existing Earth models.

Further evidence bearing on the density distribution. P.W. Bridgman and F. Birch, two American scientists, provided much experimental data on the behaviour of rocks and metals at high pressures and temperatures. Birch also developed a simplified version of finite-strain theory for application to the Earth's interior. In 1952 he estimated thermodynamical contributions to the Earth's density gradient, postulated empirical relations connecting density with P-wave velocity for materials of assigned atomic weights, provided some additional evidence on the variation of compressibility and its derivative (dk/dp) in the Earth, and formulated trial equations of state connecting density, compressibility, and pressure. His results have usefully confirmed some of the findings from other sources on the internal mechanical properties of the Earth, especially the mantle, and have supplied additional detail on a number of points (see ROCKS, PHYSICAL PROPERTIES OF).

During 1961–63, Birch used data from shock-wave experiments involving transitory pressures up to several million atmospheres to infer an upper bound of about 13 grams per cubic centimetre for the Earth's central density. Through equations (4)–(6), Bullen showed that, using the latest data on P-wave velocities, any value of the central density less than 14.7 grams per cubic centimetre would require the inner core to be solid with the rigidity mildly increasing in depth. These results gave a measure of new support to the presence of solidity in the inner core.

Solidity of inner core

Distribution of ρ, k, μ, p, g, and other factors. Representative values of the density and related properties of the Earth's interior, as inferred from the evidence given above, are shown for selected depths in the Table and Figures 2 and 3. The values relate to a spherically symmetrical Earth model and hence are global averages for the depths given. They are based principally on a procedure in which an A-type model was progressively modified to fit recent evidence and accord fairly well with results derived by other methods.

Some of the main features exhibited in the Table are: the jump in density (5.6 to 9.9 grams per cubic centimetre) at the mantle–core boundary (at a depth of about 2,900 kilometres [1,800 miles]); the currently favoured value of 13 grams per cubic centimetre for the central density; the fairly smooth variation of the incompressibility throughout the Earth; the steady increase of the rigidity μ throughout most of the mantle; estimates of rigidity in the inner core; and details on pressure and gravity. The values of Young's modulus E and Poisson's ratio σ are derived from those of k (incompressibility) and μ (rigidity) by simple standard formulas (see ELASTICITY).

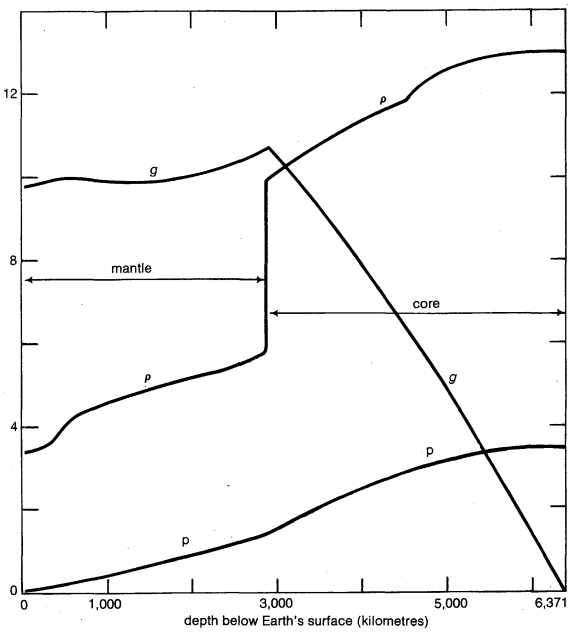


Figure 2: Variation of the density ρ (in grams per cubic centimetre), gravitational intensity g (in 100 centimetres per second squared), and pressure p (in 10^{12} dynes per centimetre squared) throughout the interior of the Earth (see text).

A broad description of the internal layering is as follows. The conventional crustal region A is taken as averaging 15 kilometres (about 10 miles) thick with a mean density of 2.84 grams per cubic centimetre. The region B (15–350 kilometres [10–200 miles] in depth) is subdivided into B' and B'' by a boundary (not closely determined) at 60 kilometres (40 miles) depth, at which small diminutions in β and μ probably occur. The average value of the coefficient η inside region B is noticeably less than unity and would correspond to the presence of a high temperature gradient (which somewhat offsets the rate of density increase due to pure compression). In the region C (350–900 kilometres [200–600 miles]), in which the coefficient η is greater than unity, there are significant changes in chemical composition or phase. The subregions C' and C'' are separated by a boundary at 650 kilometres (400 miles) depth, where there are small but fairly sharp changes in the density gradients. The changes from C'' to D are gradual, so that the boundary at 900 kilometres (600 miles) is not closely determined. The subregion D' (900–2,700 kilometres [600–1,700 miles]) appears to be nearly chemically homogeneous with η approximately equal to unity. Inside the subregion D'', the coefficient may be as high as three units. The outer core proper (region E) extends to about 4,500 kilometres (2,800 miles) in depth and may deviate a little, but not far, from chemical homogeneity in the upper part. The inner core proper (region G) has a radius of about 1,200 kilometres (800 miles). The intermediate zone F may include one or more transition regions (see EARTH, STRUCTURE AND COMPOSITION OF).

DEPARTURES FROM SPHERICAL SYMMETRY

Oblateness. The outstanding departure from spherical symmetry is the Earth's oblateness of figure—that is, flattening at the poles (see EARTH, FIGURE OF; EARTH, GRAVITATIONAL FIELD OF)—a consequence principally of the Earth's axial rotation. The measure of oblateness is ellipticity, which is equal to the ratio of the difference between the equatorial radius and polar radius (about 26.5 kilometres) to the equatorial radius. The ellipticities of surfaces of constant density and pressure within the Earth diminish steadily from the surface inward, ranging from 0.00335 at the outside surface, to about 0.0025 at the mantle–core boundary, to about 0.0024 near the centre. Allowances for oblateness have to be made in a variety of problems—*e.g.*, in deriving seismic travel-time tables.

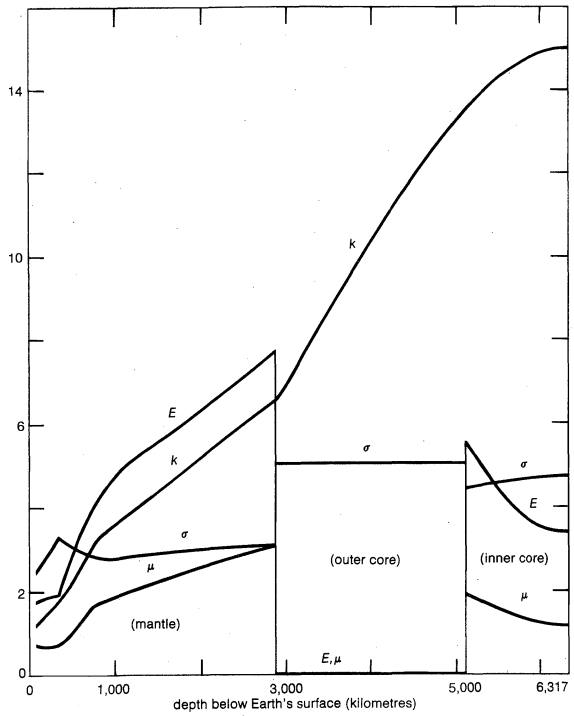


Figure 3: Variation of the principal elasticity variables throughout the interior of the Earth: k is incompressibility, μ is rigidity, and E is Young's modulus (each in units of 10^{12} dynes per centimetre squared); σ is Poisson's ratio (in units of 0.1; see text).

Lateral variations. The remaining departures from spherical symmetry occur predominantly in the crust and upper mantle. The depth to the Mohorovičić discontinuity is about 35 kilometres (about 20 miles) in continental shield areas, nearly 70 kilometres (40 miles) under some mountain chains, but only about ten kilometres (about five miles) in oceanic regions. Lateral variations associated with these crustal differences persist for some distance down into the mantle, becoming undetectable at 400–600 kilometres (250–400 miles) depth. The variations chiefly involve small differences in the rates of change of rigidity and density under different geographical regions, and they complicate the task of determining the P and S seismic velocities to close accuracy.

Distribution of deviatoric stresses

The distribution of deviatoric stresses below the Earth's surface is far from symmetrical. This is indicated by the predominant occurrence of major earthquakes in limited zones, mainly near the boundary of the Pacific Ocean Basin and on a belt running through the East Indies to the Mediterranean. Other phenomena such as mountains and certain sequences of geophysical importance are likewise unsymmetrically distributed; for example, sequences of ocean troughs, shallow earthquakes, gravity anomalies, large earthquakes, island arc systems, volcanoes and earthquakes of deep focus.

Isostasy. The visible inequalities at the Earth's surface in general produce much less disturbance in the gravitational intensity (g) than would be the case if they were simple (positive or negative) loads. Thus, elevations of the surface are "isostatically compensated" to greater or less extent by reduced density for some distance below—depressions by increased density. Isostasy is not complete, and gravity anomalies remaining after theoretical allowance for isostasy contribute further evidence on deviatoric stresses existing at present. Isostasy to some extent counterbalances some of the mechanical effects of crustal asymmetries.

ANELASTICITY IN THE EARTH

The results given above relate mainly to properties of the Earth at its present stage of evolution and ignore anelasticity (departures from perfect elasticity), and they provide a basis—in a sense supply "boundary conditions"—for investigating mechanical aspects of the past evolution of the Earth. The investigations bear on possible large-scale happenings in the past (see MOUNTAIN-BUILDING PROCESSES and CONTINENTAL DRIFT) and involve, fairly crucially, difficult questions of the anelasticity and strength of the Earth's materials.

There is evidence that all materials show some creep (gradual movement of material after the removal of applied deviatoric stress) even under quite small deviatoric stresses (P_{ij}) when the stresses are sufficiently long sustained. This carries the consequence that the strength of a material may depend not only on the P_{ij} but also on the time during which the P_{ij} are applied. Over geological time, creep may be substantial. The perfect elasticity stress–strain relations, equations (1), then cease to be an adequate model and have to be amended to allow for anelasticity. To some extent, this applies even for times well short of geological time. The mechanics of anelasticity are so complicated that, while some limited quantitative inferences of geophysical interest can be made, many conclusions in this field are qualitative and circumstantial and fall well short of the precision of the earlier results.

Damping of seismic waves. Evidence on anelasticity in response to stresses with periods up to an hour comes from the observed damping of bodily and surface seismic waves and free Earth oscillations. The results are commonly expressed in terms of the quantity Q defined so that the damping factor is the exponential of minus 2π divided by Q, $\exp(-2\pi/Q)$, for one wavelength; thus, $1/Q$ is a measure of the dissipation of the wave energy. It has been estimated that the order of magnitude of Q is 400 in the average crust, 100–200 in the upper mantle, and 1,000–2,000 at greater depths. (These estimates are simplified; *e.g.*, they ignore dependence of Q on the wave period.) The evidence suggests that Q is a minimum at about 70 kilometres (40 miles) depth, a point that bears on strength in the upper part of the mantle.

Damping of Eulerian nutation. The Eulerian nutation, an oscillation of the Earth's axis now known to have a period of about 14 months, appears to be damped with an amplitude reduction factor of 1/e in about 15 years (see EARTH, FIGURE OF). There are problems in assessing the damping accurately and determining its mechanism. It is likely but not certain that the damping is associated with energy dissipation inside the Earth's mantle. The observations have been used to supply evidence on anelasticity in respect of stress periods of the order of some years.

Region of energy damping

Suggested laws of imperfect elasticity. The model relations (1) are equivalent to a pair of relations expressing that (a) the pressure is equal to minus the product of incompressibility (k) and dilation (θ),

$$p = -k\theta; \tag{7}$$

and (b) the deviatoric stresses (P_{ij}) are equal to twice the product of the rigidity (μ) and the deviatoric strains (E_{ij}, defined analogously to the P_{ij}),

$$P_{ij} = 2\mu E_{ij} \tag{8}$$

In general, anelasticity is significantly observed only under deviatoric stress, so that only the relations (8), which for convenience will be abbreviated to $P = 2\mu E$, have to be modified.

Simple modifications are obtained by adding to both sides of the equation $P = 2\mu E$ terms that involve the rate of change of stress (dP/dt) or strain (dE/dt). An example is the model set of relations

$$P + \tau dP/dt = 2\mu E + 2\nu dE/dt, \tag{9}$$

involving the parameters μ, τ, and ν. Particular cases of this model are the firmoviscous model for which the parameter τ equals zero, and the elasticoviscous model for which μ is set equal to zero in (9). Between them, these models exhibit in a general way such properties as sudden change of strain (as in perfect elasticity) following sudden change of stress, creep or plastic flow under sustained stress, and elastic afterworking after removal of stress, with a possible permanent set. The elasticoviscous

model gives approximately perfect elastic behaviour under rapidly changing stress, and viscous fluid behaviour under slowly changing stress. Analysis of the models shows that mechanical properties such as rigidity and viscosity are not in general determined by a single parameter but depend on combinations of parameters and markedly on the period of the applied stress. The above models are too simple to agree with observational evidence in fine detail, but they are useful in special limited geophysical contexts.

Models designed to suit wider geophysical contexts are more complicated. An example is the Jeffreys model, based on an earlier model of C. Lomnitz, which for constant deviatoric stresses (P) takes the form (involving the four parameters μ, q, a, and α):

$$2\mu E = P\{1 + (q/\alpha)\,[(1 + at)^{\alpha} - 1]\}. \qquad (10)$$

This model takes account of certain laboratory evidence and data on the damping of seismic waves and nutation. When α is equal to unity, equation (10) is equivalent to the elasticoviscous model. But Jeffreys requires α to be near 0.2, a difference that has far-reaching effects. After testing this model against certain evidence on the mechanical properties of the Moon, Jeffreys concluded that it is also applicable on the geological time scale.

Completely different approaches involving more complicated theory have been made from considerations at the microscopic level, through solid-state and lattice theory and related experimental evidence. A difficult problem in these approaches is to ascertain the parameters that are likely to be important in the macroscopic Earth context. Notwithstanding considerable progress, there is as yet no comprehensive, widely accepted quantitative representation of the Earth's anelasticity.

Attempts have been made to assess viscosity in the Earth's mantle using such evidence as the Fennoscandian uplift (if attributed to deglaciation) and the excess of the Earth's ellipticity as revealed through analyses of satellite orbits over the value calculated on the hydrostatic theory. The assessments are of the order of 10^{21} poise (a unit of viscosity, dyne-second per square centimetre) and upward (the viscosity of castor oil is 10 poise and glass 10^{13} poise).

In interpreting these results, it should be noted that the reference to viscosity in the Earth context assumes rather over-simple stress–strain relations.

RESPONSE OF THE EARTH TO STRESSES OF LONG DURATION

The presence of anelasticity in the Earth affects possible large-scale movements in two ways. First, it permits the possible flow of materials that behave as ordinary solids under stresses of limited duration. Second, it causes mechanical energy to be dissipated. For large-scale flow to be maintained, this dissipation has to be overcome by sufficiently large independent sources of energy release. Also, the forces arising from the energy release have to be appropriately directed.

Sources of energy. The possible sources of energy are dynamical, thermodynamical, gravitational, and chemical, including energy latent in particular phases of materials. Of these, the thermodynamical and gravitational are the primary sources of any positive supply of energy that can be expected to be maintained over geological time. The available gravitational energy includes energy connected with lateral variations of density in the upper mantle and possible small departures from monotonic variation with depth of the "uncompressed" density (*i.e.*, the density when reduced to zero pressure). The thermodynamical energy includes energy changes that may be associated with cooling inside the Earth (resulting also in further gravitational energy being made available) and energy from heat sources (radioactivity; see EARTH, HEAT FLOW IN).

If sufficiently detailed evidence on the sources of energy and on anelasticity were available, mechanical principles could be applied to infer large-scale past movements inside and on the Earth. As yet, the usual procedure is the other way around—to set up hypotheses of past movements on independent grounds (in which much of the evidence is of a nonmechanical type) and then test against the mechanical evidence.

Mountain building and continental drift. The most evident large-scale Earth movement is the intermittent elevation of mountains (see MOUNTAIN-BUILDING PROCESSES). The long-standing theory of mountain building assumes thermal contraction in the outermost 700 kilometres (400 miles) of the Earth as the primary source of energy, though another theory assumes thermal expansion arising from internal heat sources. The formation of a major mountain system involves horizontal displacements of the order of 50 kilometres (30 miles).

The 1960s and 1970s brought a revival of interest in the hypothesis of continental drift, another type of assumed large-scale movement that had (on grounds mainly somewhat different from the present) been strongly advocated by the German meteorologist Alfred Wegener in 1912–27. The evidence for drift, which is diverse (see CONTINENTAL DRIFT), takes account of observations of paleomagnetism, details on the spreading of ocean floors in the vicinity of midocean ridges, as well as older geological and paleontological evidence on the edges of continents supposed to have been once in contact. Many geophysicists conclude that there is now strong evidence that continents have steadily drifted apart since the beginning of the Mesozoic Era, about 200,000,000 years ago, without serious distortion of the coastlines originally in contact. Followers of this view associate many phenomena with continental drift; for example, mountain building, the asymmetric distribution of earthquake sources, earthquake focal mechanisms, the present distributions of island arcs, and ocean deeps.

Apart from some controversies about some of the evidence used in its support, the hypothesis of drift faces difficulties on the question of mechanism. The only proposed mechanism that has not been quickly ruled out envisages the continents as being carried like rigid plates on top of convection currents assumed to exist some distance below the crust. The comparatively small value of the attenuation parameter Q used in assessing energy dissipation, the probable reduction in the seismic S velocity, the probably high temperature gradient, and the probably abnormally low density gradient inside the outermost 100–200 kilometres (50–100 miles) of the mantle are brought to bear as evidence for convection.

No satisfactory mechanism of convection that meets all the quantitative requirements has yet been found, however, and most attempts to do this have been grossly simplified. Arguing on the basis of the model equation (10), Jeffreys has asserted in fact that convection currents cannot occur on the scale envisaged. They could occur if the parameter α could be taken as equal to unity (corresponding to the elasticoviscous model) or greater, but Jeffreys regards as crucial his assessment that the constant α is much less than unity. Until the question of mechanism is solved to general satisfaction, it appears premature, in spite of current enthusiasms, to regard continental drift as established. In the meantime, work on the question is proceeding apace.

BIBLIOGRAPHY. Due to the intricacies of the subject, there are few nontechnical books that can be recommended. The following, though not all nontechnical, contain much matter that can be understood by the lay reader: HAROLD JEFFREYS, *Earthquakes and Mountains*, 2nd ed. (1950); SCIENTIFIC AMERICAN, *The Planet Earth* (1957); BENJAMIN F. HOWELL, JR., *Introduction to Geophysics* (1959); DAVID R. BATES (ed.), *The Planet Earth*, 2nd ed. rev. (1964); THOMAS F. GASKELL (ed.), *The Earth's Mantle* (1967); and HITOSHI TAKEUCHI, *Debate About the Earth* (1967). More advanced books are: WALTER H. MUNK and G.J.F. MACDONALD, *The Rotation of the Earth* (1960); K.E. BULLEN, *An Introduction to the Theory of Seismology*, 3rd ed. (1963); HITOSHI TAKEUCHI, *Theory of the Earth's Interior* (1966); S.K. RUNCORN (ed.), *International Dictionary of Geophysics*, 2 vol. (1967), and *The Application of Modern Physics to the Earth and Planetary Interiors* (1969); MICHELE CAPUTO, *The Gravity Field of the Earth* (1967); WILLIAM M. KAULA, *An Introduction to Planetary Physics* (1968); FRANK D. STACEY, *Physics of the Earth* (1969); and SIR HAROLD JEFFREYS, *The Earth*, 5th ed. (1970); J. TUZO WILSON (ed.) *Continents Adrift* (1973).

Earth, Physiography of

The term physiography has evolved through several different meanings. Today it generally signifies the description and genetic interpretation of the relief features of the Earth's surface. Sometimes it is offered as a synonym of geomorphology, but the latter stresses the dynamic, climatic, and structural aspects of landforms, whereas physiography is reserved mainly for the descriptive treatment of the Earth's surface features.

In an areal treatment of the Earth's surface features, a scale of magnitudes (size of features in terms of factors of ten) can be discerned. Several classification schemes have been developed and they can be combined and modified as follows:

Orders of magnitude of Earth features

First order features. The primary form of the globe, namely an idealized spheroid, is characterized by a shortening in the polar axis (7,899 miles, or 12,713 kilometres) and by an equatorial bulge (maximum diameter 7,927 miles or 12,756 kilometres). Satellite data have demonstrated a minor component of triaxial, "pear-shaped" symmetry indicating that the Northern Hemisphere is about 50 feet (15 metres) higher at the pole than predicted, and about 25 feet (8 metres) lower in the middle latitudes. In the Southern Hemisphere the polar area is 50 feet (15 metres) lower than predicted and the mid-latitudes are 25 feet (8 metres) higher. The Earth's surface area is subdivided geometrically into 360 degrees of longitude and 360 degrees of latitude; a degree of latitude is 362,780.4 feet (110,575.4 metres) at the Equator and about 1 percent more at the pole. The polar circumference is 24,860 miles (40,007 kilometres), whereas the equatorial circumference is 25,326 miles (40,757 kilometres).

Second order features. The Earth's surface is subdivided into continental masses (mean elevation: 2,755 feet or 840 metres; area: 148,100,000 square kilometres) and oceanic depressions (mean elevation: −12,500 feet or −3,810 metres; area: 361,900,000 square kilometres). Seven continents are recognized: Europe, Asia, Africa, Australia, North America, South America, and Antarctica. Four oceans are recognized here: Arctic, Atlantic, Pacific, and Indian oceans. (Some classifications group the Arctic with the Atlantic, and others distinguish a Southern or Antarctic Ocean.) The areal order of magnitude of the average ocean is 10,000,000 square kilometres (13,861,000 square miles).

Third order features. The continents and oceanic depressions are further divided into positive topographic (relief) features; namely, mountain systems on the land and rises or ridges undersea. Maximum elevation on land is 8,848 metres (29,028 feet; Mt. Everest), and maximum depression of the sea floor is 11,033 metres (36,198 feet; Mariana Trench). The principal youthful mountain systems of the world are (discontinuously) in two belts, the Circum-Pacific, and the Alpide or Mediterranean-Himalayan systems. All other mountains are more subdued and dismembered or are isolated relics of ancient and deeply eroded systems. The principal oceanic topographic highs are the system of mid-ocean ridges. This forms a spine that bisects the Arctic, Atlantic, and Indian Oceans, extends around the Antarctic region, and passes through the eastern Pacific, where it dies out in the Gulf of California. This is the largest single topographic feature in the world and is approximately 40,000 miles (64,000 kilometres) long. Negative relief features of the same magnitude are lacking elsewhere on the Earth, because it is broken up into large numbers of plains, depressions, and basins. The areal order of magnitude of the major structural units is 1,000,000 square kilometres (380,000 square miles).

Fourth order features. The secondary mountain systems, mountain chains, volcanic provinces, and plateaus, alternating with plains and basins, are features of the 4th order. With a slightly different nomenclature, similar divisions exist in the ocean. The areal order of magnitude is 100,000 square kilometres (38,000 square miles).

Fifth order features. A "tract" of land distinguished by a single rock type, a single volcano, a major ridge or escarpment, or a major valley is a 5th order feature. The areal order of magnitude is 1,000 to 100 square kilometres (380 to 38 square miles).

Sixth order features. A "site," hill, gorge, cave, cliff, or beach is at this level. The areal order of magnitude is 10 square kilometres to 10 square metres (3.8 square miles to 100 square feet).

Seventh order features. A large boulder to a grain of sand are at this level. The areal order of magnitude is 1 square meter to 1 square millimetre (10 square feet to 0.001 square inch). Smaller particles, such as a speck of dust, and those down to the level of molecules, atoms, and electrons, do not concern the topic of physiography.

This article treats the general physiography of the Earth, provides more specific coverage of the configurations of the several continents and oceanic depressions, and indicates some of the factors that govern the observed configurations. For further detail on continental features, see MOUNTAIN RANGES AND MOUNTAIN BELTS; PLATEAUS AND BASINS; VOLCANOES; HILLSLOPES; DESERTS; and JUNGLES AND RAIN FORESTS. Oceanic features are treated in CONTINENTAL SHELF AND SLOPE; ISLAND ARCS; and OCEAN BASINS. See also EARTH, FIGURE OF; and EARTH, GEOLOGICAL HISTORY OF for discussions of relevant geodetic and geological concepts. See further LANDFORM EVOLUTION for treatment of the change in Earth features through time.

HYPSOGRAPHY OF THE EARTH'S SURFACE

As noted under the second order features, there is a clear distinction between continents and oceans in terms of mean elevation, and furthermore a cumulative height frequency curve shows two distinct (bimodal) peaks, one between mean sea level and +1000 metres (3,000 feet), which intersects the coastal plain and continental shelf surface, and the other between −4000 and −5000 metres (−13,000 and −16,400 feet), the mean level of the deep sea floor. The former occupies more than 21 percent of the Earth's surface and the latter 22 percent of the surface. Although 5.5 percent of the Earth (and 7.5 percent of the ocean area) is represented by continental

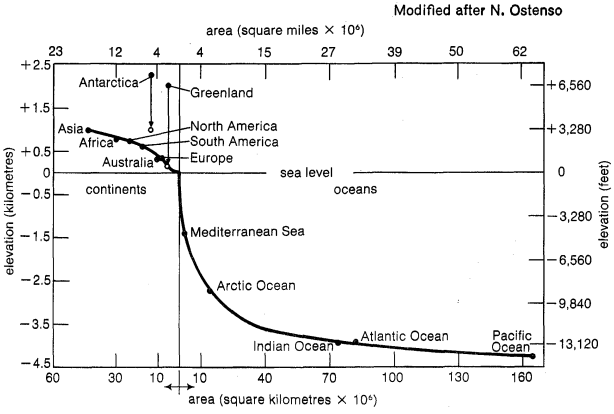


Figure 1: Comparative areas and elevations of the world's continents and oceans.

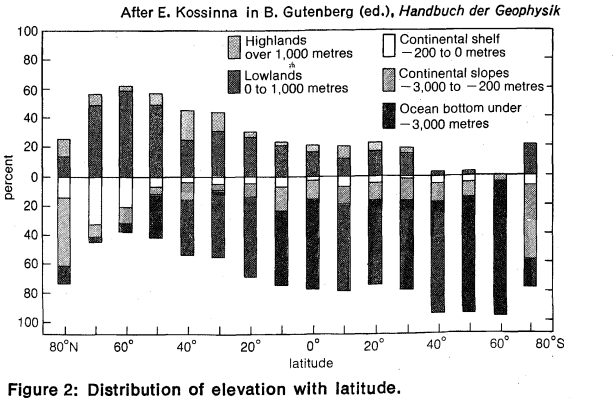


Figure 2: Distribution of elevation with latitude.

Table 1: World Distribution of Land and Water

(in 5° belts; based on the Helmert-Hayford Ellipsoid)

latitude	Northern Hemisphere				Southern Hemisphere			
	water (000 km²)	land (000 km²)	water (percent)	land (percent)	water (000 km²)	land (000 km²)	water (percent)	land (percent)
90°–85°	979	—	100.0	—	—	979	—	100.0
85°–80°	2,546	384	86.9	13.1	—	2,930	—	100.0
80°–75°	3,743	1,112	77.1	22.9	522	4,333	10.7	89.3
75°–70°	4,415	2,327	65.5	34.5	2,605	4,137	38.6	61.4
70°–65°	2,457	6,118	28.7	71.3	6,818	1,756	79.5	20.5
65°–60°	3,124	7,212	30.2	69.8	10,304	32	99.7	0.3
60°–55°	5,400	6,615	45.0	55.0	12,007	9	99.9	0.1
55°–50°	5,530	8,069	40.7	59.3	13,392	207	98.5	1.5
50°–45°	6,614	8,461	43.8	56.2	14,697	377	97.5	2.5
45°–40°	8,413	8,018	51.2	48.8	15,837	594	96.4	3.6
40°–35°	10,032	7,629	56.8	43.2	16,488	1,173	93.4	6.6
35°–30°	10,809	7,945	57.7	42.3	15,787	2,968	84.2	15.8
30°–25°	11,750	7,954	59.6	40.4	15,442	4,262	78.4	21.6
25°–20°	13,358	7,147	65.2	34.8	15,454	5,051	75.4	24.6
20°–15°	14,986	6,166	70.8	29.2	16,152	5,000	76.4	23.6
15°–10°	16,558	5,082	76.5	23.5	17,216	4,423	79.6	20.4
10°–5°	16,633	5,334	75.7	24.3	16,903	5,064	76.9	23.1
5°–0°	17,392	4,738	78.6	21.4	16,797	5,334	75.9	24.1
90°–0°	154,739	100,311	60.66	39.34	206,421	48,629	80.92	19.08

90° N–90° S: ocean = 361,160,000 km² (70.80%), land = 148,940,000 km² (29.20%)

Source: E. Kossinna, "Die Erdoberfläche," *Handbuch der Geophysik*, 1933.

shelf (0 to −200 metres [0 to −660 feet]), the depth range −200 to −1000 metres (−660 to −3,000 feet) only corresponds to 2.9 percent of the Earth, emphasizing the steepness of the continental slope.

Distribution of land and sea. Oceans cover 70 percent of the globe. Clearly, the continents and oceans of the world are not evenly distributed (Table 1). They are roughly antipodal, displaying a certain regularity. The Northern Hemisphere has the larger proportion of land (30 percent; in round figures 100,000,000 square kilometres [38,000,000 square miles]); and the Southern Hemisphere the smaller (19 percent; approximately 48,600,000 square kilometres [18,700,000 square miles]). Yet the Antarctic continent coincides symmetrically with the south polar region, in contrast to the north polar area, which corresponds to the Arctic Ocean. This curious symmetry prompted the tetrahedral theory proposed in 1875 by Lowthian Green, which depicted the four oceans as lying on the four faces of a tetrahedron. There would be three nodes in the Northern Hemisphere, 120° apart, along latitude 20° N (in The Sudan, Yucatan, and south of Japan, respectively) and a fourth at the South Pole. It was argued that the tetrahedron is the geometric figure with the maximum surface area for a given volume, in contrast to the sphere, which has a minimum area for a given volume. If the Earth had, on cooling, contracted from a perfect sphere it would tend toward the tetrahedral slope. Although no available data support this theory (and both the contraction and cooling theories have been rejected by numbers of authorities), the symmetry remains unexplained.

The tetrahedral theory of land-sea distribution

A second feature of importance concerns the shape of continents or major structural units. Several continents tend toward a triangular form—broad to the north and pointed to the south: North America, South America, Africa, the Indian subcontinent, and smaller structural blocks such as the Sinai Peninsula. In general the great fracture patterns (lineaments) tend to follow NW to SE and NE to SW patterns, except in the Pacific where they tend to be E to W. Several theories, in terms of a geological history of crustal shearing, have been proposed to explain the situation. In 1891 Wilhelm Prinz, a Belgian astronomer, in his torsion theory, compared the intersecting lineaments with the Martian "canals" and suggested that the southern continents had drifted farther eastward than the northern ones, thus anticipating in a remarkable way the theory of continental drift (*q.v.*). His delineation of the E–W belt of maximum mountain-building has now been identified with the equatorial zone of 100,000,000 years ago (displaced now due to polar wandering).

Lineaments and the torsion theory

Other theories to explain the shapes of continents and the major lineaments involve polar migration and various related phenomena.

Symmetry. There is an elegant symmetry to be seen in certain terrestrial features. The parallel shapes of the west coast of Africa and the east coast of South America were first observed by Roger Bacon. The form of the Atlantic Ocean, particularly now that the existence of the Mid-Atlantic Ridge is known, is quite remarkable, although the Antillean (Caribbean) Arc and the analogous Scotia Arc represent paired asymmetric re-entrants in the oceanic depression. The Indian Ocean floor shows a totally different pattern; its mid-ocean ridge is roughly central, but it splits in the south into a Y-shape, and in the northwest it diverges into the Aden–Red Sea Rift System. The major fracture systems of the Atlantic are E to W, but those of the Indian Ocean are N to S. The Pacific in contrast is much less symmetric, although there is a certain parallelism between the Samoa Angle and the Panama Angle and a bisymmetry to the NW and SE. The mid-ocean ridge, however, is displaced completely to the east.

On a smaller scale there is an elegant symmetry displayed by the sides of the Red Sea and the matching escarpments and highlands of the Eastern Desert of Egypt and The Sudan, and those of the Hejaz and Yemen. This type of symmetry is to be seen throughout the rift valley

Table 2: Areas in Elevation Steps; Totals, Volumes and Mean Altitudes

(000,000 km² unless otherwise specified)

continent	below 0	0–200	200–500	500–1,000	1,000–2,000	2,000–3,000	3,000–4,000	4,000–5,000	above 5,000	total area	volume (000 km³)	mean altitude (m)
Europe	0.2	5.4	2.1	1.5	0.5	0.2	0.0	0.0	—	9.9	3,360	340
Asia	0.5	10.4	9.4	10.5	8.0	2.3	0.9	1.8	0.5	44.3	42,520	960
Africa	0.1	2.9	11.6	8.4	5.8	0.8	0.3	0.0	0.0	29.8	22,400	750
Australia	0.1	3.5	3.7	1.5	0.2	0.0	0.0	0.0	0.0	8.9	3,040	340
North America	0.1	7.2	7.4	2.9	4.0	2.2	0.4	0.0	0.0	24.1	17,320	720
South America	—	6.8	5.3	3.4	1.0	0.4	0.5	0.4	0.0	17.8	10,550	590
Antarctica	—	0.9	0.4	0.7	3.1	5.3	3.7	0.0	—	14.1	31,100	2,200
Land	0.8	37.0	39.9	28.9	22.6	11.2	5.8	2.2	0.5	148.9	130,290	875

Source: E. Kossinna, "Die Erdoberfläche," *Handbuch der Geophysik*, 1933.

system of Africa and elsewhere in the world where a simple fracturing with central collapse seems to have taken place; *e.g.*, the Rhine Graben of Germany.

Average elevation of continents. If the mean elevation of each continent is plotted on a height-area curve, all of the points for the continents fall on a uniform curve except Antarctica, which is more than four times higher than expected (Table 2). Clearly this departure is related to the accumulation of its ice sheet, which today covers about 23,500,000 square kilometres (9,100,000 square miles). The uniform relations of the non-glacier-covered continents would seem to be related to the nature of fluvial gradation (erosion by running water), but the phenomenon does not appear to have been subjected to critical analysis. The mean hypsometric gradient (altitude–area relation) for Antarctica is closely matched in shape by Greenland, even though its ice area is only 2,500,000 square kilometres (960,000 square miles), about 9 percent that of the Antarctic mass.

Isostasy and land elevation

The elevation of the continents is in part a function of isostasy (natural buoyancy), which requires the Earth's crust to be thicker in areas of relatively less dense rocks, and vice versa, thus insuring that the Earth's crust in general is in hydrostatic equilibrium. If the entire crust were of the same composition, the elevation would correspond to an ideal spheroid, but during the course of geologic history the continents have been segregated as masses of lighter specific gravity (average about 2.7), whereas the crust of the ocean floors consists of denser materials (average about 3.3). The continental crust is characterized by rock of approximately granitic or granodioritic composition (silica-alumina rich, termed sial) and the oceanic crust is marked by basalt or peridotite rocks (silica-magnesia rich, termed sima). According to the concept of isostasy, the sial must float higher than the sima, and according to the hypsographic curve, the actual difference in mean level is about 4,000 metres (13,000 feet). The reason that mountains stand higher than the rest of the continent is because in their formation they develop deep roots of light (granitic) crustal material that enables them to float even higher than the average continental elevation.

Most of the Earth's crust is approximately in hydrostatic equilibrium in this way, so that when erosion occurs and rivers transport large quantities of weathered material away from the upland areas to be deposited in the oceans, there is a tendency for the hinterland to rise isostatically, and for the adjacent ocean floor to sink. Uplifted terraces that represent old river valleys or coastlines testify to the uplift, and deep oil drilling on the continental shelf demonstrates that it has become progressively downwarped through time. The rates of uplift and downwarping vary according to relative relief but average 0.1 to 1.0 millimetre (0.004 to 0.04 inch) per year.

Plainlands and shield areas

Very ancient plainlands, as in the interiors of North America, Central Asia, Africa, Brazil, and Australia (the ancient geological "shields"), have reached an approximate equilibrium; they have very little relative relief, tend to be stable, and suffer a minimum of erosion. Even in the ancient shields, however, the equilibrium is easily upset by an event such as an ice age. The load of ice depresses the crust in a ratio of about 1:3 (the respective specific gravities of ice and rock) and after glaciation has passed, the ice melts and the crust tends to return to its former position. This is never completely achieved because the glaciation leaves behind loads of boulders and other debris, which retard the isostatic readjustment. There may also be permanent modification to the subcrustal mass distribution. In any case, the major continents are marked by numbers of broad, dish-shaped depressions, many of which appear to be the sites of former glacial depressions. Hudson Bay and the Baltic Sea fit this model; in Africa, the Taoudenni, Chad, Congo, and Kalahari depressions seem to represent the sites of Paleozoic (570,000,000 to 225,000,000 years ago) ice ages.

Coastlines. Not counting islands, the continental coastlines of the world measure approximately 312,000 kilometres (193,000 miles; see Table 3). Many aspects of the physiography are directly related to the proximity of the ocean (Table 4). These include climate (oceanicity versus continentality), stream profiles, cyclic salt transport (and thus lake and soil salinity), and other phenomena.

Table 3: Coastlines of Continents*

continent	area (000,000 km²)	coastal length (km)
Europe	9.32	37,200
Asia	41.48	70,600
Eurasia	50.70	107,800
Africa	29.20	30,500
Australia	7.60	19,500
North America	19.98	75,500
South America	17.63	28,700
Antarctica	14.14	50,000 (?)
Total		312,000

*Excluding islands.
Source: Adapted from E. Kossinna, "Die Erdoberfläche," *Handbuch der Geophysik*, 1933.

From the point of view of geological structure, the coastal plains of the world are in continuity with the continental shelves. The widest coastal plains usually match the widest shelves. The present mean sea level approximates a medial line between the respective inner and outer limits of the two. This appears to be in part a function of isostasy, which tends to raise the interior as it is eroded and depress the shelf as it is loaded by sediment. Both features are, however, additionally modified by the glacio-eustatic rise and fall of world sea level that has marked the advance and retreat of glaciers during the last 2,000,000 years, the Quaternary Period, marked by the Great Ice Age. The Ice Age was not a single event but was marked by repeated advances and retreats and in fact is not yet over; the next glacial advance is predicted within a few thousand years (see CLIMATIC CHANGE). The areas of the continental shelves (as shown by Table 5) and of the coastal plains vary remarkably from continent to continent. In Africa, for example, the coastal plains (below 200 metres [660 feet]) represent about 10 percent of the total area, the shelves only 4 percent. For Europe, in contrast, the coastal plains represent over 55 percent and the shelves nearly 30 percent. This relationship is attributed to geologic history. In the case of Africa, the bulk of the coastline is a zone of fractures that originated at the time when continental drift split the former "megacontinent" of Gondwanaland apart, separating Africa from South America, Antarctica, India, and Australia; the last critical fracturing occurred about 100,000,000 years ago. Europe, too, was affected by drifting when it split off from North America, but this was more than 200,000,000 years ago and affected only the Atlantic courses. The other coasts of Europe are marked either by the drowning of broad shelf areas, notably to the north, or by mountain-building (orogenic) processes, along the Mediterranean borders with only minor segments of drift-fractured coasts.

Coastal classification and continental drift

A very broad classification of the world's coastlines, therefore, names the orogenic coasts the "Pacific type," and the fractured coasts the "Atlantic type." These designations are only just labels, of course, for examples of each type are found in all parts of the world. Nevertheless, the general pattern applies as follows:

1. Pacific Type—Circum-Pacific coasts: western borders of the Americas, Aleutians, Japan, Philippines, New

Table 4: Continentality—Distances from Oceans

continent	distance from coast (km)		regions near ocean (percent)	regions far from ocean (percent)
	average	maximum		
Europe	340	1,550	62	38
Asia	780	2,400	61	39
Africa	670	1,800	53	47
Australia	350	920	55	45
North America	470	1,650	58	42
South America	550	1,600	56	44

Source: E. Kossinna, "Die Erdoberfläche," *Handbuch der Geophysik*, 1933.

Table 5: Land Areas and Continental Shelf Areas of the World
(000,000 km²)

continent	land areas	shelf areas
Europe	9.90	3.11
Asia	44.18	9.38
Africa	29.23	1.28
Australia	8.51	2.70
North America	24.06	6.74
South America	17.78	2.43
Antarctica	14.16	0.36

Source: E. Kossinna, "Die Erdoberfläche," *Handbuch der Geophysik*, 1933.

Guinea, New Zealand; Indonesian Arc; Crete and other Mediterranean shores; Caribbean Arc and Scotia Arc.

2. Atlantic Type—coasts of Atlantic (except as stated above), northwestern Mediterranean, most of Indian Ocean borders, Tasman Sea coast of Australia, and certain other western Pacific marginal seas.

This physiographic classification, as indicated above, has its roots in geologic structure and history. Throughout the Paleozoic Era (570,000,000 to 225,000,000 years ago), what is now the South Pole, the "Gondwana Pole," lay over Africa, and what is now the North Pole lay in the mid-Pacific. Polar migration accompanied by continental drifting has split "megacontinents" in the Northern and Southern hemispheres, leaving them with Atlantic type coasts. The continents spread much farther apart in the Southern Hemisphere than in the Northern Hemisphere. At the same time, an ancient Mediterranean Sea (Tethys) was crushed along the latitudinal Alpide (Alpine) Belt, and an E–W belt of mountains was produced.

PHYSIOGRAPHY OF THE CONTINENTS

Europe. The principal orographic and regional feature of Europe is the Alpine System, which embraces not only the Alps of France, Switzerland, Austria, and Italy, but also the structurally related or continuous ranges that include the Pyrenees and Betic belts of Spain, the Apennines of Italy (and Sicily), the Carpathians, Dinarides, Rhodope, and Hellenide chains of eastern Europe.

The stable Precambrian (older than 570,000,000 years) shield region of Europe is exposed in the Fennoscandian sector (Norway, Sweden, Finland, and northwestern U.S.S.R.), and the same stable block forms the deeper foundations of the Russian Platform, only outcropping locally in the southwestern Ukraine.

Older Paleozoic belts of folded rocks, the Caledonides, are today topographically much reduced. They form the western border of the Fennoscandian Shield and extend to Scotland and northern Ireland. Younger Paleozoic foldbelts, the Hercynides, are found preserved in isolated areas of western Britain and Ireland, the Spanish Meseta, Corsica and Sardinia, the French Massif Central, Brittany, the Ardennes, central Germany, and Bohemia. Distributed between, or overlapping, these numerous structural blocks, there are basins and flatlands in great number, for the most part still accumulating sediment today; examples include the London Basin, Paris Basin, North German Plain, the Hungarian (Pannonian) Plain, and the Po Plain.

European glaciation and present appearances

A key factor in determining the landscape has been the Quaternary history of multiple glaciations, which scoured and grooved the surface of the northern highland areas from Scandinavia to the British Isles, as well as the younger mountain belts of the Alps, Pyrenees, and Carpathians. The northern lowlands are thus widely characterized by moraine landscapes and broad sandy outwash plains. Particularly in the eastern parts of Europe and Asia, widespread plains were coated in loess, a wind-blown dust derived from the glacial meltwater regions, giving them a characteristically rounded, gentle topography.

Large numbers of lakes dot the former glaciated regions. Outside of the latter, there is only one large lake in Europe, Lake Balaton (230 square miles [590 square kilometres]), situated in the Hungarian Plain. The major rivers are the Danube, the Rhine, the Vistula, and the Volga. Extensive marshes also occur in the glaciated lowlands, notably the Pripet Marshes and others in Russia and Finland.

Asia. The world's largest continent, Asia, is geologically and geographically closely tied to Europe and the two are often treated as Eurasia. The great Alpine System of youthful mountains continues eastward through Anatolia in a series of double arcs, splitting well apart in Iran to form the Elburz in the north and the Zagros in the south. The ranges reunite in Afghanistan and northern Pakistan in the Hindu Kush and Pamirs, passing around a hairpin bend to form the Himalayan arc on the south and into the Karakorum and Kuen Lun chains to the north. Rejoining in Yunnan, these chains continue south through Burma, and then in a more interrupted manner through the island arcs of the Sunda and Banda islands (Indonesia), where they are joined by the Circum-Pacific belts. Throughout this mighty mountain system, the most important on the globe, there is a systematic pattern of divergence and convergence, surrounding nuclei of interior high plains or plateaus: Anatolia, Iran, Tibet, Indochina. The entire eastern border of Asia from the Verkhoyansk ranges of Yakutia to Chukotski Peninsula (with connections to Alaska) also consists of youthful mountain belts, but of the Circum-Pacific System. They extend southward as island arcs through Kamchatka to the Kurile Islands and Sakhalin, to Japan, Taiwan, the Philippines, Borneo, the Celebes (Sulawesi), and Halmahera, where they join the Australasian belts of New Guinea and New Zealand.

The old crystalline shields of Asia are multiple—the most important is the Angara Shield of Siberia (with its Anabar nucleus), largely overlapped by subhorizontal younger formations, leading to extensive plateau landscapes. In the Soviet east, there is the Kolyma Block; in southeastern China, the Sinian Block; and south of the Alpide zone, the "subcontinental" shields of peninsular India and Arabia.

Interior basins and their drainage

Fold belts of ages comparable to the Caledonian and Hercynian belts of Europe extend from the Urals to the Altai and others in Central Asia, continuing through China to the Pacific coast and into Korea. Between all these ranges are numerous large interior basins. Between the Urals and the Angara Shield lie the West Siberian Lowlands and to the south the plains of Kazakhstan, and the Kara-Kum depression of Turkmenia and Uzbekistan. In Sinkiang there are the Tarim and Dzungaria basins. In Mongolia there is the Gobi depression. Smaller depressions in China include the Szechwan Basin and the Yellow Sea lowlands. South of the Alpide belt, the Indian Shield is separated from the mountains by the Indo-Gangetic depression and in a similar way the Zagros are separated from the Arabian Shield by the Mesopotamian-Persian Gulf depression.

Asia possesses many areas of interior drainage and inland seas and lakes: the Caspian Sea, Aral Sea, Lake Balkhash, and Lake Baikal (the world's deepest lake, containing more water than all of the North American Great Lakes together). In Iran, Tibet, Sinkiang, and northwestern China, there are innumerable ephemeral, saline lakes; an example is Lop Nor, the borders of which are variable and bear traces of more extensive lake floors indicating formerly moister climates. Unique among the salt lakes is the Dead Sea, the surface of which is −392 metres (−1280 feet).

Extensive deserts occur in Arabia, notably the Rub'al-Khali, in northwestern India (Thar Desert), and in Mongolia (Gobi). The great rivers of Asia are the Ob, Yenisey, and Lena flowing to the Arctic Sea; the Amur, Huang Ho, Yangtze, and Mekong flowing to the Pacific; the Ganges, Brahmaputra, Indus, Tigris, and Euphrates, which flow to the Indian Ocean; and some small streams, such as the Orontes and Meander (Büyür Menderes Nehri), flowing into the Mediterranean and thence into the Atlantic. All four oceans are thus fed by the rivers of Asia.

Africa. Youthful fold mountains in Africa are limited to parts of the Atlas, a continuation of the Betic-Sicilian

sector of the Alpide System. Most of the continent consists of an immense Precambrian crystalline shield, subdivided into broad basins: Taoudenni, Chad, Congo, Upper Nile, Kalahari, and Karroo depressions. The shield is impressively split along its entire eastern margin, by rifts from the Red Sea, Aden, Ethiopia, and East Africa southward to Natal. Only in the Cape Ranges (Capides) are there some Paleozoic mountains, now greatly reduced in relief.

African deserts and jungles

Africa is marked on the one hand by its great deserts, the Sahara in the north, passing into the Libyan and Arabian deserts, and the Kalahari in the south; and on the other hand by its great tropical forests and jungles, notably in West Africa and the Congo. High savanna-covered plateaus extend over most of the subtropical latitudes as well as the high, terraced landscapes of East Africa. During the last ice age much of Africa was reduced to a desert, and traces of old sand dunes, now tree-covered, are widely traceable, even crossing over the Congo both from the south and from the north. Great rains then set in, reaching a peak around 12,000 years ago, leading to torrential floods along the Nile and the growth of lakes to enormous size. The ancestral Lake Chad, for example, was at least ten times larger than today. High terraces reflect former stronger precipitation feeding the East African lakes: lakes Victoria, Rudolf, Tanganyika, and Nyasa (Malawi). The great rivers of Africa, the Nile, Senegal, Niger, Congo, Zambezi, Limpopo and Orange, nearly dried up during the glacial maximum, and some became lost in swamplands, for example, in the Sudd (Nile), in the middle Niger and in the Okavango Swamps of the Kalahari depression.

Australia. Youthful mountain belts are absent from the Australian continent but run through the great Australasian subcontinental islands: New Guinea–New Caledonia–New Zealand, where Mt. Cook reaches 12,349 feet (3,764 metres) elevation. The high mountains of central New Guinea were crowned by a small ice cap during the last glaciation and New Zealand was extensively glaciated.

The (West) Australian Shield occupies two-thirds of the continent and is marked by broad savannas or sandy deserts (Simpson and Victoria deserts). Like Africa, traces of former sand ridges cover much of the northern part of the country, though now they are largely tree covered. Innumerable ephemeral salt lakes dot the interior, marking former river systems now dried out and truncated by drifting sand. Most important is the Lake Eyre Basin, which is below sea level, being fed by ephemeral rivers—Coopers Creek and the Diamantina.

A subdued Paleozoic mountain belt forms the Great Dividing Range and Australian Alps along the eastern margin of the country. Only a small area around Mt. Kosciusko (7,316 feet [2,230 metres]; the highest mountain in Australia) was glaciated, but in Tasmania to the south there was an extensive ice cap. Major rivers are the Murray and the Darling; in the tropical north there are the Fitzroy and the Catherine.

Canadian Shield, fold belts, and structural features

North America. The youthful mountains of North America belong to the Circum-Pacific system, extending through Alaska and British Columbia to form the Rocky Mountains and the Coast Ranges of the western United States, and the Sierra Madre Orientale and Occidentale of Mexico. In Guatemala, the fold belts swing eastward to form the discontinuous island arcs of the Caribbean—Jamaica, Cuba, Hispaniola, Puerto Rico, and the Leeward Islands.

The old crystalline nucleus of the continent is the Canadian Shield, which breaks up in the north to form the Canadian Arctic Archipelago and Greenland. Before continental drifting during the geologic past, Greenland fitted snugly against Labrador and Baffin Island.

Wrapped around the margin of the shield in the east there are the subdued mountains of the Paleozoic fold belts, the Appalachians, which in Newfoundland show evidence of having originally been connected to the Hercynian and Caledonian belts of Ireland. To the northwest of the Canadian Shield there is another old fold belt (the Franklin Geosyncline) extending through Banks and Ellesmere islands. A narrow belt of northeast Greenland extends this system and seems to have been originally separated by drifting from Spitsbergen, Norway, and Britain.

Marginal to the shield there are numerous large interior basins and plains, typical of the Prairie Provinces of Canada and the Midwest of the U.S. There are broad coastal plains only in the Atlantic states and on the Gulf Coast (which continue into Mexico). To the interior of the western fold belts there are interior plateaus and depressions structurally comparable with those of Iran and the Alpide Belt. Most important are the Great Salt Lake Basin of Utah, the Colorado Plateau, and the Basin and Range Province of Nevada and adjacent states that continue southeastward through Chihuahua to form the central plateau (altiplano) of Mexico. Extensive lava plains (of flood basalt) mark the landscape of the northwestern states, punctuated along its western border by youthful volcanoes in the Cascade Range. The volcanic belt is picked up again from southern Mexico to the Central American states.

Extensive glacial sheets covered much of North America (more than any other temperate continent) during the last ice age, leaving an immense number of closed drainage systems and major lakes. In the Rocky Mountains, numbers of pluvial (ice age) lakes formed (notably lakes Bonneville and Lahontan), just as in Africa, near the end of each glacial stage. Their sites are marked today by abandoned terraces. Major river systems in North America include the Mississippi (with its tributaries the Missouri, Illinois, Ohio, and others), St. Lawrence, MacKenzie, Athabasca, Churchill, Yukon, Fraser, Columbia, Snake, Colorado, Rio Grande, Hudson, and Susquehanna.

South America. The Circum-Pacific youthful mountain system extends, as the Andes, along the entire western border of South America, swinging easterly in Venezuela to join the Caribbean Arc in Trinidad, and again in the far south, in Tierra del Fuego, to join the Scotia Arc that takes in South Georgia, the South Shetlands, and South Orkneys to join the Antarctic Peninsula in Grahamland. Interior high plateaus, comparable to North America's, form the altiplano regions of Peru and Bolivia, marked in the same way by large lakes; *e.g.*, Lakes Titicaca and Poopo, that were once much more extensive.

Relation of South Americar and African features

The Precambrian shield region lies to the east and is almost continuous but customarily is divided into three parts: the Guianan Shield in the north, the Brazil Shield in the centre (and much the largest), and the Patagonian Shield in the far south. They are separated by transverse basins (respectively, those of the Amazon and the Rio de la Plata) that merge along the inner foot of the Andes with a continuous belt of piedmont depressions. Traces of the older, Paleozoic fold belts in South America are not as clear as on other continents, having been largely overwhelmed and incorporated in the younger Andean folds. A subdued branch in Argentina, however, swings eastward from near Córdoba to south of Buenos Aires, and it seems likely that this is a continuation of the South African Cape ranges in the pre-continental-drift history (earlier than about 200,000,000 years ago).

The great interior deserts of the other continents are largely absent from South America, although a well known arid coastal region is represented by the Atacama Desert of northern Chile and southern Peru. The great rivers all flow to the Atlantic—the Orinoco, Amazon, Paraná, and Paraguay (converging on the Rio de la Plata).

Antarctica. The Antarctic continent is largely ice covered because the Quaternary ice age continues unabated, but there are sufficient projecting mountains and nunataks (isolated peaks) to discern the general structure of the continent. The Circum-Pacific youthful fold belts continue through the Antarctic Peninsula into Marie Byrd Land (West Antarctica), but they were apparently separated by continental drifting from their original connection to New Zealand.

Three quarters of Antarctica consists of the ancient Pre-

cambrian shield terrain of East Antarctica, on which is situated the South Pole. Some 2 to 3 kilometres (up to about 10,000 feet) thickness of continental ice covers the basement (bedrock) here, which is isostatically depressed. When the ice eventually melts the continent will rebound somewhat from the removal of the load. Thus, though some parts of the basement floor are below sea level today, they would not remain an archipelago after deglaciation.

Between the East Antarctic Shield and the fold mountains of the western part there is a belt of depressions with scattered hills, extending from the Ross Sea to the Weddell Sea. This is largely a rift zone, as indicated by the presence of volcanoes (Mt. Erebus, 12,450 feet [3,795 metres]) near McMurdo Sound. The lower parts of the depression are covered by extensive ice shelves.

THE EARTH'S OCEANIC REGIONS

Although oceanic regions cover 70 percent of the Earth's surface, very little was known about the floor of the ocean until the "Challenger Expedition" of the last century brought back a series of wire soundings that showed there were depths in the ocean at least as great as the Himalayas are high. Only with modern echo sounding techniques have the sea floors become well known.

Mid-ocean ridge system, deep-sea trenches, and other submarine landforms

Principal oceanic features. Just as the continents are united by the continuity of the youthful Alpide and Circum-Pacific fold mountain belts, there is a united "suture" around the globe in the form of the Mid-Ocean Ridge: the site of a high seismic (earthquake) activity, high heat flow from the Earth's interior, and youthful injections of basaltic lava. Its central rift is the site of the splitting apart of the ocean floor that is now referred to as sea-floor spreading (*q.v.*) and is going on at a rate ranging from 1 to 5 centimetres (0.4 to 2 inches) per year. Numerous transverse fracture zones displace the ridge distances of 50 to 500 kilometres (30 to 300 miles).

Topographically the ridge is a broad swell, 1 to 3 kilometres (0.6 to 1 mile) in elevation, 2,000 to 4,000 kilometres (1,200 to 2,500 miles) wide, and 55,000 to 65,000 kilometres (34,000 to 40,000 miles) long. It extends from the termination of the Verkhoyansk Range in eastern Siberia, crosses the Arctic Ocean, bisects the Atlantic Ocean, turns east to bisect the Indian Ocean, one branch disappearing into the Aden–Red Sea rift and the other continuing south of Australia, to cross the southeastern Pacific, and disappears into the Gulf of California.

The structural margins of continents do not terminate at the coastline, but are continued in the continental shelves. They pass at the break in slope (usually between 100 and 200 metres [300 and 600 feet] depth) into the continental slope, with a declivity of 3 to 6°. Below about 1,500 metres (5,000 feet) this passes into the continental rise, with a much more gentle slope, about 1:1,000.

Between the rise and the mid-ocean ridge there are relatively flat or low hilly regions at depths of 3,000 to 5,000 metres (9,800 to 16,000 feet). The flat areas are the abyssal plains and the hummocky sectors are called the abyssal hills. Deep rifts, often parallel to island arc sectors, are called deep-sea trenches or troughs (the less extreme ones).

Most of the ocean floor falls into one of these three major categories: mid-ocean ridge, continental margins, and abyssal depressions. Apart from these, there are certain less regular features—continental borderlands, oceanic rises and aseismic ridges, oceanic plateaus, and seamounts (a flat-topped seamount is a guyot). Certain erosional features (submarine channels and canyons) also characterize the shelf, slope, and rise areas.

Summary of oceanic physiography. The principal features of each ocean are described below.

Arctic Ocean. The Mid-Ocean Ridge crosses the sector north of Spitsbergen to the New Siberian Islands, separating the Nansen Basin on the south from the Eurasia Deep, which underlies the North Pole. Parallel to these in the trend of longitudes 40° W and 140° E (a great circle) is the Lomonosov Ridge, separated in turn by the Makarov Deep from the Alpha Ridge. From here almost to the Alaskan side is the broad Canada Basin, with smaller embayments in the Hyperborean Basin and Beaufort Deep, separated respectively by the Chukchi Cap and the Beaufort Plateau.

Atlantic Ocean. Bisecting the ocean symmetrically is the Mid-Atlantic Ridge; its northernmost sector is called the Reykjanes Ridge. In the Northern Hemisphere, to the west lie (from N to S) the Labrador Basin, Newfoundland Basin, North American Basin (with the Puerto Rico Trench), and the Guiana Basin. In the east there are the West Europe Basin, Iberia Basin, Canaries Basin, Cape Verde Basin, Sierra Leone Basin (with Romanche Trench), and Guinea Basin. In the Southern Hemisphere, on the west lie the Brazil Basin and Argentine Basin, and on the east the Angola Basin, Cape Basin, and Agulhas Basin. In the Southern Ocean sector south of the ridge (here called Atlantic-Antarctic Basin), the Sandwich Trench is in the west, adjacent to the Scotia Arc.

Principal ridges and plateaus on the west side are the Greenland-Iceland Rise, Labrador Rise, SE Newfoundland Rise, the Bermuda Rise, the Blake Plateau, the Pará Rise, the Rio Grande Rise, and the Falkland Plateau. On the east side are the Iceland-Faeroe Ridge, the Biscay Rise, the Azores Rise, the Azores Plateau, the Canary Island Rise, the Cape Verde Plateau, the Sierra Leone Rise, the Liberia Rise, the Guinea Ridge, Walvis Ridge, and Cape Rise.

Marginal seas to the Atlantic include the Greenland, Norwegian, North, Baltic, Mediterranean, and Black seas in the north and east, with the Gulf of Mexico and Caribbean Sea on the west, and the Scotia Sea and Weddell Sea in the extreme south.

Indian Ocean. The Mid-Indian Ridge bisects the ocean in an inverted Y pattern, with a deflection to the northwest as the Carlsberg Ridge. To the West are the Somali Basin, Mascarene Basin, and Madagascar Basin, with the Madingly Rise, Mascarene Plateau, Madagascar Ridge and Agulhas Plateau. East of the Mid-Indian Ridge are the Oman Basin, Arabian Basin, Ceylon Basin, Java Trench, Bali Trough, Timor Trough, and the Wharton Deep, North Australia (Argo) Basin, Cuvier Basin and West Australia Basin, South Australia Basin, and Diamantina Trench. Positive features include the Chagos-Laccadive Plateau, the Ninetyeast Ridge, Cocos Rise, Christmas Rise, Karma Rise, Roo Rise, Exmouth Plateau, Wallaby Plateaus, Broken Ridge, and Naturaliste Plateau. South of the mid-ocean ridge are the Valdivia or Enderby (eastern Atlantic–Indian-Antarctic) Basin, Kerguelen Basin, and the Wilkes (Indian-Antarctic) Basin. High features include the Crozet Rise, Amsterdam-St. Paul Plateau, Kerguelen Plateau, Banzare Plateau, and Gribb Plateau.

Marginal seas to the Indian Ocean include the Red Sea, Gulf of Aden, Persian Gulf, Andaman Sea, Timor Sea, Arafura Sea, and Gulf of Carpentaria.

Pacific Ocean. The mid-ocean ridge in the Pacific embraces a number of sectors—the southern one is the Pacific-Antarctic Ridge, the centre sector merges with the East Pacific (Easter Island) Rise, and the northern one with the Albatross Plateau. The Main or Central Pacific Basin has not been subdivided for the most part but occupies a vast area of the North Pacific, extending westward to a zone of trenches that borders the "Andesite Line" island arcs, including the Aleutian Trench, Kuril Trench, Japan (and Bonin) Trench, Mariana Trench, and Yap Trench; from north of New Guinea and the Bismarck Archipelago, the British Solomon Islands to Tonga, there are no trenches and the border is poorly defined. Instead, there are a series of *en echelon* troughs (including the Vitiaz) within the margin of the Melanesian Borderland. South of about 20° S is the South-western Pacific Basin. Two small basins near the western limits are identified as the West Caroline Basin and East Caroline Basin.

To the east of the mid-ocean rise, from north to south, are the Middle America Trench (Acapulco Trench and Guatemala Trench), and off South America the Peru-Chile Trench, with several intermediate basins, the Guatemala Basin, South-eastern Pacific Basin, Bellingshausen (Pacific-Antarctic) Basin, and Balleny Basin.

Numerous ridges occur in the Pacific; in the east there are the Cocos Ridge, Carnegie (Colón-Ecuador) Ridge, San Felix-Juan Fernandez Ridge, and South Chile Ridge, Emperor Seamounts, Marcus-Necker Rise, Christmas Island Ridge, Marquesas Ridge, Tuamotu Ridge, Society Ridge, Austral Ridge, Tokelau Rise, Phoenix Rise, Samoa Ridge, Ellice Ridge, Marshall-Gilbert Rise, Caroline-Solomon Rise, Ontong-Java Rise, and Eauripik-New Guinea Rise, Magellan Seamounts. In the southwest are the Chatham Rise and New Zealand Plateau, the Macquarie-Balleny Ridge, South Tasmania Ridge, and Mill Rise.

Marginal seas to the Pacific include the Bering Sea (with the Aleutian Basin), Sea of Okhotsk, Sea of Japan, the East China Sea, Yellow Sea, South China Sea, Philippine Sea (with Nansei Shoto or Ryukyu Trench and Philippine or Mindanao Trench), the Sulu Sea, Celebes (Sulawesi) Sea, Flores Sea, Savu Sea, Banda Sea, Ceram Sea, and Molucca Sea. In the Southwest Pacific are the Bismarck Sea, Solomon Sea (with New Britain Trench and San Cristobal Trench), the Coral Sea (with Coral Sea Basin, Queensland Plateau, New Hebrides Basin and Trench), Tasman Sea (with Lord Howe Rise, Norfolk Ridge, Norfolk Trough, Three Kings Basin and North Cape Rise), and Fiji Sea (with Melanesian Border Plateau, Pandora Basin, South Fiji Basin, Fiji Plateau, Lau Ridge, and many smaller features).

BIBLIOGRAPHY. Books of a general nature on this subject include: R.W. FAIRBRIDGE (ed.), *Encyclopedia of Oceanography* (1966), containing many articles on the oceans, their basins, and the general physiography of the submerged portions of the Earth, *Encyclopedia of Geomorphology* (1968), presenting separate articles on the many types of landforms that occur on the Earth and thus comprise its physiography; and R.F. FLINT, *Glacial and Quaternary Geology* (1971), covering the final molding of the Earth's surface by glacial events and processes during the Pleistocene Epoch. Fundamental data on the topography of the Earth is given in the article "Die Erdoberfläche" (the Earth's surface) by E. KOSSINNA in B. GUTENBERG (ed.), *Handbuch der Geophysik*, vol. 2 (1933).

(R.W.F.)

Earth, Structure and Composition of

An understanding of the constitution and origin of the Earth is derived from a synthesis of knowledge from many fields, including the basic sciences of physics, chemistry, and mathematics and such derivative fields as geology, geophysics, geochemistry, astronomy, and cosmogony. Progress in the earth sciences requires the rigorous application of fundamental sciences combined with a broad approach that enables the relationships between results obtained from several fields to be recognized and integrated.

Until about 1950 the fundamental fields of solid earth science were seismology and geology. The principal internal provinces of the Earth had been characterized from the determination of the way in which the velocities of compressional (*P*) waves and shear (*S*) waves varied with depth (Figure 1); compressional waves are characterized by volume, and hence density, changes and by particle motion in the direction of travel of the wave; shear waves are distortional secondary or transverse waves. Through calculations based on the constraints of moment of inertia (the tendency of matter to remain at rest or in uniform motion in the same straight line unless acted upon by an external force) and mass (the property of a body that causes it to have weight in a gravitational field), a close approximation to the distribution of density (Figure 2), pressure, and elastic properties throughout the entire interior had been derived. Geologists had been concerned primarily with reconstructing the history and evolution of the crust, a task in which they had been extremely successful. Progress in understanding the chemical and mineralogical nature of the mantle and core had been less successful.

Advances in data resources

Because of advances in several fields, this situation changed drastically after 1950. Studies of the abundances of elements in the Sun, in meteorites, and in deep-seated terrestrial rocks, together with new concepts regarding

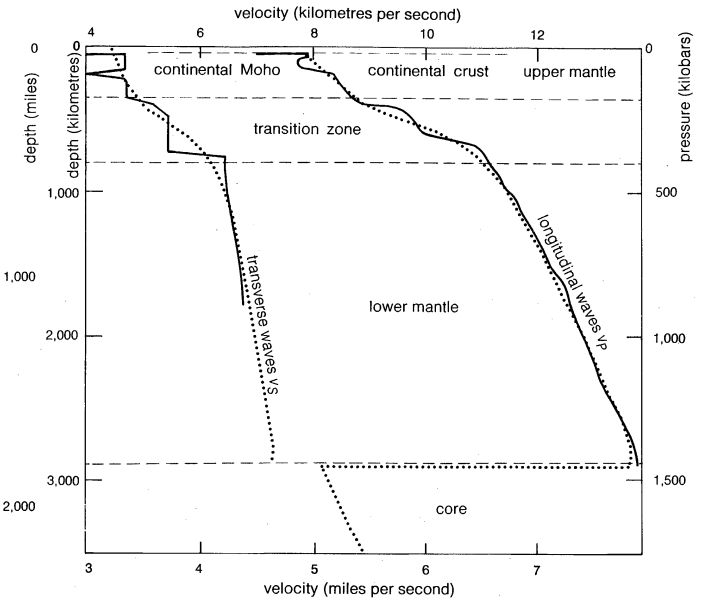


Figure 1: Seismic *P*- and *S*-wave-velocity distributions (v_P and v_S) and principal subdivisions of the Earth. Dotted lines show earlier data and full lines modern values. Subdivision of core occurs at 5,080 kilometres.

the origin of the Earth, placed important limitations upon the chemical composition of the Earth. Expanding geochemical research provided greater understanding of the distribution of elements within the Earth. Advances in knowledge of the elastic properties of silicates and oxides permitted the formulation of equations of state that were applicable to conditions in the Earth's deep interior and were employed to infer the occurrence of major phase transformations in the mantle. Development of experimental high-pressure–high-temperature techniques permitted the stability fields of mantle minerals to be studied at pressures corresponding to depths up to 600 kilometres (370 miles). Use of analogue systems provided information about the probable behaviour of silicates throughout the entire mantle. Developments in experimental petrology (study of the origin and natural history of rocks) made it possible to study under controlled conditions the physical chemistry of the processes within the Earth's interior that generate basaltic, andesitic, and other magmas (mobile rock materials from which igneous rocks form by solidification). Shock-wave methods resulted in the experimental determination of equations of states of silicates, oxides, and metals at pressures and temperatures up to those at the Earth's centre.

New developments in seismology, particularly the use of arrays of instruments and of nuclear explosives as sources of Earth vibrations, together with interpretation of surface-wave and free-oscillation data, resulted in greatly improved data on seismic-velocity distributions, including regional variations. That information, in turn, added further understanding of other important physical properties, such as density. Parallel advances were made in other branches of geophysics—*e.g.*, heat flow, gravity, electrical and magnetic properties, rheological properties (related to the flow of materials) and tectonophysics, which is concerned with forces responsible for movements and deformations in the Earth's crust. Manned exploration of the Moon and retrieval of lunar-surface samples, combined with space-probe studies of Mars and Venus, stimulated studies of the relationships between the Earth and other members of the solar system.

In the period between 1950 and 1970, agreement was reached on the answers to many long-standing fundamental questions regarding the Earth's internal constitution, and methods and techniques for the solution of other problems were developed. Development of theories of sea-floor spreading and plate tectonics (*i.e.*, study of broader structural features of the Earth) provided an integrated theory of the behaviour of the crust as recon-

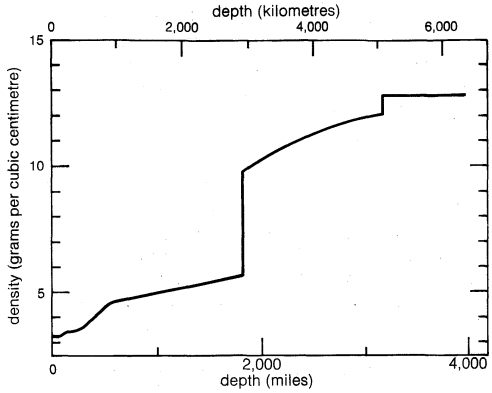


Figure 2: Density distribution within the Earth assuming that the upper mantle is composed of pyrolite.
From S.P. Clark and A.E. Ringwood, *Reviews of Geophysics* (1964)

Table 1: Dimensions and Properties of the Earth

Equatorial radius (a)	6,378.2 km
Polar radius (c)	6,356.2 km
Mean radius (r)	6,371 km
Volume V	1.083×10^{27} cm^3
Mass M	5.977×10^{27} gm
Mean density M/V	5.517 g/cm^3
Moments of inertia	
About polar axis (C)	8.043×10^{44} gm cm^2
Mean	8.025×10^{44} gm cm^2
Coefficient of moment of inertia $\left(\frac{C}{Ma^2}\right)$	0.3309
Equatorial gravity (g_e)	978.032 cm sec^{-2}
Ratio $\frac{\text{mass of Sun}}{\text{mass of Earth}}$	3.33×10^5
Ratio $\frac{\text{mass of Earth}}{\text{mass of Moon}}$	81.3
Mean distance from Sun	1.49×10^8 km
Mean distance from Moon	3.844×10^5 km
Mass of atmosphere	5.1×10^{21} gm
Mass of oceans	1.4×10^{24} gm
Mass of crust	2.4×10^{25} gm
Mass of mantle	4.1×10^{27} gm
Mass of core	1.9×10^{27} gm

structed by geologists, in terms of the dynamics of the mantle (see CONTINENTAL DRIFT; SEA-FLOOR SPREADING).

Three major regions of the Earth

Distribution of seismic velocities with depth divides the Earth into three principal regions: crust, mantle, and core (Figures 1 and 2). The crust is about 40 kilometres (25 miles) thick in continental regions and is highly heterogeneous, composed largely of rocks varying in composition between granite and basalt. Beneath the deep oceanic basins, the crust is about five kilometres (three miles) thick on the average and is believed to be predominantly of basaltic composition. The seismic *P*-wave velocities in the crust mostly vary from six to seven kilometres (about four miles) per second. Below the crust lies the mantle, which extends downward for 2,900 kilometres (1,800 miles). In most regions the crust is separated from the mantle by the Mohorovičić Discontinuity, or Moho, a narrow zone over which the seismic *P*-velocity increases suddenly to about 8.2 kilometres per second (5.1 miles per second). The mantle is solid—not liquid—and thus transmits both *P*- and *S*-waves. It is believed to be dominantly composed of magnesium silicates, with lesser amounts of iron, calcium, and aluminum, and is divided into a number of layers characterized by different types of mineral assemblages. At a depth of about 2,900 kilometres (1,800 miles), a major seismic discontinuity occurs. Below that level, *S*-waves are no longer transmitted, and *P*-wave velocity decreases from 13.7 to eight kilometres (8.5 to five miles) per second. The region below this discontinuity is known as the core and is believed to be predominantly composed of metallic iron in the molten state. At a depth of 5,080 kilometres (3,160 miles), a small increase of *P*-velocity occurs. This marks the boundary of the inner core that may be solid as the result of freezing of the iron under the high pressure existing (about 3,200,000 atmospheres).

This article treats the physical and chemical properties of the Earth's crust, mantle, and core, the structure of the Earth that is deduced from these properties, and the probable course of development of the Earth. For further information on the Earth's composition, see MINERALS; IGNEOUS ROCKS; METAMORPHIC ROCKS; SEDIMENTARY ROCKS; and ELEMENTS, GEOCHEMICAL DISTRIBUTION OF. For detail on the Earth's physical properties, see EARTH, MECHANICAL PROPERTIES OF; EARTH, MAGNETIC FIELD OF; EARTH, GRAVITATIONAL FIELD OF; EARTH, HEAT FLOW IN; and EARTHQUAKES. See also GEOCHEMICAL EQUILIBRIA AT HIGH TEMPERATURES AND PRESSURES; HIGH-PRESSURE PHENOMENA; and PHASE CHANGES AND EQUILIBRIA.

PHYSICAL PROPERTIES AND ZONAL STRUCTURE OF THE EARTH

Velocities of seismic P- and S-waves. Some important general physical properties of the Earth are given in Table 1. Variation of these properties within the Earth will be discussed in the following sections. The fundamental observational data required are the times of travel of seismic waves from sources (earthquakes, nuclear explosions) to seismometers spread over a wide range of distances. From these data, travel-time curves are constructed. In general, the farther the seismometer is from the source, the deeper the seismic rays penetrate into the Earth, carrying information about the integrated velocity variations along their paths. By analysis of the basic travel-time data, the variations with depth of *P* (compressional) and *S* (shear) waves are obtained (Figure 1). The fundamental seismic structure of the Earth was thereby explained before 1940, and the results provided the basis for subdivision of the Earth into crust, mantle, and core. In addition, the mantle was divided into three regions: upper mantle, transition zone, and lower mantle, each with a characteristic velocity–depth structure. Also, two distinct regions of the core were recognized.

Improvements in techniques of measurement

After World War II, a number of important advances were made, including the employment of improved long- and short-period seismometers and arrangements of seismometers, signal-filtering techniques, nuclear explosives as seismic sources (eliminating uncertainties in source location and times), and surface-wave-dispersion techniques, combined with data reduction by large electronic computers. Increased accuracy and separation of signals, including the use of second arrivals, resulted in important refinements of the existing velocity–depth curves. In particular, regional differences between velocity distributions beneath major geologic provinces were established; the widespread existence of a minimum in seismic velocities under oceans and active continental regions at depths of approximately 70–150 kilometres (40–90 miles) was confirmed; and the existence of at least two important seismic discontinuities in the mantle at depths of about 400 and 650 kilometres (250 and 400 miles) was established. Evidence of minor discontinuities and fine structure in the lower mantle and in narrow transition zones between core and mantle and between the outer and inner core was also obtained. The mean variation of *P*- and *S*-wave velocities with depth in the Earth and the principal internal subdivisions is shown in Figure 1.

Density distribution. In any region of the Earth that is chemically uniform, knowledge of the seismic *P* and *S* velocities immediately yields through what is called the Williamson–Adams equation the rate at which density varies with depth within the particular layer. In the 1930s it was demonstrated that, when this equation was applied in conjunction with the moment of inertia and mass of the Earth and with the density at the top of the mantle as estimated from other geologic-geophysical considerations, the density variation in most regions of the Earth could be defined within narrow bounds. Refinements have been proposed, based upon improvements in seismic-velocity distributions and determinations of the periods of the Earth's free oscillations; the most important modifications have been consideration of minimum average density between 70 and 150 kilometres (40 and 90 miles) and of the occurrence of rapid increases in density associated with the seismic discontinuities at 400 and 650 kilometres (250 and 400 miles). Whether the

mean density of the outer mantle is closer to 3.3 or 3.5 grams per cubic centimetre is not known, but the former value is probably the more nearly correct one. The density distribution within the Earth is shown in Figure 2.

Electrical conductivity. Magnetic storms caused by solar activity induce electric currents in the mantle. These currents in turn induce secondary magnetic fields at the Earth's surface. It has been demonstrated that the rate of decay of these secondary magnetic fields yields information on the electrical-conductivity distribution within the mantle. Most interpretations have had certain key features in common: an outer zone of relatively low conductivity extending to depths of a few hundred kilometres followed by a rapid increase in conductivity by factors of 100 to 1,000 at depths around 400–700 kilometres (250–450 miles). Additional considerations show that this rapid increase must flatten off at greater depths. When a more thorough understanding of the electrical conductivity of mantle minerals is obtained, the inferred conductivity distribution will be of value in determining the temperature distribution within the mantle.

Conductivity distribution within the mantle

Temperature distribution. Temperature distribution within the Earth is a function of heat flow within materials of varying composition and is related to the ultimate sources of heat within the Earth. These sources include residual heat from the time of the Earth's formation, heat from tidal friction (due to the gravitational interaction of the Sun, Moon, and Earth) and the braking effect of the Earth's rate of rotation, and energy emitted by radioactive decay of elements within the Earth. Radioactive decay can release substantial amounts of heat. It has been estimated that radiogenic heat production from granitic rocks in the Earth's crust is about 2.6×10^{-13} calories per gram per second, and that from basaltic rocks, which underlay most of the oceanic areas, about 3.8×10^{-14} calories per gram per second, on the average. The temperature distribution within the Earth is of critical importance to many aspects of geophysics, but at many depths it has not been determined, especially for those greater than about 400 kilometres (250 miles). Down to about 100 kilometres (60 miles), the approximate temperature distribution may be calculated from the theory of heat conduction, using as boundary conditions the amount of heat observed to be conducted to the surface of the Earth, the thermal conductivities of rocks, and the depth distributions of the heat-producing radioactive elements: uranium, thorium, and potassium. Between 100 and 200 kilometres (60 and 120 miles), the principal considerations are imposed by the experimentally observed melting behaviour of mantle silicate minerals at high pressures combined with knowledge that temperatures in this region must be high enough to explain the production of widespread basaltic magmas throughout geological time. Further limitations upon temperature are imposed by the necessity of explaining the occurrence of the low-velocity zone, believed to be caused by a small degree of partial melting.

Theories of heat distribution for greater depths must take into account not only the inferences that the outer core is mainly composed of liquid metallic iron and that the boundary between outer and inner core is caused by the freezing of iron but also large differences in estimates of the melting point of iron at very high pressures. These considerations suggest that the temperature at the base of the mantle may be about 3,000° C (5,400° F), whereas at the centre of the Earth the temperature may be in the vicinity of 4,000° C (7,200° F). The uncertainties in these estimates are probably ±1,000° C (2,000° F).

Rheological properties. Most rheological (flow) properties of materials such as viscosity, plasticity, strength, and creep rates are dependent upon the degree of proximity to melting temperature. The evidence on temperature distribution within the Earth and experimental data on the effect of pressure upon the melting temperatures of silicates imply that their melting point is approached most closely at depths between about 70 and 200 kilometres (40 and 120 miles). Accordingly, this zone could be expected to be the "weakest" in the Earth and one of high mobility that might take an important part in the dynamics of the mantle, continental drift, and adjustment mechanisms of isostasy, in which pressures from all sides are equal. Evidence suggests that this is indeed the case. The outermost cool shell, or lithosphere, some 70 to 100 kilometres (40–60 miles) thick, possesses substantial long-term strength and appears to slide relatively freely in the form of a series of large rigid "plates" over the underlying weak asthenosphere. Below 200 kilometres (120 miles), the melting-point gradient appears to increase more rapidly than the actual temperature gradient; in other words, the high-temperature material in the lower mantle is well below the melting point and probably has a much higher viscosity, or creep strength, than the region between 70 and 200 kilometres.

CHEMICAL AND MINERALOGICAL COMPOSITION OF THE EARTH

The continental crust. The thickness of the continental crust as defined by the depth to the Mohorovičić Discontinuity is usually between 35 and 45 kilometres (20 and 30 miles). Seismic *P*-velocities in the upper crust are usually between 5.8 and 6.4 kilometres (3.6 and four miles) per second and are believed to increase generally with depth; velocities in the vicinity of 6.6 to 7.2 kilometres (4.1 to 4.5 miles) per second are commonly inferred for the lower crust in stable regions. Higher velocities are found in some regions, however, particularly in those that have been the site of active mountain-building activity during the last 100,000,000 years. The nature of the downward velocity increase is not settled; in some regions there is evidence for a seismic discontinuity—the Conrad—whereas in others the increase is probably continuous.

Seismic *P*-velocities of 5.8 to 6.4 kilometres per second in the upper crust correspond to those of acidic igneous and metamorphic rocks and have led to the inference that below a veneer of sedimentary rocks, which may or may not be present, the upper crust or basement complex has an average composition approaching that of a granitic rock, more specifically granodiorite. This view is supported by the close agreement between both geological observations and geochemical estimates of the mean chemical composition (see Table 2). A granodiorite rock consists principally of the minerals quartz, feldspar, mica, and amphibole. Gravity investigations also lead to the conclusion that the upper crust has a density of approximately 2.70–2.75, which is consistent with the inferred composition.

Table 2: Chemical Compositions of Continental Shield Crystalline Surface Rocks and Pyrolite Model
(in weight percent)

component	A*	B†
SiO_2	66.4	45.4
TiO_2	0.6	0.3
Al_2O_3	15.5	3.5
Fe_2O_3	1.8	0.5
Cr_2O_3	—	0.4
FeO	2.8	8.1
MnO	0.1	0.1
MgO	2.0	37.6
NiO	—	0.2
CaO	3.8	3.1
Na_2O	3.5	0.6
K_2O	3.3	0.1
P_2O_5	0.2	0.1

*Continental shield crystalline surface rocks; composition A after Poldervaart. †Pyrolite model; composition B after Ringwood.

Large regions of the lower crust are characterized by *P*-velocities in the range 6.5–7.2 kilometres (four to 4.5 miles) per second, velocities similar to those of basalts, dolerites, and gabbros, and for many years it had been widely accepted that the lower crust is indeed composed of such rocks. Recent experimental investigations, however, demonstrate that gabbroic rocks are not thermodynamically stable under the pressure (five to ten kilobars)

and temperature (300°−700° C, or 400°−1,300° F) conditions of the lower crust. The experimental results, in conjunction with inferred densities and seismic velocities, suggest two classes of rocks for the lower crust: (1) in highly evolved regions, such as shields, the lower crust is probably highly heterogeneous and comprised on the average of groups of rocks of intermediate chemical composition (in granulitic or eclogitic metamorphic facies) that have been altered in composition, texture, or structure, by changes in pressure, heat, or chemical environment; (2) in crustal regions that have had a simpler and lower temperature history, the lower crust may consist dominantly of common rock-forming minerals called amphibolites.

Two classes of rocks in the lower crust

The mean chemical composition of the entire continental crust in most regions probably approaches that of the intermediate volcanic rock andesite, which is the most abundant volcanic rock erupted in many island arc regions. Geochemical investigations indicate that the crust is the result of extremely efficient chemical and magmatic differentiation processes that must have affected much of the mantle. It is estimated that between 30 percent and 60 percent of the uranium and barium originally present in the entire Earth has since become concentrated in the crust. This type of concentration is also displayed by some other elements and is in part attributable to large radii or ionic charge or both that inhibit the ready entry of such elements into the major minerals that comprise the mantle.

The ages of continental crustal rocks display a more or less continuous distribution from the present time to about 3,000,000,000 years. A few older ages beyond that figure—up to 3,800,000,000 years—have been recorded; but no rocks have been found with ages between 3,800,000,000 and 4,700,000,000 years, the estimated age of the Earth. Conditions on the Earth during the first 1,000,000,000 years of its history evidently differed in important respects from those in latter times. Increasing evidence suggests that the crust has formed more or less continuously from the mantle over geological time by magmatic differentiation processes and that perhaps this differentiation process did not commence until about 3,800,000,000 years ago.

Crustal profile of deep oceanic basins

The oceanic crust. The average crustal profile in deep oceanic basins consists of about 0.5 kilometre (0.3 mile) of sediments underlain by a layer of 1.5 to 2.5 kilometres (one to 1.5 miles), which has a compressional-wave (*P*-) velocity of about five kilometres (three miles) per second and is believed to represent a series of submarine flows of basalt characterized by substantial porosity. Next below is the main oceanic crustal layer with a thickness of about five kilometres (three miles) and mean *P*-velocity of 6.7 kilometres (4.2 miles) per second. Two hypotheses have been proposed about the nature of this layer: one holds that it consists of partly serpentinized peridotite; the other, that it consists of basalts, gabbros, and other mafic (composed largely of magnesium silicates) igneous rocks. Both concepts explain the observed seismic *P*-velocities, but the first does not satisfactorily explain the sharpness of the suboceanic Mohorovičić Discontinuity. In addition, the observed Poisson's ratio (ratio of the fractional transverse contraction to the longitudinal extension of a body under tensile stress) of serpentine (0.38) is much higher than that of the oceanic crust (0.27), which is, however, similar to that of mafic rocks. Most scientists favour the hypothesis that the oceanic crust consists of mafic igneous rocks, most likely a heterogeneous mixture of basalts, dolerites, gabbros, amphibolites, and greenschists.

According to the theory of sea-floor spreading, the oceanic crust is generated at midoceanic ridges (see Figure 4) and migrates away from the ridges for great distances, ultimately reaching oceanic trenches where it subsides and is consumed in the mantle. The crustal generation process consists of the partial melting of primary mantle material beneath midoceanic ridges forming basaltic magmas and leaving behind residual peridotite. The subsurface structure near the axes of ridges probably consists of a complex and highly heterogeneous mixture of mafic and ultramafic (peridotitic) rocks. Because of the relatively high temperatures that greatly weaken the crust in these regions, the large, highly dense peridotite bodies subside back into the mantle as the crust moves away from the ridge axis, and ultimately a relatively sharp boundary between crustal mafic rocks and mantle ultramafic rocks is established. This boundary constitutes the Mohorovičić Discontinuity.

The Mohorovičić Discontinuity. Beneath most stable continental regions the Mohorovičić Discontinuity is marked by a rapid increase in *P*-velocity, from 6.5 to seven kilometres (four to 4.4 miles) per second characteristic of the lower crust to values of eight to 8.4 kilometres (five to 5.2 miles) per second characteristic of the upper mantle. The detailed structural nature of the transition is not well understood and probably varies widely from place to place. In some regions, there is evidence of reflections from the Moho, indicating that the transition boundary is relatively sharp (about 0.3 kilometre, or 0.2 mile wide). In other regions, there is evidence of the existence of a transition zone. It is widely believed that in most regions, a large part of the increase of velocity from about seven to 8.2 kilometres (4.4 to 5.1 miles) per second occurs within a depth range probably less than five kilometres (three miles) thick. Observations of near vertical reflections and critical reflections combined with lack of evidence of the presence of a positive-velocity gradient in the main oceanic crust layer indicates that beneath deep oceanic basins the Mohorovičić Discontinuity is usually a relatively sharp feature, less than one kilometre (0.5 mile) thick and sometimes less than 0.3 kilometre (0.2 mile).

The nature of the Moho—two hypotheses

Two hypotheses regarding the nature of the Moho have been widely discussed. The first maintained that the Moho was caused by an isochemical phase transformation of a gabbroic (pyroxene + plagioclase) lower crust into an eclogitic (jadeitic pyroxene + garnet) upper mantle. As eclogite has a density of about 3.5 grams per cubic centimetre, or some 15 percent greater than that of gabbro, it could provide the required seismic velocity for the upper mantle. This hypothesis attracted attention largely because of the possibility that a phase change at the Moho, involving the reversible transformation of gabbro to eclogite, could be responsible for vertical movements of the Earth's crust and other tectonic phenomena. The second hypothesis maintained that the Moho represents a sharp change in the chemical makeup of the material above and below the Moho—from an intermediate or mafic lower crust into an ultramafic mantle composed of peridotite (olivine + pyroxene). Peridotite, with a density of 3.3 grams per cubic centimetre, could also provide the observed mantle seismic velocities.

The phase transformation hypothesis was ultimately tested by means of laboratory investigations of the gabbro–eclogite transition. The experiments indicated that a transitional mineral assemblage (garnet granulite) intervened between the stability fields of gabbro and eclogite. As gabbro became transformed through garnet granulite to eclogite, pressures ranged from three to 12 kilobars, with densities and seismic velocities distributed throughout this zone. Furthermore, gabbros with only slightly different chemical compositions displayed very large differences in the pressures required to cause transformation. Considering the foregoing characteristics of the gabbro–eclogite transformation, it is impossible to explain the relative sharpness of the Moho in most continental and oceanic regions. Observed relationships between surface-heat flow and crustal thickness were also contrary to expectations based upon the phase transformation hypothesis. In most continental and oceanic regions where it is clearly defined, the Mohorovičić Discontinuity cannot be explained in terms of a reversible chemical change from gabbro to eclogite; it is accordingly considered that the discontinuity is caused by a difference in chemical composition between crustal rocks and the underlying mantle rocks. The composition of the mantle rocks is discussed below.

The upper mantle. Seismic velocity distributions (see Figure 1) provide the basis for subdividing the mantle

The three regions of the mantle

into three regions. The upper mantle extends to a depth of 400 kilometres (250 miles) and is characterized by regional differences in velocity profiles and by generally low-velocity gradients except in the vicinity of the low-velocity zone beneath oceans and some geologically unstable regions. The transition zone extends downward from about 400 to 1,000 kilometres (250 to 600 miles) and is characterized on the average by rapid increase of velocity with depth. There are at least two major seismic discontinuities, at depths of about 400 and 650 kilometres (250 and 400 miles). The lower mantle extends between 1,000 kilometres and 2,900 kilometres (600 and 1,800 miles) and is characterized by relatively gentle and uniform increases of velocity with depth, although high-resolution seismic studies provide some evidence of minor irregularities in this region also. On the basis of seismic velocities, this threefold subdivision of the mantle is likely to reflect corresponding variations in other physical and chemical properties and must be explained in terms of the physicochemical properties and thermodynamic stability of the minerals present.

Compressional-wave velocities. Seismic *P*-velocity in the uppermost mantle usually falls within the range 8.2 (±0.2) kilometres, or 5.1 (±0.1) miles per second. This property, together with certain broad petrological and mineralogical limitations, indicates some combination of olivine, pyroxene, garnet and, in restricted regions, amphibole. The two principal rock types carrying these minerals are peridotite (olivine-pyroxene) and eclogite (pyroxene + garnet). Both types may carry some amphibole. These relationships have given rise to the alternative hypothesis that the upper mantle is of peridotitic or eclogitic composition; choice between these alternatives would require accurate knowledge of the density of the upper mantle. The densities of peridotitic rocks, which might qualify as major upper mantle constituents, are close to 3.3 grams per cubic centimetre, whereas eclogites with seismic velocities of 8.2 kilometres (5.1 miles) per second possess densities close to 3.5 grams per cubic centimetre. Geophysical evidence on the density of the upper mantle is, however, somewhat conflicting. Interpretations of gravity observations combined with the theory of isostasy (which holds that all large portions of the Earth's crust are balanced as though floating on a denser underlayer) are considered to favour a density closer to 3.3 than 3.5 grams per cubic centimetre, and thus the less dense peridotitic composition. Observations demonstrating that in the oceanic upper mantle seismic velocity varies substantially with azimuth or orientation also support the peridotite conclusion, because such a phenomenon (anisotropy) is characteristic of many peridotites but much less pronounced, or absent, in eclogites. Some estimates of density of the upper 200 kilometres (120 miles) of the mantle, obtained by inversion of free-oscillation data for the Earth, on the other hand, indicate a density of about 3.5 grams per cubic centimetre, suggestive of eclogitic composition; but conflicting results have been obtained by the method.

Conflicting geophysical evidence on density

Composition of the upper mantle. Geological and petrological observations provide more definitive information. An important indication of the composition of the upper mantle is provided by the study of xenoliths (inclusions) in kimberlite pipes, cylindrical igneous bodies. In Africa and Siberia, where the pipes have been studied most intensively, they carry numerous xenoliths of crustal rocks that they are known to have intruded during rapid upward emplacement. Such pipes also contain many inclusions of rocks not known to occur in the vicinity, particularly peridotites and eclogites. It is generally believed the latter xenoliths are from the mantle and represent a random sample of mantle rock types that have been intersected by the pipes. The occurrence of diamonds (evidently formed under great pressures) in the pipes and in some of the inclusions implies that the pipes are derived from depths of at least 150 kilometres (about 90 miles); occurrence of other high-pressure phases suggests some pipes come from depths as great as 300 kilometres (about 180 miles). In nearly all cases in which sampling has been adequate peridotitic inclusions were found to be much more common than eclogitic inclusions —strong indication that the upper mantle sampled by the pipes is predominantly of peridotitic composition, with eclogite a minor but widely distributed constituent.

Alpine-type peridotites

Investigations of the occurrence and origins of large bodies of so-called Alpine-type peridotites emplaced in the Earth's crust are a second source of information on the composition of the upper mantle. Peridotite rock intrusions, some covering several thousand square kilometres, have been found in New Caledonia, New Guinea, the Philippines, Cyprus, Greece, British Columbia, and many other localities. Alpine-type peridotites are characteristically intruded close to the axis of maximum deformation along mountain-building belts and island arcs. It is known that intrusion is frequently controlled by major faults, which may extend for hundreds of miles and almost certainly extend into the mantle.

Dredging on the mid-Atlantic and mid-Indian oceanic ridge systems has also revealed widespread occurrences of peridotites. Great thicknesses of these rocks have been discovered in fracture zones that offset the mid-Atlantic Ridge.

The nature and significance of Alpine-type peridotite occurrences have been considerably clarified by development of the theories of sea-floor spreading and plate tectonics. The outcrop of large bodies of peridotites along midoceanic ridges, combined with geophysical observations of mantle-type seismic velocities at shallow depth beneath some ridge localities, strongly indicates that such peridotites represent actual mantle material intruded to high levels. Similarly, detailed mapping of peridotite and associated gabbro and basalt occurrences in New Guinea, Cyprus, and Greece provides evidence that relatively complete sections of oceanic crust and underlying peridotitic mantle have been overthrust into the crustal environment. Other peridotite occurrences are believed to have originated at midoceanic ridges and been carried into their present localities in island arcs (*q.v.*) and continental margins by the migrating sea floor.

The occurrences of many Alpine-type peridotites, as interpreted in the light of plate tectonic theories, provide strong indication that these rocks are representative of the uppermost suboceanic mantle, whereas there is no comparable evidence that eclogites have comparable significance. Eclogites are rare compared to peridotites and occur in very small bodies. According to the results of detailed studies, many eclogites found in the crust have been formed there by metamorphism of mafic rocks and do not represent tectonic slices of the upper mantle.

Basaltic magmas have been erupted abundantly over both oceanic and continental regions of the Earth's surface throughout geologic time. The source of basalt is believed to be the upper mantle, of which a fundamental property is, clearly, that it should be capable of yielding basalt magmas when subjected to suitable partial melting processes. Because they are generally so low in potassium, uranium, thorium, barium, strontium, lanthanum, and other trace elements—as well as in the major elements of sodium, calcium, and aluminum—the compositions of Alpine peridotites cannot yield basalt magmas by partial melting. Rather the compositions are strongly suggestive of refractory residues remaining after partial melting processes in which basaltic magmas have been produced.

From such considerations it can be concluded that below the residual layer of Alpine-type peridotite must exist a more primitive rock that is capable of yielding basalt magmas by partial melting, leaving behind a residual unmelted peridotite. The composition of this hypothetical primitive material, called pyrolite, must lie somewhere between the compositions of peridotite and basalt. Several types of evidence indicate that the primitive composition is closer to that of peridotite and is perhaps not far from a mixture of three parts peridotite to one part of basalt (see Table 2). The exact nature of the composition is not known, however, and it is thought that it may vary substantially from place to place. Experimental investigations on the crystallization of basaltic magmas at high pressures and the formation of basalts by partial

melting of pyrolite have led to a generally consistent interpretation of physicochemical relationships between types of basalt and their complementary peridotites.

The above considerations lead to the concept of a chemically zoned upper mantle consisting of a layer of residual peridotite underlain by pyrolite. Beneath oceans the peridotite layer is believed to be about 70 kilometres (40 miles) thick; beneath highly evolved continental regions such as Precambrian shields the thickness would be closer to 150 kilometres (90 miles).

Low-velocity zone. In many oceanic regions, the upper 70 kilometres (40 miles) of the upper mantle exhibits an *S*-wave (shear-wave) velocity of about 4.6 kilometres (2.9 miles) per second. Below this "lid," velocity decreases rather sharply, reaching a minimum of about 4.3 kilometres (2.7 miles) per second at about 100 kilometres (60 miles). This zone of decreased velocity extends to about 150–200 kilometres (90–120 miles), and a corresponding low-velocity zone is also believed to occur for *P* (compressional) waves. The low-velocity zone is more strongly evident in some regions of recent mountain-building activity, as in the Alps. In other active regions (*e.g.*, Japan, or the Basin and Range Province of the United States) the 70-kilometre (40-mile), high-velocity lid may not be present, and the low-velocity zone may extend upward to the Moho. Beneath stable continental areas such as Precambrian shields, a low-velocity zone is probably not present for *P*-waves, and the *S*-minimum is much less pronounced, occurring at greater depths.

A number of possible causes of the oceanic and tectonic low-velocity zones have been considered, and it is believed that the observed magnitude of the velocity change together with the inferred velocity gradients indicates a small degree of partial melting in the region. The presence of as little as 1 percent of liquid in the form of thin films along grain boundaries would cause a marked decrease in both *P* and *S* velocities and also the increase in attenuation of seismic waves that is also observed. Experiments lead to the suggestion that partial melting is caused by the effect of water, believed to be present in the mantle in amounts of about 0.1 percent. At depths less than 70 kilometres (40 miles) the water is principally held in the mineral amphibole; at greater depths, however, amphibole becomes unstable because of the high pressure, and the water is liberated as a free phase. As a result, the temperature at which melting begins is lowered by 200°–400° C (350°–700° F), and a small amount of interstitial melt occurs between 70 and 150 kilometres (40 and 90 miles). Below that depth other minerals capable of incorporating water in their structure become stable, and hence the water-vapour pressure is reduced, so that melting does not occur.

Partial melting in the low-velocity zone has important implications for the dynamic behaviour of the mantle; the low-velocity zone is a region of very low mechanical strength and high mobility. The less pronounced *S*-wave low-velocity zone beneath Precambrian shields is not attributed to partial melting; it is believed to be the result of the temperature gradient in the region, which is not, however, critical enough to affect *S*-waves.

Phase transformations in the upper mantle. An important property of rocks near the pyrolite composition is that they may crystallize in four distinct mineral assemblages characterized by the coexistence with olivine of amphibole, plagioclase, aluminous pyroxenes, and garnet. These mineral assemblages have different physical properties, with densities and seismic velocities varying by about 4 percent; accordingly, their distribution has a significant effect upon the fine structure of the upper mantle. The stability ranges of the different types of pyrolite have been determined, leading to a general understanding of the conditions governing their occurrence.

Mineral assemblages and structure

Garnet pyrolite is stable in the upper mantle beneath Precambrian shields; pyroxene pyrolite probably occurs as a relatively thin layer in the uppermost mantle beneath oceanic and tectonically active regions, and it transforms into garnet pyrolite at depths of 60–100 kilometres (35–60 miles) beneath these regions. Garnet pyrolite is characterized by the highest velocity and density and remains stable between depths of approximately 70 and 350 kilometres (40 and 200 miles); no further phase transformations occur in this region. Regions of occurrence of plagioclase and amphibole pyrolites are limited in the uppermost suboceanic mantle; plagioclase is restricted to conditions of low pressure and high temperature such as, for example, immediately beneath midoceanic ridges.

The transition zone. The region of the mantle between 400 and 1,000 kilometres (250 and 600 miles) below the Earth's surface is called the transition zone and is characterized in general by an exceptionally high rate of increase in seismic velocities with depth and has at least two major seismic discontinuities (see Figure 1). It was demonstrated in 1936 that the density in the transition zone increased much more rapidly with depth than would be expected in a uniform self-compressed layer. A detailed study of the elastic properties of the mantle in 1952 led to the conclusion that the abnormal changes of elastic properties and density in this zone were probably caused by a series of major phase transformations; that the principal minerals of the upper mantle (olivine, pyroxenes, and garnets) were not stable under the high pressures existing, and that those minerals transformed into a new set of denser polymorphs in which the atoms were more closely packed and different in crystal structures.

A series of experimental investigations between 1956 and 1966 confirmed 1952 inferences that olivines, pyroxenes, and garnets were unstable at high pressures, transforming into denser polymorphs. A systematic study of these transformations in silicates using both direct, static, high-pressure techniques (up to 200 kilobars), and indirect methods utilizing thermodynamic studies provided indications of the probable series of phase changes occurring throughout the transition zone.

Assuming that the transition zone composition is similar to that of the upper mantle (pyrolite), a number of transformations are to be expected (see Figure 3). Near 400

From A.E. Ringwood, *Earth and Planetary Science* Letters (1972), 14, pp. 233-241

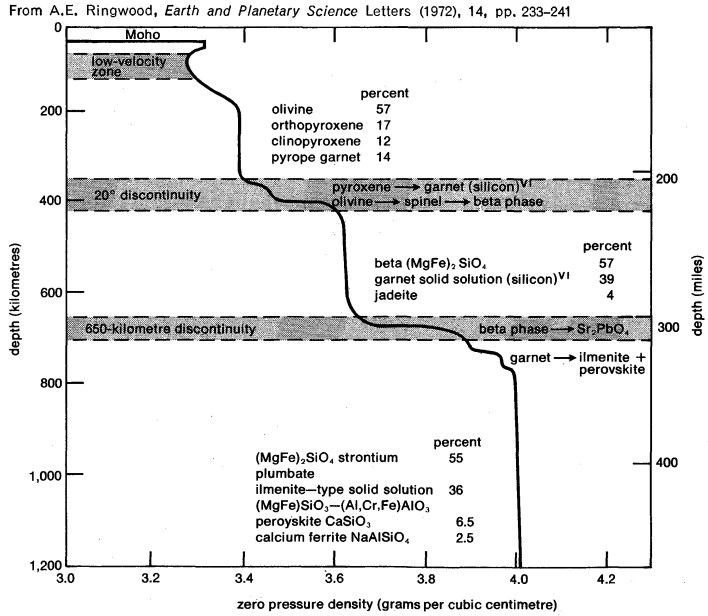


Figure 3: Mineral assemblages and densities given by pyrolite model in outer mantle.

kilometres (250 miles), two major transformations occur: (1) the pyroxene component of the upper mantle becomes unstable and transforms to a new type of garnet structure, in which about one quarter of the silicon atoms are surrounded by six atoms of oxygen; (2) at slightly greater depths, olivine transforms into a spinel-like structure. These transformations would result in an increase of density of about 8 percent and an accompanying increase of about 0.7 kilometre (0.4 mile) per second in seismic velocity, which is in close agreement with observed seismic velocity changes at the 400-kilometre (250-mile) discontinuity.

Two major transformations near 400 kilometres

Between about 450 and 600 kilometres (300 and 400 miles) the rate of increase of seismic velocities is not indicative of further phase transformations. Experimental studies reveal that the spinels and garnets formed near 400 kilometres (250 miles) remain stable through the interval to 600 kilometres (400 miles). Direct techniques are unable to reproduce pressures and temperatures occurring below about 600 kilometres; detailed interpretation of phase changes at greater depth rests heavily upon studies of analogue compounds and shock-wave methods. Such studies reveal that garnets and spinels will transform to denser phases at pressures of about 200–300 kilobars; the spinel phase may transform to the strontium plumbate structure or it may disproportionate (*i.e.*, convert itself into two derivatives) into a mixture of periclase and ilmenite structure. The garnet phase probably disproportionates into a mixture of ilmenite and perovskite structure. Minor phases that are probably present possess the calcium ferrite structure and hollandite structures.

These transformations are unlikely to occur at exactly the same depth. By far the most important transformation would be of magnesium silicate, Mg_2SiO_4 (spinel-like), to the strontium plumbate structure, which may well occur within a relatively narrow depth range and could be responsible for the occurrence of the 650-kilometre (400-mile) seismic discontinuity (Figures 1 and 3). Transformation of garnets to ilmenite and perovskite structures involves the disproportionation and formation of very complex solid solutions, probably smeared out more than the strontium plumbate transformation; it would be coincidental if both were to occur at exactly the same pressure. Apparently, they contribute to a rather complex fine structure between 650 and 1,000 kilometres (400 and 600 miles) that has yet to be resolved by seismology.

From the crystallographic standpoint the principal characteristic of transformations occurring between 400 and 1,000 kilometres (250 and 600 miles) is the replacement of typical upper mantle silicate minerals by a new assemblage of phases in which the structure is octahedral (each silicon atom being surrounded by six oxygen atoms) instead of the original tetrahedral structure (each silicon atom being surrounded by four oxygen atoms). Closer packing of the atoms is primarily responsible for the 20 percent increase in density (referred to zero pressure) associated with the transformations. The mean density of the assemblage of silicate polymorphs with ilmenite, perovskite and strontium plumbate structures, which is thought to occur at depths of 800–1,000 kilometres (500–600 miles), is similar to that of a mixture of oxides of magnesium, iron, calcium, aluminum, and silicon ($MgO + FeO + CaO + Al_2O_3 + SiO_2$), provided the density of SiO_2 is taken as that of the high-pressure, rutile-type polymorph of SiO_2, "stishovite," of octahedral structure.

Increases in densities with transformations

The lower mantle. The general nature of the phase transformations responsible for the large increases in density, seismic velocities, and other physical properties in the transition zone of the mantle, between 400 and 1,000 kilometres (250 and 600 miles), is fairly well understood. Further transformations to denser states of materials in the lower mantle must also be considered possible. Although the rate of increase of seismic velocities in the lower mantle had been considered similar to that to be expected for self-compression of homogeneous material without chemical or phase changes, seismic velocity distributions show regions of anomalously high gradients near depths of 1,000, 1,230, 1,540, and 1,910 kilometres (600, 760, 960, and 1,190 miles). About 15 percent of the total velocity increase between 1,000 and 2,300 kilometres (600 and 1,400 miles) can be accounted for by those regions, and it is possible that minor-phase changes could be the cause. Further phase transformations in the deep mantle over the pressure range of 300 to 1,300 kilobars are not unlikely in view of the number of transformations that occur at pressures of less than 300 kilobars. Nevertheless, the density distribution and elasticity of the deep mantle are considered to permit only small density changes.

Studies of the density distribution of the lower mantle and of shock-wave data on the densities of silicates at comparable pressures indicate the density of the lower mantle to be about 5 percent higher than expected for the pyrolite composition crystallizing as a mixture of phases with ilmenite, strontium plumbate, perovskite, and periclase structures. A possible explanation is, as already suggested, that further minor transformations have resulted in the formation of a mineral assemblage intrinsically denser than the assemblage above it. An alternative explanation is that the lower mantle is characterized by a slightly higher iron/magnesium ratio than the upper mantle and is thus denser. Combined experimental and seismic investigations have developed a generally consistent and satisfying explanation of some of the major physical properties of the mantle in terms of the natures and structures of the minerals present and their stability fields.

When due allowance is made for the effects of phase transformations, the concept of pyrolite composition for the upper mantle explains the observed physical properties of the transition zone and lower mantle; it also suggests that major changes in chemical composition do not occur throughout the mantle. This appears explicable if the mantle had been rather well "stirred" and homogenized during the process of core-segregation early in the Earth's history; strong cosmochemical arguments can, in fact, be made in support of the view that the composition of the entire mantle may approximate that of pyrolite.

Concept of pyrolite composition of upper mantle

Seismic velocities in the lower 200 kilometres (120 miles) of the mantle may decrease slightly with depth, as might be caused by the presence of metallic iron. It has been suggested that the core and mantle may be out of chemical equilibrium and that chemical reactions near the core–mantle boundary may result in precipitation of metallic iron in the region.

The core. The boundary of the core is marked by a major seismic discontinuity close to 2,900 kilometres (1,800 miles), at which *P*-wave velocity drops from 13.7 kilometres (8.5 miles) per second to approximately eight kilometres (five miles) per second and below which *S*-waves are no longer transmitted (see Figure 1). The latter observation implies that the rigidity of the outer core is zero (*i.e.*, that it is liquid), a conclusion that is reinforced by tide investigations and theories of the origin of the Earth's magnetic field. The seismic discontinuity at 2,900 kilometres (1,800 miles) is accompanied by a corresponding discontinuity in density, which changes from 5.7 grams per cubic centimetre in the lowermost mantle to approximately 9.8 grams per cubic centimetre in the outermost core (see Figure 2).

The dominant component of the Earth's magnetic field can be demonstrated to originate within the Earth. Intensity of the field varies from place to place in a rather irregular manner, and in addition there are gradual changes ("secular variation") over periods as short as ten years. The short time scale of these changes makes it improbable that they originate in the solid mantle, and it is believed that their source lies in the liquid region of the core. It has been demonstrated that fluid motions in the core in the presence of an initial magnetic field may lead to a self-sustaining dynamo action, whereby a relatively stable system of electric currents is induced within the core. These currents in turn generate a magnetic field, part of which leaks to the Earth's surface to provide the observed field and part of which interacts with the moving liquid to sustain the dynamo action.

The Earth's magnetic field and the liquid region of the core

Originally, it was proposed that the motions of the liquid outer core were convective in nature, caused by heating from a high concentration of radioactive elements in the inner core. As the required distribution of radioactive elements was improbable on geochemical grounds, however, a modified theory was proposed. It held that convection was driven by cooling at the core–mantle boundary accompanied by freezing at the boundary of the inner core, which liberated its latent heat, thus providing the required energy.

The dynamo theory of the Earth's magnetic field implies that the electrical conductivity of the outer core is very

high and similar to that of a metal, and this property together with the high density suggests that the core may be composed principally of liquid iron. Cosmochemical evidence on the relative abundances of iron in the Sun and in the most primitive chondritic meteorites supports such a hypothesis. The extraterrestrial abundances are similar to the relative abundance of iron in the Earth, providing the core is assumed to consist mainly of iron (see further, EARTH, MAGNETIC FIELD OF).

Composition of the core

All available evidence thus supports the hypothesis that the core is principally composed of metallic iron, but it is probable that additional components are also present. Studies of meteorites and cosmic-abundance tables indicate that the core probably contains about 10 percent of metallic nickel. Careful comparisons of the density and elasticity of the core with those of pure iron, as provided by shock-wave investigations at core pressures, indicate that the density of the core is 10–15 percent less than that of pure nickel–iron under comparable pressure and temperature conditions, whereas the seismic velocity is substantially higher than that of pure iron. Such observations suggest the presence of a third component in the core, capable of decreasing density and increasing the velocity; on geochemical grounds the most likely component is sulfur or elementary silicon. About 20 percent of either in solution in the core would explain the observed densities and velocities, and would have interesting implications for the processes by which the core was formed, and indeed for the origin of the Earth.

At a depth of about 5,100 kilometres (3,200 miles) the seismic *P*-velocity increases by about 10 percent (see Figure 1). That depth marks the boundary of the inner core. From limitations on the density of iron at the centre of the Earth—as indicated by shock-wave data, experiment, and calculations—it can be demonstrated that the inner core is probably substantially rigid, which would imply that it is solid. The inner core may thus be regarded as being caused by the freezing of iron under very high pressures (approximately 3,000–5,000 kilobars).

Detailed seismic studies reveal the presence of a rather complex fine structure at the boundary of outer and inner cores and, further, that at least two seismic discontinuities may occur within a transition zone several hundred kilometres thick. The structure of the transition zone may be a consequence of chemical inhomogeneity and phase changes during the crystallization of a relatively pure iron–nickel inner core from a less pure liquid containing larger amounts of a third component (sulfur or silicon) in solution.

DEVELOPMENT OF THE EARTH'S STRUCTURE AND COMPOSITION

The origin of the Earth is one of the oldest problems of natural philosophy, one that has captured the interest not only of scientists from many different disciplines but also of mankind generally. Although the subject remains controversial in many respects, a great deal has been learned since 1950, largely through the consideration of geochemical and cosmochemical evidence that has accompanied the expansion of knowledge in these fields. A new dimension of data has been added by the retrieval and study of lunar-surface samples.

Theories of the origin of the Earth

Prior to 1950 the most popular theories of the origin of the Earth and other planets were based on cataclysmic events, such as removal of material from the hot Sun by some means; for instance, tidal interactions from a star passing nearby. In addition to the improbability of such processes, it was recognized that many intractable geochemical problems followed from the assumption that the Earth had condensed entirely out of gaseous solar material.

Astronomers had developed theories of star formation by accretion processes in vast interstellar clouds of dust and gas. Such concepts were extended to the Earth, suggesting that it, too, had accreted from dispersed solid matter within a primordial solar nebula. Many of the processes occurring during accretion fall in the realm of chemistry rather than physics, and many of the key boundary conditions that must be satisfied by an acceptable theory of formation of the Earth are geochemical in nature. Several hypotheses have been developed in attempts to satisfy these and other boundary conditions. Three distinct categories can be recognized and are discussed below.

Homogeneous accretion of the Earth. The various theories of formation of the Earth as a result of the accretion of metal and silicate particles in the solar nebula have certain essential elements in common. It is assumed that a flattened solar nebula of dust and gas (mainly hydrogen and helium) first developed some 4,600,000,000 years ago, then passed through a high-temperature stage, during which the dust component recrystallized as a mixture of metallic iron and silicate particles. Some physical fractionation process (magnetic fields is one suggestion) then caused varying degrees of separation of metal particles from silicate particles, so that the metal/silicate ratio varied in different parts of the nebula. (This assumption is necessary to explain the differing densities of the planets.) The metal and silicate particles present at this stage are believed to resemble those found in chondritic meteorites.

The Earth then is assumed to have accreted from the mixture of metal and silicate particles in a localized region of the nebula. Accretion occurred slowly enough (over 10^8 years) so that enough energy was dissipated by radiation to allow formation of the Earth in a cool, unmelted state at a mean temperature of probably less than 1,000° C (1,800° F). Radioactive elements (potassium, uranium, thorium) then slowly heated the Earth until partial melting occurred in the interior, perhaps 10^9 years after accretion. When this stage was reached, the metal particles, which had previously been somewhat uniformly distributed throughout the Earth, then melted and collected together, ultimately sinking and segregating to form the core; differentiation and degassing of the mantle, leading to the gradual formation of the crust, hydrosphere, and atmosphere, accompanied the melting process.

This hypothesis offers no satisfactory explanation for the initial physical fractionation of metal particles from silicate particles within the solar nebula. The accretion of the Earth from an intimate mixture of small metal and silicate particles would have led to a state of local chemical equilibrium (in which the products formed by the reaction themselves react with each other to reform the original reactants) between the two phases. But the Earth's mantle seems to contain too much nickel and other siderophile elements (those that have weak affinity for oxygen and sulfur and go readily into solution in molten iron); it also is too highly oxidized (as shown by the amount of ferric iron present) to have been in equilibrium with metallic iron before the core separated. It is difficult, too, to understand the trapping of water and carbon dioxide in the mantle (subsequently to be liberated into the atmosphere and hydrosphere), because these compounds would be mostly reduced to hydrogen, carbon, and carbon monoxide in the presence of metallic iron at high temperatures. Some recent experimental measurements on the partition of lead between metallic iron and silicates, combined with knowledge of the isotopic composition of lead in the upper mantle-crust system, have shown conclusively that the core must have segregated about 4,600,000,000 years ago, close to the time of accretion of the entire Earth. Formation of the core soon after accretion would have required temperatures far higher than are consistent with the "cool accretion" hypothesis.

Inhomogeneous accretion of the Earth from metal and silicate components. The above considerations have resulted in substantial modifications of the homogeneous accretion hypotheses. One such modification that has received some consideration is the proposal that accretion of the Earth was a sequential process occurring over a wide range of temperatures: iron core first at high temperatures; silicates of the mantle at a later stage; the hydrosphere, atmosphere, and other volatiles at the terminal stage at low temperature. According to this version, silicate components did not accrete upon the Earth until formation of the metallic core was nearly complete, so

Sequential cooling of Earth components

that the core thus became surrounded by a silicate mantle; and at the lowest temperatures, the material condensing or precipitating from the solar nebula became more oxidized and rich in volatiles, including water and carbon dioxide. The latter material was plastered onto the outside of the mantle at the latest stage of accretion and ultimately provided the Earth's atmosphere and hydrosphere.

The early formation of the Earth's core and the lack of chemical equilibrium between the upper mantle and core are accounted for by this hypothesis, which, however, still has many shortcomings. For example, if the Earth's core condensed from the solar nebula at a temperature of about 1,500° C (2,700° F) and if subsequent accretion of the mantle (depleted in the heat-producing radioactive element, potassium) occurred at lower temperatures, then the temperature at the centre of the Earth after accretion was not much higher than 2,000° C (3,600° F). Under the prevailing pressure, the core would have been well below the melting point throughout, and there does not appear to be any way in which subsequent heating of the deep interior could occur at a sufficient rate to cause the melting of the outer core some 2,000,000,000 years later, a rate that would be required to explain the generation of a magnetic field as inferred from the magnetizations of Precambrian rocks.

Furthermore, if the Earth's mantle condensed sequentially from a gaseous solar nebula, the first components to condense would be calcium oxide and aluminum oxide, which should form a seismically distinct layer several hundred kilometres thick near the base of the mantle; there is no evidence of such expected seismic zoning. The upper mantle would be expected to be strongly depleted in calcium, aluminum, and other involatile elements; but the compositions of widespread basalts derived from the upper mantle show that such is not the case.

Single-stage accretion of the Earth from primordial material. In addition to the specific shortcomings mentioned, both of the preceding hypotheses assume that complex chemical and physical fractionation processes occurred in the solar nebula before the Earth accreted. An attempt to avoid these assumptions and difficulties proposes that the terrestrial planets accreted directly from primitive solid material resembling the Type 1 carbonaceous chondrites in a cool solar nebula; that there were no complex high-temperature physical and chemical fractionations in the nebula prior to accretion.

Type 1 carbonaceous chondrites are believed to be the most primitive type of meteorites known and to contain the primordial or solar abundances of most elements, except those that are highly volatile. All of the iron occurs in the oxidized state, which is to be expected under conditions of thermodynamic equilibrium in a nebula of solar composition at low temperature. Furthermore, the meteorites contain about 20 percent of water, up to 10 percent of complex organic compounds, and large quantities of other relatively volatile components. In many respects the compositions of these meteorites resemble those expected for the primitive dust particles in a low-temperature solar nebula.

It has been suggested that the Earth, Mars, and Venus accreted directly from planetesimals of this composition and that, during accretion, carbonaceous compounds reduced the associated oxidized iron to form metallic iron. This process occurred to varying extents on the three planets, and it is believed that the differences in planet densities were primarily caused by differences in oxidation states rather than in the relative abundances of metals. On this model, Mars is more oxidized than the Earth, and Venus is intermediate.

Assuming that accretion of the Earth occurred in a period of a few million years or less, the chemical equilibria in the accreting material were largely controlled by the liberation of gravitational potential energy as the primitive planetesimals fell upon the Earth. This energy amounted to 10,000 calories per gram for the entire Earth, but it was not liberated uniformly. Gravitational energy per gram would have been zero at the start of accretion, increasing as the Earth grew, and reaching more than 15,000 calories per gram during the final stages of accretion. Early stages of accretion would have occurred under relatively cool conditions, permitting the trapping of volatiles and oxidized iron deep within the interior; temperature would have risen during later stages because of the very large amounts of gravitational potential energy dissipated; carbon reacted with oxidized iron to form metal and to complete the degassing that was taking place. The degassing resulted in the formation of a large primitive atmosphere, perhaps one-fifth the mass of the Earth and composed mainly of carbon monoxide and hydrogen, with perhaps about 10 percent of volatilized silicates and metals.

According to this model, the Earth accreted in a state grossly out of chemical equilibrium, the temperature in the outer region exceeding the melting point during the final stages and permitting metallic iron to segregate into large bodies that then sank toward the centre. The process of core segregation is highly exothermic (*i.e.*, formed with the evolution of heat), releasing an additional 600 calories per gram of gravitational potential energy, enough to heat the entire Earth by 2,000° C (3,600° F). Segregation of the core would have been too rapid for complete chemical equilibrium between core and mantle to be attained, and both would have been homogenized by convection after formation of the core.

Model for simple chemical development

The foregoing model appears to be capable of satisfying some important chemical and physical boundary conditions for the Earth's formation, specifically the early formation of the core, implying a high-temperature Earth immediately after accretion and the lack of chemical equilibrium between mantle and core. The hypothesis provides a ready explanation of the densities of Mars and Venus, within the framework of a rather simple chemical development of the solar nebula prior to accretion, but does not provide a satisfactory explanation of how the massive primitive atmosphere was lost from the Earth. A combination of two causes has been suggested for that loss: (1) rotational instability and (2) the effects of intense solar particle radiation when the Sun passed through a very young (T-Tauri) phase. Whether those mechanisms were capable of removing the postulated primitive atmosphere is not yet known; future investigations may be expected to provide a crucial test of the single-stage hypothesis. As a by-product of this hypothesis, it was suggested that the formation of the Moon may perhaps be linked with the loss of the massive, hot primitive atmosphere, that on expanding and cooling the silicate components of the atmosphere would be precipitated to form a ring of planetesimals orbiting the Earth, ultimately becoming unstable and coagulating to form the Moon.

Evolution of crust from mantle. Much geochemical and petrological evidence suggests that the crust has evolved gradually from the mantle. The theory of seafloor spreading, combined with developments in experimental petrology, has begun to provide an understanding of the processes of crust formation.

Pyrolite diapirs (folds of pyrolite produced by a rising, mobile core material; see Figure 4) rise from the low-velocity zone beneath midoceanic ridges and undergo partial melting to produce basaltic magmas and residual peridotite. The basaltic magmas segregate and crystallize to form a heterogeneous oceanic crust composed mainly of basalt, gabbro, amphibolite and greenschists. The oceanic crust and underlying ultramafic (*i.e.*, composed dominantly of magnesian rock-forming silicates) lithosphere moves outward from the ridge, ultimately reaching a trench near an island arc or continental margin. The lithosphere then subsides into the mantle.

A second stage of partial melting of the former oceanic crust, possibly caused by frictional dissipation along a band of earthquake foci known as the Benioff zone, may then occur. Experiments indicate that the products of this second stage of partial melting resemble the dominant volcanic rocks in most island arc regions, the compositions of such rocks being similar to average continental rocks. A parallel differentiation process occurs in the wedge between the Benioff zone and the surface. In this

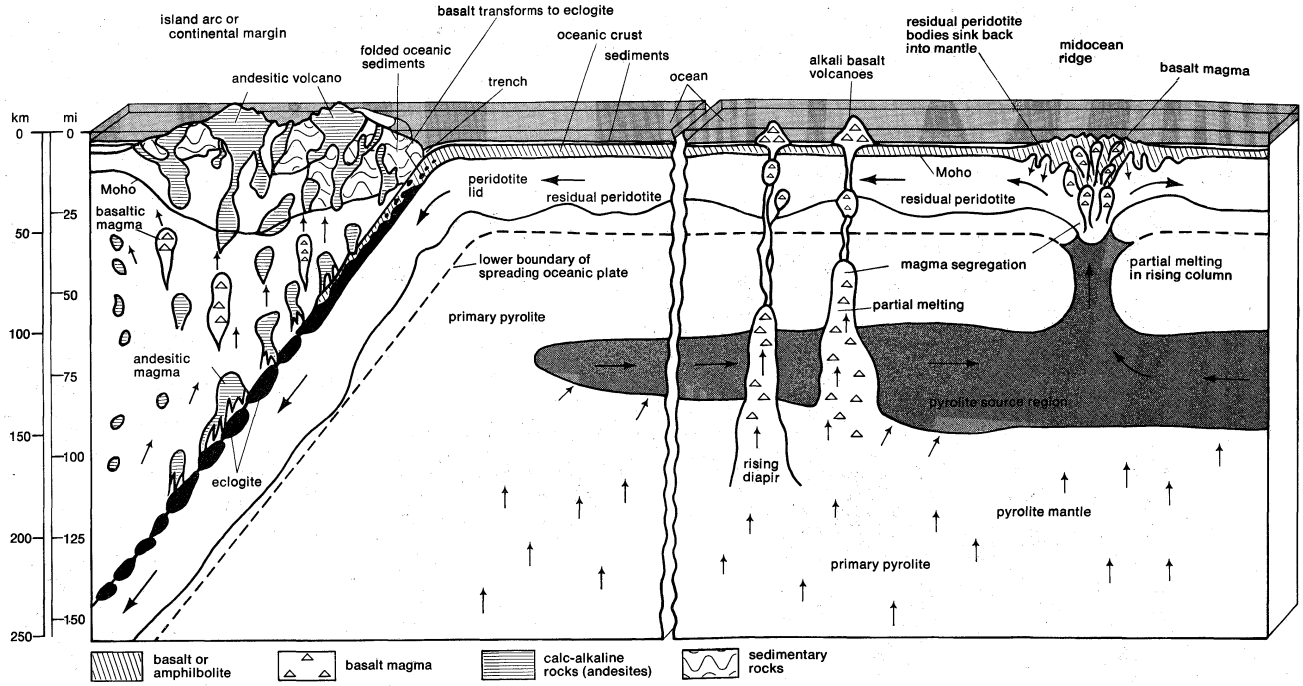


Figure 4: Petrological model of sea-floor spreading, differentiation of the mantle, and crustal evolution.
From A.E. Ringwood, *Geophysical Monograph*

region, water carried down into the mantle in the form of hydrous minerals from the oceanic crust is released by dehydration and enters the overlying pyrolite, lowering the melting point and initiating the production of basalt magmas, which under wet conditions may also differentiate toward andesitic compositions.

It is seen that formation of crustal rocks from the mantle proceeds via two distinct processes characterized by vertical and lateral differentiation, respectively. Vertical differentiation involves partial melting of the mantle and transport of the low-melting components into the crust immediately above. The result is the formation of a residual ultramafic zone in the mantle, a process that has reached its maximum development in Precambrian shield regions. In lateral differentiation oceanic crust is formed at midoceanic ridges; the differentiated mantle and crust is then transported, perhaps for thousands of kilometres, to a trench, where the material sinks into the mantle and remelts at depth to produce andesites.

There is no evidence that crustal rocks were formed during the first 1,000,000,000 years of Earth history, perhaps because the basic sea-floor spreading process did not occur during this period. It is significant that the formation of continental crust entails at least two major stages of petrological differentiation, whereas the formation of oceanic crust is a single-stage process.

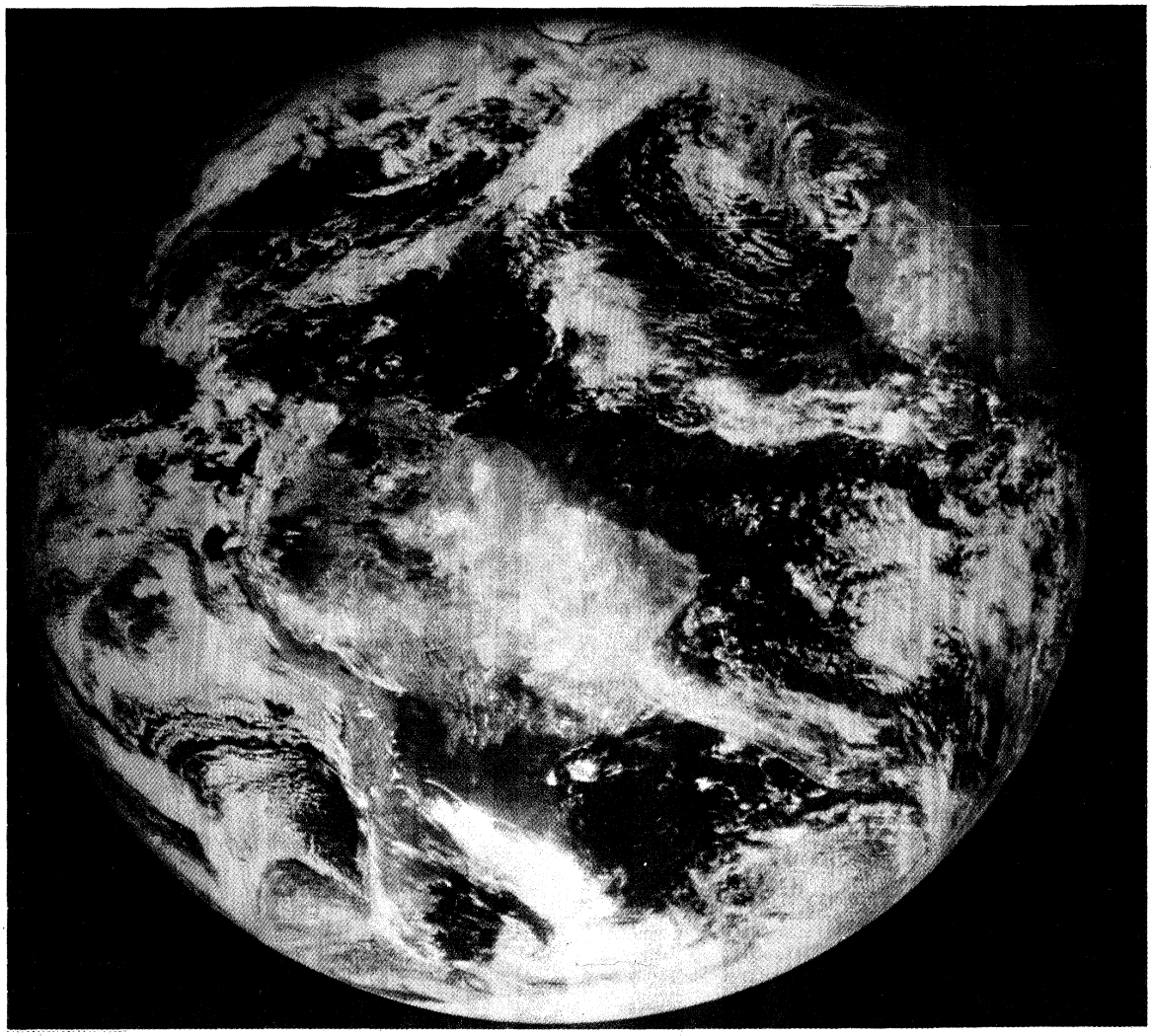
It is apparent from Figure 4 that, petrologically speaking, the sea-floor spreading process may be regarded as being one of irreversible differentiation of an initially homogeneous mantle. The principal products of this differentiation process are the continental crust, on the one hand, and the residual and complementary peridotite and eclogite; on the other. The peridotite and eclogite sinks deeply into the mantle and displaces relatively unfractionated material upward into the low-velocity zone, the immediate source region for the midoceanic ridges. It is improbable that the residual peridotite and eclogite ever participate in further magma generation. It has been estimated that perhaps 30 to 60 percent of the Earth's total uranium and barium has been segregated into the Earth's crust by differentiation processes, and it can thus be argued that a comparable proportion of the Earth's mantle has so far been subjected to the irreversible process of differentiation involved in crustal evolution.

BIBLIOGRAPHY. Useful reviews of the physical properties and internal structure of the Earth are given in the following books: K.E. BULLEN, *An Introduction to the Theory of Seismology*, 3rd ed. (1963); T.F. GASKELL (ed.), *The Earth's Mantle* (1967); P.J. HART (ed.), *The Earth's Crust and Upper Mantle* (1969); B. GUTENBERG, *Physics of the Earth's Interior* (1959); H. JEFFREYS, *The Earth*, 4th ed. (1962); and F.D. STACEY, *Physics of the Earth* (1969). An additional key reference is, "Elasticity and Constitution of the Earth's Interior," *J. Geophys. Res.*, 57:227–286 (1952). The following books contain chapters devoted to several aspects of the chemical and mineralogical composition of the Earth: T.F. GASKELL, *op. cit.;* P.J. HART, *op. cit.;* A. POLDERVAART, *Crust of the Earth* (1955); A.E. RINGWOOD and D.H. GREEN (eds.), *Phase Transformations and the Earth's Interior* (1969); J.S. STEINHART and M.A. TUVE (eds.), *The Earth Beneath the Continents* (1966); and P.J. WYLLIE (ed.), *Ultramafic and Related Rocks* (1967). Papers by S.P. CLARK and A.E. RINGWOOD, "Density Distribution and Constitution of the Mantle," *Rev. Geophys.*, 2:35–88 (1964); D.H. GREEN and A.E. RINGWOOD, "The Genesis of Basaltic Magmas," *Contr. Miner. Petrol.*, 15:103–190 (1967); and A.E. RINGWOOD and D.H. GREEN, "An Experimental Investigation of the Gabbro-Eclogite Transformation," *Tectonophysics*, 3:383–427 (1966), deal with topics less adequately covered in the texts. Principal hypotheses relating to the origin and early evolution of the Earth are advanced in the following references: P.M. HURLEY (ed.), *Advances in Earth Sciences* (1966); H.C. UREY, *The Planets* (1952); F. BIRCH, "Speculations on the Earth's Thermal History," *Bull. Geol. Soc. Am.*, 76:133–154 (1965); A.E. RINGWOOD, "Chemical Evolution of the Terrestrial Planet," *Geochim. Cosmochim. Acta*, 30:41–104 (1966).

(A.E.R.)

Earth as a Planet

The Earth is the third planet in distance outward from the Sun. Many consider the Earth–Moon system to be a unique double planet because the Moon is much larger in comparison with its primary, the Earth, than most other planetary satellites. In the more usual view, the Earth has a single natural satellite, the Moon. Like some other planets of the solar system, the Earth has radiation belts, an ionosphere (part of the atmosphere containing electrically charged particles, or ions), an atmosphere, and a lithosphere, or crust (outer surface); beneath the crust, the deep interior is divided into layers called the mantle and the inner and outer core (see Figure 1). Much more is known about details of the Earth's structure and composition than about those of the other planets because of its accessibility. The Earth has, however, been much changed and differentiated locally because of its active and turbulent history, and,



The Earth from 22,300 miles in space. The National Aeronautics and Space Administration's ATS III satellite transmitted this colour photograph to a ground station at Rosman, North Carolina. It was received at 10:30 AM EST on November 10 from its on-station position at 47° west longitude on the Equator over Brazil. Four continents (portions of North and South America, Africa, and Europe and the Greenland ice cap) can readily be seen, while the Antarctic continent is blanketed under cloud cover. Major weather over the central United States, stretching from the Great Lakes to Mexico, represents a cold front moving eastward. A tropical storm (bottom centre) with a cold front extending into Argentina can be seen.
By courtesy of the National Aeronautics and Space Administration

thus, care is needed in making interpretations. Processes such as volcanism, mountain building, and sea-floor spreading are still going on.

The curvature of the Earth can be demonstrated from ground level in many simple ways—for example, by measurements of Eratosthenes of Cyrene in antiquity (see below). Photographs from rockets in the 1950s clearly showed the curvature of the edge of the Earth (at the horizon), as did later photographs taken from the space vehicle Gemini 4. In the 1960s, for the first time, the Earth was photographed from sufficient distance to show it as a planet, more beautifully coloured and more varied than any other. (See accompanying photograph.) In this picture, cloud patterns show clearly the scale of atmospheric circulation. Below the clouds, the oceans appear blue and the outlines of continents can be seen. The Earth's oceans are probably unique among the planets because of the small temperature and pressure range at which liquid water can occur.

Like other planets, the Earth is approximately a sphere. It spins on its axis (the line joining the points at the surface called the north and south poles) once a day relative to the Sun. This is equivalent to once in about 23 hours 56 minutes 4 seconds relative to the stars (the sidereal day). The sidereal day is shorter than the solar day (24 hours 3 minutes 57 seconds) because the Earth, revolving at the same time about the Sun, must rotate farther, after completing the sidereal day, before the solar day is completed.

This article presents an overview of the Earth as a whole, ignoring the details of its surface features, the oceans, and the atmosphere. Those topics are treated in the articles EARTH, PHYSIOGRAPHY OF; OCEANS AND SEAS; and ATMOSPHERE. For additional information on the Earth's interior, surface configuration, physical properties, and chemical composition see EARTH, STRUCTURE AND COMPOSITION OF; EARTH, FIGURE OF; EARTH, MECHANICAL PROPERTIES OF; and ELEMENTS, GEOCHEMICAL DISTRIBUTION OF. See also the separate articles treating the Earth's gravitational and magnetic fields and its heat flow. The history of the Earth is summarized in the article EARTH, GEOLOGICAL HISTORY OF and the origin and nature of the Earth's principal surface features are treated in MOUNTAIN-BUILDING PROCESSES; SEA-FLOOR SPREADING; CONTINENTAL DRIFT; and OCEANIC RIDGES, among others.

BASIC DATA

Figure of the Earth. The radius from the centre to the poles is not quite the same as that from the centre to a point on the Equator (see the Table), but the concept of the Earth as a sphere is a good approximation. The distance a traveller has to proceed northward in order to make the elevation of the Pole Star increase by one de-

Earth Data

a (equatorial radius)	6,378.188 km (3,963.221 mi)
b (polar radius)	6,356.784 km (3,949,921 mi)
R (mean radius)	6,371 km (3,959 mi)
f (flattening)	0.033523
e (eccentricity)	0.081813
Circumference, equatorial	40,075.51 km (24,902.45 mi)
M (mass)	5.976×10^{27} g
V (volume)	1.083×10^{27} cu cm
Mean density (M/V)	5.517 g per cu cm
Mass ratios: Sun to Earth to Moon	333,400 : 1 : 0.01228
Mean distance from Sun	1.496×10^{8} km (92,900,000 mi)
Earth's speed in its orbit	29.8 km/sec (18.5 mi/sec)
Revolution around Sun	365.25 days
Rotation	23 hr 56 min 4.09 sec
Area	5.096×10^{8} sq km (197,000,000 sq mi)
Land	1.48×10^{8} sq km (57,000,000 sq mi; 29% of Earth)
Highest point	9,521 m (29,028 ft; Mt. Everest)
Lowest point	−10,912 m (−35,800 ft; Marianas Trench)
Lowest point on land	−397 m (−1,302 ft; Dead Sea)

gree is the length of one degree of latitude. The length of a degree of latitude depends slightly on the latitude at which it is measured. This dependence indicates that there is a deviation of the Earth from spherical shape. Sir Isaac Newton was the first to demonstrate that the centrifugal force of the Earth's rotation would tend to make the Earth bulge out at the Equator, giving it the shape of an oblate spheroid (the surface formed by an ellipse when it is rotated about its short axis). Such a surface is flatter near the pole than near the Equator, and one degree of latitude is therefore longer in high latitudes than it is in low ones. Gravity also varies over the Earth's surface, primarily with latitude, being somewhat higher by about 0.5 percent at the poles than at the Equator.

Rotation and precession. Because the Earth revolves around the Sun in 365.25 days, the sidereal day (the exact length of which is 23 hours, 56 minutes, 4.099 seconds of mean solar time) is shorter than the conventional mean solar day by about one part in 365. Astronomical observations show that the Earth is not an ideal timekeeper. The length of the day is subject to certain tiny changes. The most important are (1) a gradual slowing down of the Earth's rotation—*i.e.*, lengthening of the day; (2) small annual or seasonal variations; (3) small irregular variations. The magnitude of the gradual slowing down found from ancient astronomical observations, especially from ancient eclipses and from the more accurate astronomical data available since about 1600, amounts to an aggregated time change of about 30 seconds per century. If this phenomenon is assumed to have existed throughout the Earth's lifetime, the Earth must once have rotated much more rapidly than it does now. The cause of the slowing down has not been established with complete certainty, but most probably it is due to the effect of tidal friction in the oceans. The loss of angular momentum (that is, momentum acquired through rotation) by the Earth is compensated by that gained by the Moon's orbit, which then moves farther away from Earth. The seasonal fluctuations in the length of the day are tiny, of the order of some thousandths of a second per year; they can be explained by changes in the size of the polar ice caps and changes in atmospheric circulation. The irregular changes are equally tiny and can apparently be traced to motions in the Earth's core.

Factors in the changing length of the day

Lacking any external influences, the direction of the Earth's axis of rotation in space would remain the same for all time. Gravitational forces of other members of the solar system, in particular the Moon and the Sun, however, influence the Earth. Because the Earth is flattened, the Moon's gravity tends to tilt the Earth's axis, so that it becomes perpendicular to the Moon's orbit, and to a lesser extent the same is true for the Sun. A spinning body, such as the Earth or a gyroscope, subject to such a force will not, however, yield to it, but the axis of rotation will precess—*i.e.*, move around the surface of a cone in space. This motion of the Earth's axis is known as the precession of the equinoxes (times of equal day and night in spring and autumn). It was discovered by the astronomer Hipparchus in 120 BC and first theoretically explained by Sir Isaac Newton in the 17th century. Precession implies that the pole is not always in the same place among the constellations but moves about; for instance, 12,000 years hence it will coincide with the star Vega. The length of one period of the precession is 26,000 years. The precession makes the Sun's orbit in the heavens (the ecliptic) and its intersection with the celestial equator (the equinoxial points) move in such a direction as always to meet the Sun, at a rate of 50.26″ per year. This change amounts to about 29° since Hipparchus' time; hence the ancient Greek's description of the Sun's yearly path across the constellations no longer fits. The force tending to alter the tilt of the axis is not constant; when the Sun or Moon is crossing the Equator, for instance, it has no effect. The path of the pole is therefore not exactly a circle but contains a superimposed smaller oscillation, the nutation, with a period of 18.6 years and an amplitude of 9.2 seconds of arc. The Chandler wobble, another gyroscopic effect of extremely small amplitude and a period of 436 days, is accompanied by a slight deformation of the Earth's body. The solid body of the Earth undergoes a slight daily tidal deformation caused by the gravitational action of the Moon and Sun, about 25 centimetres (10 inches) at most.

The Chandler wobble

Numerical data. Some quantities relating to the Earth are known with a precision exceeding one part in 1,000,000. Others are known only to within a much lower error limit, and those derived from the latter have corresponding uncertainties. The one most important quantity of the former type is the constant of universal gravitation, $G = 6.67 \times 10^{-8}$ ($cm^3\ g^{-1}\ sec^{-2}$) units, known to about one part in 2,000. The numbers with many digits given in the Table might some day have to undergo revision in the last one or two decimal places, but they are internationally agreed upon as the most likely values.

STRUCTURE OF THE INTERIOR

Data from seismic waves. Two phenomena provide information about the deeper layers of the Earth's interior: seismic waves generated by earthquakes (*q.v.*) and, to a lesser degree, the Earth's magnetic field. There are two main types of seismic waves, the P-waves (primary, because they travel fastest) and the S-waves (secondary, because they travel more slowly and, hence, arrive later at the station than the P-waves). The P-waves are compressional (longitudinal) in nature, whereas the S-waves are shear (transverse) waves. The existence of two types of waves greatly increases the amount of information about the Earth's interior.

Types of seismic waves

The most conspicuous feature of the Earth's interior is the existence of a sharp boundary at which reflection and refraction occur and which encloses the part known as the core; it represents an extremely well-defined discontinuity. The mean radius of the core (ignoring flattening) is 3,470 kilometres (2,156 miles), or nearly 55 percent of the Earth's radius. The most outstanding feature of the core is that it never transmits S-waves. Because the failure to transmit transverse waves is a distinctive char-

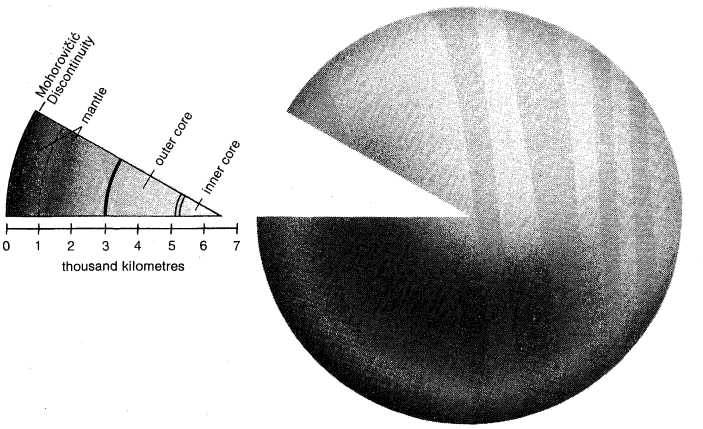


Figure 1: Main layers of the Earth.

acteristic of liquids, it is generally accepted that the core material is in a liquid state, probably mainly of molten iron. Existence of the core was first quantitatively established in 1913. Later studies have shown an additional boundary, somewhat diffuse, which encloses the so-called inner core. Its radius, not precisely known, is about 1,300 kilometres (800 miles). Evidence indicates that the inner core is solid iron (solidified under extreme pressures).

Seismic velocities, density, pressure. In the mantle and the core a progressive stiffening of the material with increasing pressure makes the velocity of seismic waves increase at greater depths. The density distribution cannot be uniquely determined from such data, but certain rather restrictive conditions are imposed on it thereby. A further condition results from a knowledge of the moment of inertia of the Earth about its axis, known with great accuracy from a combination of astronomical and geodetic observations. The moment of inertia about an axis through the centre of a homogeneous sphere, for which m is the total mass and a is the radius, is 0.4 ma^2. Observations show that the moment of inertia of the Earth is 0.334 ma^2, which indicates a rather pronounced increase of density toward the central parts.

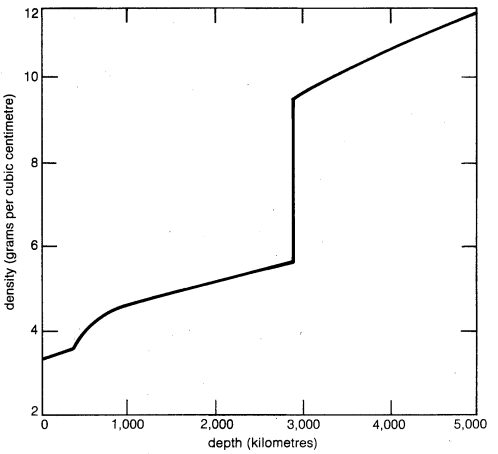


Figure 2: Known density distribution in the Earth down to the limit of the inner core.

The pressure inside the Earth rises rather regularly with increasing depth. In the mantle the rise is at a rate of about 470 atmospheres per kilometre (one atmosphere of pressure equals 14.7 pounds per square inch). The pressure at the boundary of the core is 1,370,000 atmospheres; at the centre of the Earth it is 3,700,000 atmospheres. The strength of gravity is also a function of the depth; it clearly must vanish at the centre. Gravity is remarkably constant throughout the mantle, to within 1 or 2 percent; but, beginning at about the boundary of the core, it decreases steadily to zero.

ORIGIN, COMPOSITION, AND AGE

Formation of the Earth. Astronomers have speculated about the origin of the solar system ever since it was suggested in the 18th century that the stars and planets arose from the condensation of vast gaseous masses filling cosmic space. The formation of the Sun may have occurred much earlier than that of the planets. It is now generally accepted that the planetary system developed out of a disk-shaped mass of gas surrounding the Sun. The original cosmic matter consisted mainly of hydrogen (and some helium), and most of these gaseous components must have been lost to space early from the disk. The less volatile matter condensed, mainly as silicates and metallic iron, these being the prime constituents of present-day meteorites as well as of the interior of the Earth. At the time the gas escaped, the solid matter probably consisted of bodies much smaller than the present planets, because the larger bodies would hold on to a gaseous envelope by gravitation, at least to some of the gases heavier than hydrogen. Most probably, the original solid components were dust grains.

While this picture is crude and speculative in its details, it brings out one crucial point common to all modern theories as distinct from the views held in the 19th century: the original matter from which the planets condensed must have been relatively cold. The temperature in this primary lens-shaped body of gas and dust was probably quite low, because it should readily have been cooled by radiation that could escape in a direction perpendicular to the plane of the disk. When dust congregated to form the Earth, gravitational energy was released and could be transformed into heat, but it is not known what fraction was available for actual heating, because this amount depends critically on the speed of aggregation. If the aggregation was slow, most heat should have escaped by radiation, and this is the most likely case.

Chemical composition. The composition of the Earth's crust can be investigated directly by quantitative chemical analysis of samplings from many rocks. The pioneer in this field was V.M. Goldschmidt, who in the early decades of the 20th century undertook the painstaking task of determining the average proportions of the chemical elements in the Earth's crust. To appreciate the significance of the results it is necessary to compare them with what is known about the chemical composition of cosmic objects other than the Earth. Such objects as meteorites fall upon the Earth from outer space. Meteorites are generally considered to be fragments of a larger solar system body or bodies. Many chemical analyses of meteorites have been made. Next, quantitative determinations of the relative proportions of the chemical elements can be made for the Sun by spectroscopic methods. Finally, though with greater limitations, the same can be done for other stars.

One conclusion long drawn from the vast mass of observations is that the gross chemical composition of matter in the universe is relatively uniform. There might be small variations from one star to the next, but the impression of overall uniformity is strong. Scientists are thus led to think of a definite gross composition of primordial matter from which the Earth and the other planets were formed. The figures given below for the relative proportions (known as cosmic abundances) of the most important elements are from a classical compilation by H. Suess and H.C. Urey (1956). Since the abundances are only relative, they are expressed as the numbers of atoms existing per one atom of a reference substance. The latter is taken to be the element silicon, since rocks consist largely of compounds of silicon. These abundances are as follows:

hydrogen	40,000.0	aluminum	0.09
helium	3,100.0	silicon	1.00
carbon	3.5	phosphorus	0.01
nitrogen	6.6	sulfur	0.37
oxygen	21.5	argon	0.15
neon	8.6	calcium	0.05
sodium	0.04	iron	0.60
magnesium	0.91	nickel	0.03

Remarkably enough, only six of the remaining elements have abundances that lie between 0.001 and 0.0001; all other elements are so rare that they exist only in minutest traces on the cosmic scale.

Location of concentrations of elements in the Earth's layers

On the whole, the lighter compounds are more concentrated in the crust, including oxides and more complex compounds of sodium, magnesium, aluminum, and calcium that form many important constituents of common rocks. Geologists are able to discriminate between rocks of the surface layers and others that have been pushed up from lower layers and should therefore resemble more closely the chemical composition of the Earth's mantle. On the basis of such data it is commonly believed that the Earth's mantle consists primarily of certain compounds (silicates) containing oxygen combined with silicon, magnesium, and iron, three of the most common nonvolatile elements. Sulfur most probably exists in the interior of the Earth in the form of iron sulfide. Geochemical evidence, however, indicates that sulfur is far less abundant on the Earth than the overall cosmic abundance figure given above.

That the core is most probably composed of molten iron is a generally accepted view held for many years; lately the presence of a little sulfur and, in addition, some silicon has been made extremely probable; also there is a small amount of nickel. The view that the core consists of iron fits all the known physical and chemical facts. It is in agreement with the seismic data, and because the melting point of iron at high pressures is much lower than that of the compact silicates, it explains the appearance of a sharp discontinuity at the core's boundary. Iron is by far the heaviest of the common nonvolatile elements; hence it will try to find its way toward the more central parts of the Earth.

Composition of the core

It is at present fairly well established that the inner core consists of solid iron or iron–nickel in almost pure form.

Radioactive age determinations. Geologists early arrived at minimum figures of well over 100,000,000 years for the main part of geological history. The discovery of radioactive decay around 1900 provided an entirely new and much more reliable method for dating events in the Earth's history. Radioactivity consists of the transformation of one species of atomic nucleus into another. In the case of naturally radioactive substances, this transformation occurs spontaneously, and no chemical or physical condition of the environment has any effect on the rate of radioactive decay. In any given interval of time a fixed fraction of the atoms of the mother substance decays into atoms of the daughter substance. Radioactive decay, therefore, furnishes a natural and trustworthy type of clock wherewith to measure the distant past. Only the most long-lived of radioactive substances are useful for dating earlier periods in the history of the Earth (see DATING, RELATIVE AND ABSOLUTE).

The age of the Earth

Attempts have been made to estimate the total age of the Earth under certain plausible assumptions, using data on the relative proportions of lead isotopes in very old rocks and in meteorites. The age arrived at is about 4,600,000,000 years, a figure now well established and believed to be accurate to within a very small percentage of possible error.

MAGNETISM AND TEMPERATURE

Characteristics of the Earth's magnetic field

Earth's magnetic field. The Earth's magnetic field (*q.v.*) has two outstanding characteristics. The first is its irregularity, as seen for instance in the variation of the magnetic declination from place to place, the overall pattern of which is quite complicated. The second is its gradual change in time, a phenomenon well-known to all navigators because the declination of the needle changes continuously. Magnetic maps indicating the declination are seriously in error after a few years and must be redrawn periodically from new observations. This change is known as the secular variation of the Earth's magnetic field. If a variation of 10′ per year continues over, say, 60 years, it corresponds to the considerable change of 10° in declination. The pattern of change is, however, itself variable.

These rapid changes make it clear that the field does not originate in the solid mantle. The magnetic field is caused by electric currents that flow in the fluid, metallic core. The secular variation indicates that the fluid of the core is in motion. This motion carries the electric currents about and thus produces the variation of the field observed at the surface. The details of this variation, rather irregular, may be interpreted as representing so many large-scale eddies in the fluid. From the secular magnetic variation may be obtained a value for the average speed of the fluid iron: 0.03 centimetre per second or about 10 kilometres (six miles) per year. Further development of these ideas leads to the conclusion that the fluid motion not only modifies the electric currents but also can produce and maintain them. This process resembles closely the one that takes place in a rotating generator of electric currents as found in power stations, in which current is produced by the motion of metallic wires past each other. In the core, streams of molten metal are the equivalent of the moving wires. This explanation of the Earth's magnetic field, known as the dynamo theory, has become accepted generally.

Rock magnetism. When a lava melt is solidified in a magnetic field and the field is removed later, the solid rock retains a slight amount of so-called remanent magnetism along the direction in which the original magnetic force pointed. The remanent magnetism is due to small grains of magnetic material (iron oxide) embedded in the rock. It becomes possible to infer from this the direction of the Earth's magnetic field at the time an ancient lava flow occurred. Similarly, sandstones have been shown to retain a small remanent magnetization along the direction of the field in which the material was deposited. The study of rock magnetism (*q.v.*) has been systematically developed, giving information about the character of the magnetic field at earlier times in the Earth's history. There seems to be little doubt that, during the last 500,000,000 years or so that are accessible to detailed geological study, the Earth has had a magnetic field similar to its present one.

Past magnetic reversals

Work on rock magnetism has shown, surprisingly, that apparently the magnetic field has reversed its polarity many times, perhaps hundreds of times during the past history of the Earth—that is, the magnetic south pole became the north pole, and conversely. A detailed explanation of this phenomenon is lacking; it can be said merely that its occurrence is conceivable in terms of the dynamo theory, because a self-excited generator can produce electric current in one direction as well as in the other, depending on its starting conditions. This phenomenon is also seen in some stars.

Temperature and heat transport. There is no direct way of determining the temperature distribution inside the Earth, but it is well-known that the temperature rises in the lower levels of deep mines. The rate at which heat is delivered to the surface is given by the rate at which the temperature rises (the thermal gradient), multiplied by the thermal conductivity, which is a characteristic constant of the material. Thus, by measuring the two last-named quantities, the heat flowing out of the Earth can be determined. The results of heat-flow measurements on the ocean bottom indicate that, even where the oceans are very deep, the heat flow is comparable to that flowing out on the continents. The average thermal gradient in the outermost layers of the Earth is about 25°–30° C per kilometre (70°–90° F per mile) of depth. The average heat flow near the surface of the Earth is about 1.2×10^{-6} calories per square centimetre per second.

Assuming, say, average thermal gradients of 14° per kilometre at some greater depth, the temperature at a depth of 300 kilometres would be about 4,200° C (7,600° F), far above the melting point of any silicate, even at the pressures prevailing at such depth. Hence the temperature must level off; that is, the thermal gradient must radically diminish even at moderate depths inside the Earth. It long has been concluded that most of the heat flowing out of the Earth is produced by radioactivity (mostly from the decay of the uranium and thorium families) and that the radioactive material must now be strongly concentrated in the top layers of the Earth.

It was long believed that most of the radioactive material is in the crust, and because the crust is thick under the continents but thin or almost absent under the oceans, this occurrence would mean that most of the heat flow would occur in continental areas. Heat-flow measurements on the ocean bottom have shown that this view is wrong; it is now usually assumed that heat, especially under the oceans, is also transported by mechanical, convective motions.

In the deeper layers of the Earth it is necessary to rely largely on guesswork, but any assumed distribution is subject to certain conditions. It must remain below the melting point of the simple silicates throughout the mantle, it must be above the melting point of iron in the outer core, and if the inner core consists again of iron the temperature must rise to the melting point of iron at the boundary of the inner core and exceed it inside the latter. The best way to obtain the temperature inside the

Earth is, therefore, to estimate the increase of melting point with increasing pressure. The existing estimates of the temperature in the Earth's interior tend toward relatively low values, perhaps in the neighbourhood of 3,000° C (5,400° F) at the boundary of the core and near 4,000° C (7,200° F) at the centre of the Earth. These estimates, however, may be in error by many hundreds of degrees.

ORIGIN OF THE SURFACE FEATURES

The surface of the Earth shows a remarkable degree of irregular structure. The surface levels cluster around two preferred heights: (1) sea level—most of the areas of the continents and of the offshore continental shelves lie near it—and (2) ocean bottoms, with an average depth of about 5 kilometres (3.1 miles). Levels between these are rare. Another remarkable regularity has to do with the asymmetrical distribution of land and water masses. If a diameter is drawn arbitrarily across the Earth's centre, in more than 90 percent of all cases, one end is on land, the other on the sea. The fact that conclusive explanations, going beyond mere speculation, of such simple regularities cannot be given is indicative of lack of understanding of most of the basic physical and chemical processes that have shaped the surface of the Earth into its present form. The irregularities mentioned are treated in detail in EARTH, PHYSIOGRAPHY OF.

Isostasy. The crust considered as a load on the mantle is less dense; it is specifically lighter than the substratum. Such a load will sink only to a depth at which it is floating freely with respect to vertical displacements; it resembles an iceberg floating in the water with the larger part submerged beneath the surface. According to the principle of isostasy there is a tendency toward vertical equilibrium of the crust. Quantitatively speaking, the total mass encountered on going from the surface down toward the Earth's centre tends to be the same for any point of the surface. Where there is a mountain there seems to be more mass, but isostasy indicates that there must be a deficiency of mass farther down; the mountain has a root consisting of matter less dense than the matter lying under lowlands and the oceans. The root is bigger than the mountain itself; the lighter material extends down to about six times the mean elevation of the mountain.

This idea of isostatic (vertical) compensation was first developed on the basis of gravity observations (see EARTH, GRAVITATIONAL FIELD OF). If ideal isostatic balance prevailed, gravity would be the same over the oceans and continents (apart from variation with geographical latitude). Actually, most mountainous elevations and most oceanic deeps are fairly well compensated isostatically. Deviations, known as gravity anomalies, occur in the regions in which active mountain building occurs, such as the Pacific island areas (*e.g.*, the Japanse islands).

Volcanism and earthquakes

Volcanic activity is concentrated in these same regions, as are earthquakes. Many earthquakes originate fairly near the Earth's surface, but in active regions earthquakes are observed to originate at depths down to 700 kilometres (430 miles). Thus the upper mantle seems to participate to at least this depth in mountain-building.

Deformation of the crust: mountain building. Four agencies have been held responsible for mountain building, either individually or jointly, though it is probable that several (especially the last three) have been acting together.

1. *Contraction*—The Earth resembles a shrunken apple, the mountains and valleys being likened to the wrinkles of its skin. This is the oldest hypothesis, but since the discovery of radioactive heating of the Earth it has been difficult to believe that the Earth has shrunk appreciably in geological times.

2. *Convection*—If the mantle were a viscous fluid heated from below, convective upwelling would occur at a slow rate. If the thermal conditions are right, there must certainly be a tendency toward convection, though the mantle is a plastic solid and not a viscous fluid, and this might make actual patterns look quite different from those of a fluid.

3. *Phase change*—A change in the crystalline structure of minerals, which may occur as a result of either temperature or pressure changes is a phase change. Much evidence shows that at high pressures certain more compact forms of minerals that do not exist at the surface of the Earth are stable. It has been claimed that the Mohorovičić Discontinuity at the bottom of the crust represents such a phase change. Temperature might rise by radioactive heating and this rise could shift the discontinuity downward. Temperature falls again and the discontinuity rises if the heat is carried off by volcanism and other agencies. The tendency toward isostatic equilibrium might then produce fairly large vertical displacements.

4. *Accretion*—Because at increasing depths in the Earth the temperature first rises rapidly and then more slowly, the silicates of the mantle might be closer to melting at a depth of a few hundred kilometres than elsewhere. If melting occurs locally, the material will separate into its lighter components on top and its heavier below. Lighter material might gradually make its way through the overlying layers to the top, and in this manner the amount of crustal material can grow. Volcanism appears to be a process of this kind. The continental blocks of the crust may thus have only gradually accumulated through geological ages. Strong evidence exists that the ocean and the atmosphere are secondary products of accretion—that is, have been slowly formed in the course of the Earth's history from material coming out of the rocks (see OCEANS, DEVELOPMENT OF; ATMOSPHERE, DEVELOPMENT OF).

A more definitive theory of mountain building arose in the 1960s and 1970s. The discovery of the extent and nature of the Earth's mid-oceanic ridges and the fact that sea floor materials became older along paths normal to the ridges led to the concept of sea-floor spreading—that new ocean floor is generated by the upwelling of the mantle along the ridge crests, that it subsequently moves laterally away from the ridges and toward the continents, and that it is eventually subducted or dragged downward back into the mantle along continental margins. These processes are thought to result in both continental drift and mountain building and are further discussed in the article MOUNTAIN-BUILDING PROCESSES.

BIBLIOGRAPHY. The subject of this article is covered by a variety of sciences with a great deal of overlap. There are two major sciences, astronomy and geology, and some minor ones, mainly geodesy, seismology, geochemistry, and geophysics, in its academic rather than its economic context. More detailed bibliographies may be found under the names of the sciences mentioned. In addition, up to about 1965, the subject presented here was almost invariably dealt with in the introductory chapters of treatises on geology, but most of this is definitely obsolete in view of the discoveries of recent years. Therefore, this bibliography is limited to a few relatively recent selected texts arranged in order of ascending technical difficulty.

ARTHUR STRAHLER, *The Earth Sciences*, 2nd ed. (1971), about half of the content deals with matters covered here (nonmathematical); I.G. GASS, P.J. SMITH, and R.C.L. WILSON (eds.), *Understanding the Earth* (1971), some 25 articles by as many authors who are leading experts (readable, nonmathematical); SCIENTIFIC AMERICAN, *Readings in the Earth Sciences*, 2 vol. (1969), an excellent collection containing 74 offprints of articles that appeared in the *Scientific American* between 1948 and 1969—see also current issues of this periodical for further valuable articles; OWEN M. PHILLIPS, *The Heart of the Earth* (1968), introductory for a college science student; FRANK D. STACEY, *Physics of the Earth* (1969), an outstanding text for the advanced undergraduate or beginning graduate student of science with a corresponding mathematics background, including an extensive and carefully selected bibliography; GEORGE D. GARLAND, *Introduction to Geophysics* (1971), a thorough and reliable work that covers several subjects not adequately treated by Stacey; J.A. JACOBS, R.D. RUSSELL, and J.T. WILSON, *Physics and Geology* (1959), a classic text to introduce geology students to the underlying physics with emphasis on those aspects that have direct bearing on large-scale geology; S.K. RUNCORN *et al.* (eds.), *International Dictionary of Geophysics*, 1 vol. and 1 vol. of maps (1967), articles contributed by nearly a thousand authorities, with short bibliographies following the articles.

(W.M.E.)

Earth Movements on Slopes

Hillslopes undergo a permanent development caused by various processes that mold their shape. Of major importance are earth movements that occur under the influence of gravity when the stability of a slope is disturbed either by natural forces or by human interference. Earth movements of the most varied nature depend on the interaction of a number of factors, including the angle of slope, nature of materials, and time.

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

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

Significance to man

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

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

Dealt with in this article are the factors that produce earth movements on slopes and the several types of movements that occur in nature. A section on some of the better known historical disasters and their causes also is included. For further information on the types of hillslopes that exist in nature see HILLSLOPES. The strength of rocks and their general resistance to erosion are covered in the articles ROCKS, PHYSICAL PROPERTIES OF; and WEATHERING. Certain aspects of the mechanics of slope failure are covered in SOIL MECHANICS, APPLICATIONS OF.

FACTORS PRODUCING EARTH MOVEMENTS

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

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

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

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

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

Effects of earthquakes, groundwater, and climate

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

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

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

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

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

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

TYPES OF EARTH MOVEMENTS

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

Origin and occurrence of rockfalls

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

Hundreds of rockfalls have been recorded from young mountain ranges, such as the Alps, Carpathians, Himalayas, Andes, and Rocky Mountains. One of the largest was the rockfall in the valley of Bartang River, in the Pamir Mountains in 1911. A rock mass of about 4,800,000,000 cubic metres (6,300,000,000 cubic yards) fell and dammed the valley, creating a lake 75 kilometres (47 miles) long and 262 metres (859 feet) deep. When large rock masses drop into a lake or fjord, dangerous huge waves flood the coast. In the Norwegian fjords numerous rockfalls have occurred, and catastrophic results were caused in part by the suddenness of the event.

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

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

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

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

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

Creep results in the bending of beds (Figure 1). Friction that is active between the creeping debris and the surface of the bedrock produces a gradual bending of the bed

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

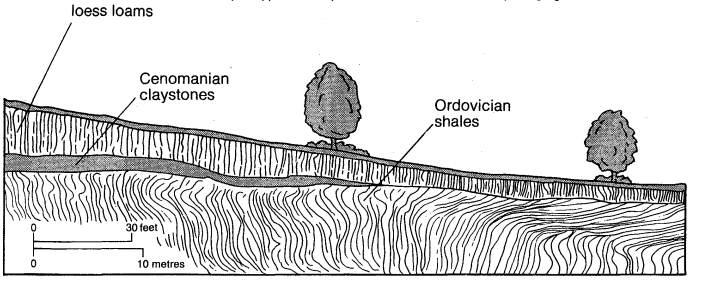


Figure 1: Creep and bending of beds in a loam pit at Prague, Czechoslovakia (see text).

faces cropping out at the surface. The dragged-out and disrupted layers of the bedrock become part of slope deposits, thus increasing their thickness.

Deformation by bulging

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

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

Figure 2 shows an example of bulging in a river valley near Ostrava (Czechoslovakia). The valley occurs in

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

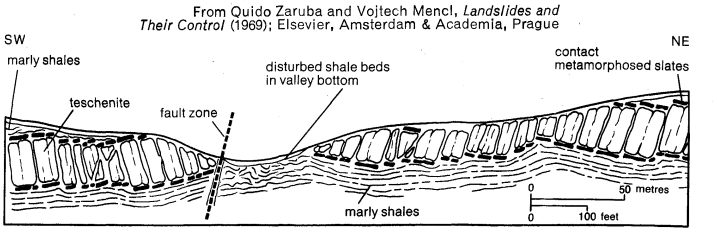


Figure 2: Squeezing-out of marly shales on the valley bottom of the Lucina River, Czechoslovakia.

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

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

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

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

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

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

Slide characteristics

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

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

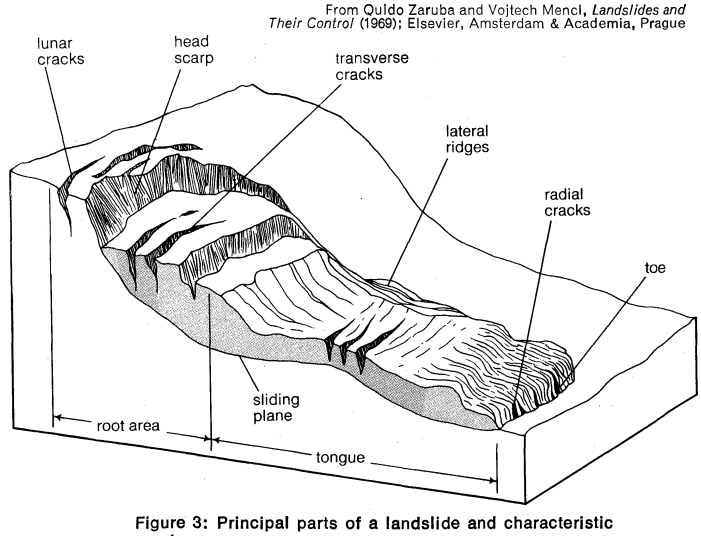


Figure 3: Principal parts of a landslide and characteristic cracks.

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

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

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

Form and occurrence of slumps

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

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

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

Earthflows generally originate in a large basin in the upper part of the slope, where debris and weathering ma-

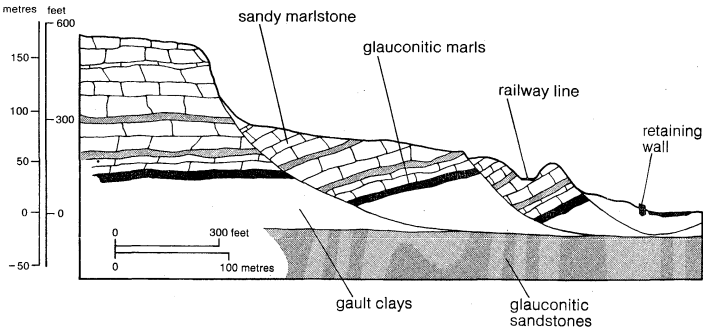


Figure 4: Landslide on the seashore near Folkestone, Kent.

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

Role of water saturation

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

With increasing velocity earthflows grade into debris flows. Mountainous debris flows are sometimes called avalanches or Muren, a term currently used in the Alpine countries. As a rule, they originate above the timberline in gorges filled with rock detritus. During torrential rains, debris and larger stones travel at a tremendous speed down the stream channels. The material is unsorted, and the ratio of solid particles to water is about 1:1. The rates of flow are so great that moving trains have been trapped and buried beneath debris avalanches. The Muren are very disastrous and often result from too great deforestation of mountain slopes. Scarcity of vegetation is also responsible for much-feared debris and mudflows in arid and semi-arid regions. This group of slope movements also includes volcanic mudflows. Volcanic explosions are usually accompanied by torrential rains that wash ash and ejected material downward as a mushy mass. Debris flows occasionally originate on the slopes of dormant volcanoes that are covered by glaciers. When volcanism recurs, the increase in heat before and during the eruption suddenly melts the glaciers, and the waters bring much debris downhill.

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

Failure of quick clays

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

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

CHARACTERISTICS OF UNSTABLE SLOPES

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

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

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

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

HISTORICAL DISASTERS

The Hope, British Columbia, rock slide

No warning devices can be effective in the case of a slope failure provoked by an earthquake; the so-called Hope rock slide in the Cascade Mountains, Canada, was such an occurrence. In January 1965 a block of about 140,000,000 tons buried the British Columbia Highway for more than three kilometres (two miles) to a depth of about 90 metres (300 feet). The movement was so rapid that the rock mass buried three cars that were in motion, and some people lost their lives. It is one of the largest slides recorded in North America, involving a volume of 47,000,000 cubic metres (61,000,000 cubic yards). The coarse rock detritus descended from an altitude of about 1,200 metres (3,900 feet) in several surges and was deposited on the valley floor. Muds were driven up the opposite slope, and the effect of the air blast and dust cloud extended far along the valley. Under contributive geological conditions—such as schistosity and jointing parallel to the surface and readily weathering serpentinite and carbonate rocks filling the fissures—the movement was triggered by an earthquake. It is of interest that a much stronger earthquake in 1872 did not threaten the stability of this slope. The earthquake in 1965 proved to be more effective, probably because it had a shallow focal depth and a suitable frequency and orientation of the horizontal shock-wave component.

The largest rock slide known to have occurred in Central Europe occurred near Flims, in the Graubünden Canton, in Switzerland. The stability of the mountain slope, formed by Jurassic marly limestones dipping moderately (7°–12°) toward the Rhine, probably was disturbed during an interglacial stage of the Pleistocene, after the retreat of the valley glacier. A huge block of rock, about 12 cubic kilometres (three cubic miles) in volume, slid down a bedding plane, and the head scarp, called Flimserstein, is still clearly visible. The moving mass must have attained a great velocity because it rose 150 metres (500 feet) up the opposite valley slope. The Rhine Valley was blocked over a distance of nearly 15 kilometres (nine miles), and a lake was impounded. Only when the river had again deepened its channel did the lake dwindle away.

The Mt. Rossberg, Switzerland, rock slide

The disastrous slide that took place on the slope of Mt. Rossberg in Switzerland in 1806 buried the town of Goldau. A complex of Tertiary conglomerates about 20,000,000 years old slid along a bed of bituminous marls dipping 19°–21° downslope. The loosened block was 1,700–2,000 metres (5,600–6,600 feet) long, 300 metres (1,000 feet) broad, and 60–100 metres (200–320 feet) thick. Its total volume was estimated at 40,000,000 cubic metres (52,000,000 cubic yards). In this case, several people were eyewitnesses to the movement. Some saved their lives by jumping over the marginal crack, but within one or two minutes the rock stream had reached a stupendous velocity and destroyed several villages in the valley. Because the slope had always been stable before, the question of the forces that had disturbed its equilibrium was extensively discussed. The origin of the rock slide was explained by three main factors: (1) orogenic forces might have caused a gradual increase in the angle of slope until the coefficient of friction on the bedding plane underlying the loosened block was surpassed; (2) friction may also have been reduced as a result of weathering of rocks on the bedding plane; (3) hydrostatic pressure of water in the bedding joints may have increased after rainy years and caused the uplift of the loosened block. A reinvestigation in 1950 of the reasons for the catastrophe led to the conclusion that it could most plausibly be explained by the increase in the inclination of slope by tectonic movements or by the decrease in the bond between the slide block and the substratum due to weathering.

The catastrophic Vaiont landslide in the Italian Alps in 1963 is analogous. A huge mass of Jurassic limestone of about 260,000,000 cubic metres (340,000,000 cubic yards) slid along a bedding plane into a reservoir, soon after it had been filled for the first time. The greatest damage was produced by a water wave 100 metres (330 feet) high, which overflowed the dam and destroyed and damaged several towns in the Piave Valley.

One of the oldest reports on slope failure is that of Marius of Aventicum, who was bishop in Lausanne toward the end of the 6th century. In his chronicle he describes a rockfall, which took place in the year 563 on Mt. Tauretunum. The rock mass fell into Lake Geneva, producing a large wave that devastated the shore and destroyed all settlements far inland. The rockfall that occurred in Italy, near Rovereto on the river Adige in 883 must have been of an extreme size. It is reported that the rock mass buried a town and blocked the river so that the town of Verona was deprived of water for a long time. Even if Dante's account of this catastrophe in his *Divine Comedy* is poetically exaggerated, the disaster was undoubtedly enormous, as evidenced by the debris field extending in the valley to a distance of several kilometres (Slavini di San Marco).

The Folkestone, England, landslide

An example of extensive sliding on sea coasts is the well-known landslide near Folkestone, in southern England. Upper Cretaceous rocks overlie the Gault Clays on a steep bank. During the last hundred years a number of large landslides have interrupted the Dover–Folkestone railway line several times. The motivating force is marine abrasion of clays at sea level, which results in unloading of the foot of the steep bank and, consequently, in slumping along the curved slide surfaces. Stability is decreased by hydrostatic water pressure in the fissures and joints traversing the Cretaceous rocks. Tests on samples taken in borings have revealed that the Gault Clays near the slide surface possess only a small part of the strength of the undisturbed clays. During every new landslide, clays at the toe are squeezed out and heaved into a ridge that rises as an island from the sea; in the course of time the ridge is removed by abrasion until it disappears completely. If the erosion of squeezed-out clays could be prevented, the danger of sliding would be diminished, because their weight contributes to the stability of the slope. Therefore, thick retaining walls with transverse buttresses have been built along the shore to protect it from abrasion and make possible the deposition of coastal sediments.

Disastrous slope failures were often provoked by ill-founded quarries. In 1618 the undermining of the slope of Monte Conto, north of Lake Como, by pits on talc schists promoted a rockfall that destroyed Plurs, a rich and well-built town, and buried more than two thousand inhabitants. Near Elm, in the Swiss Alps in 1881, a rockfall was caused by the undermining of the slope by a slate quarry. In a few minutes more than 10,000,000 cubic metres (13,000,000 cubic yards) of rock fell; 83 houses were demolished, and 115 lives were lost.

The Frank disaster

Mining operations set the stage for the rock slide in 1903 that destroyed part of the town of Frank, below Turtle Mountain, in Canada (Figure 5). The mountain ridge above the town consists of solid limestones that are thrust upon relatively soft sandstones and shales. Limestones dip 50° toward the slope and are strongly jointed; the sandstones and shales are intensely folded and steeply inclined. Part of the slope broke off across the bedding plane, and more than 30,000,000 cubic metres (39,000,000 cubic yards) of rock debris covered the valley plain. The failure was almost instantaneous, lasting less than two minutes. Movement was provoked by deformation following the exhaustion of a coal seam at the foot of the slope and by the wedging effect of freezing water in the fissures.

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

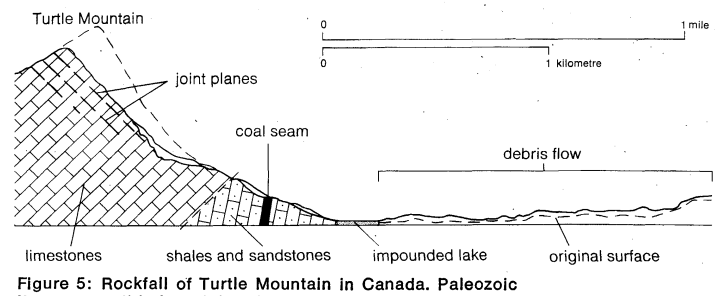


Figure 5: Rockfall of Turtle Mountain in Canada. Paleozoic limestones slid along joint planes.

Slope failures may likewise be induced by careless excavation works for transport routes and dam abutments, by sudden lowering of the level of lakes, and others. This type of human interference represents the trigger for many slope failures; the basic cause is a particular combination of topography, geology, and climate.

BIBLIOGRAPHY. ROBERTO ALMAGIA, *Studi geografici sopra le frane in Italia*, 2 vol. (1907–10), a comprehensive book on landslides in Italy, particularly in relation to geographical and climatic conditions; LAURITS BJERRUM, *Mechanism of Progressive Failure in Slopes of Overconsolidated Plastic Clays and Clay Shales* (1966), a modern theoretical analysis of slope stability in overconsolidated clays; EDWIN B. ECKEL (ed.), *Landslides and Engineering Practice* (1958), a summary of research on the origin and correction of landslides in the United States with respect to highway construction; ALBERT HEIM, *Bergsturz und Menschenleben* (1932), a report on rock falls and mountain slides in the Alps, demonstrated by many examples; S.E. HOLLINGWORTH, J.H. TAYLOR, and G.A. KELLAWAY, *Large-Scale Superficial Structures in the Northampton Ironstone Field* (1944), the first description of slope deformations originated by bulging in valley bottoms; Г.Б. ПАЛЬШИН, Оползни (1963), an analysis of extensive blockslides in the Angara and Ilima river valleys (Siberia), studied in

connection with the construction of the Bratsk Dam; RAIMONDO SELLI *et al.*, *La frana del Vaiont* (1964), a monograph on the disastrous landslide into the Vaiont reservoir (Italy) in 1963; C.F.S. SHARPE, *Landslides and Related Phenomena* (1938), a survey of slope movements, their classification, and relationship to geomorphological cycles; KARL TERZAGHI, *Mechanics of Landslides* (1950), a classic work on processes leading to landslides and considerations on modern preventive measures; QUIDO ZARUBA and VOJTECH MENCL, *Landslides and Their Control* (Eng. trans. 1969), a book summarizing the economic importance of landslides, their relation to geological conditions, and controlling measures—includes a comprehensive list of literature.

(Q.Z.)

Earthquakes

An earthquake commonly is described as a sudden individual tremor within the Earth that creates shaking at the surface. Many earthquakes disturb the surface only to a degree observable with sensitive instruments, however, and some authorities therefore refer to earthquakes as the generation of elastic waves (seismic waves) in the Earth rather than to physical shaking.

In addition to sudden events, there are gradual effects such as slow creeping of earth materials that are closely related to the production of earthquakes. Earthquakes often occur in succession, or in earthquake swarms, the effect of which is that of continuous disturbance. Other causes, largely meteorological, produce waves that are recorded on seismograms and often go on for hours or days; these are known as microseisms, but ordinarily they are not considered as earthquakes. Recently, certain extremely small individual events have been called microearthquakes. A final category includes artificial earthquakes, which are produced by man through disturbance of subsurface conditions by any of several means.

Earthquake damage and precautions

Earthquakes have been objects of concern and superstitious awe through most of recorded history. No other natural event is disastrous over so large an area in so short a time. Lives lost as a result of a single earthquake often range to the tens of thousands, sometimes to the hundreds of thousands. Nearly all this loss and most of the enormous accompanying damage are unnecessary and preventable; it generally is caused by the collapse of weak structures that would not be erected under any modern system of building regulation and inspection. Persons are crushed and buried under these falling structures, or are burned to death in fires that rage out of control following building failures. In addition, many casualties and much damage have resulted from tsunamis, the great sea waves that have swept coastal areas on the heels of large earthquakes.

The most generally important application of earthquake study is directed toward abatement of earthquake risks. Knowledge of the behaviour of structures during earthquakes has led to recommendations for safe construction of ordinary small buildings. Geographical study has revealed the areas of greatest earthquake risk, which therefore should be avoided unless special precautions are taken. It is generally known today, for example, that soft foundations—those in alluvium, sands, or artificial fill—are much more unstable than those in solid rock.

Specific earthquake warnings are at best only a hope for the distant future; but tsunami warnings are a present reality. Tsunami waves travel much less rapidly than the earthquake disturbance that creates them, and seismological stations often record an earthquake long before the sea waves can reach populated coasts. A Pacific warning service is in operation with headquarters near Honolulu.

Many other important applications result from earthquake or seismic study. Earthquake waves are used to investigate the interior of the Earth as a whole as well as the geology and subsurface structures of an area being investigated for oil or minerals. There exists a highly developed industry and profession of geophysical prospecting by seismic methods, using recordings of small artificial explosions. Similar procedure on a global scale, using the relatively high energies released in nuclear detonations, has advanced man's knowledge of the innermost parts of the Earth. Much effort also has been directed toward the detection and identification of nuclear tests; this involves distinguishing nuclear explosions from natural earthquakes.

This article treats of the causes and distribution of earthquakes and their magnitude, motion, and energy. For further information on the instrumentation involved, see SEISMOGRAPH; for the utilization of earthquake data in general Earth studies, see EARTH, STRUCTURE AND COMPOSITION OF; and ISLAND ARCS; and for additional details of the nature and transmission of seismic waves, see WAVE MOTION; ELASTICITY; and EARTH, MECHANICAL PROPERTIES OF.

CAUSES OF EARTHQUAKES

Natural causes: rock fracture and volcanism. Earthquakes can originate in various ways but one principal cause is responsible for all large earthquakes and for a majority of small ones. This is the fracturing of rocks in the outer part of the Earth as a result of the gradual accumulation of strain during geological processes. Sudden fracture occurs when strain exceeds the strength of the rocks; generally, rupture follows established lines and surfaces of weakness, which are active geologic faults. Fracture usually begins at depths of 17 to 34 kilometres (11 to 21 miles) or much more; but in large earthquakes, it often extends to the surface, and relative displacements of the two sides are observed along the line or zone of intersection between the fault and the surface. This fault displacement may be vertical, horizontal, or both (Figure 1).

From John S. Shelton, *Geology Illustrated* (© 1966); W.H. Freeman and Co.

Figure 1: The Garlock Fault, California, looking west.

The ultimate cause of distortion and faulting is the same as that responsible for the warping and breaking of the crust of the Earth during geological time, thereby producing mountain ranges and ocean basins. An earlier generation of workers confidently attributed all this to the wrinkling of the surface of a cooling and shrinking Earth. That idea gradually became untenable and is being replaced by variations of a hypothesis that the flow of heat within the Earth causes transfer of material by convection, and that these slow but inexorable motions provide the driving forces for geologic changes. The latest and most fruitful development, called the new global tectonics, has revived the temporarily neglected idea of continental drift and combined it with the new notions of sea-floor spreading and transform (offset) faulting, phenomena that are substantiated by the paleomagnetic history of oceanic rocks (see further SEA-FLOOR SPREADING).

The most important of the minor causes of earthquakes is volcanism. Earthquakes are caused not merely by explosions and other eruptive processes, but also by gradual displacements of subterranean volcanic material, or magma.

Artificial earthquakes. Artificial earthquakes include all quake events that are in some way induced by or related to man's activity. These range from the vibrations produced by machinery and common vehicular traffic to the shocks of underground nuclear explosions; the former yield a seismogram pattern that is akin to that associated with naturally occurring microseisms and the latter produce a pattern quite like that of natural major quakes. Two other means of inducing earthquakes involve dam and reservoir construction and the underground injection of liquids through wells. The earliest relevant observations of shocks associated with dams and reservoirs were made in the area of Lake Mead on the Colorado River. Special recording over a number of years showed that these events correlated with changes of water level in the reservoir, being most frequent when the level was changing most rapidly. This indicated that instability existed among the Earth's crustal blocks in the area but that the quakes were artificially triggered. Several other large reservoirs were monitored during the construction and filling periods, but no similar results were obtained. Local earthquakes, however, have been correlated with the filling of large reservoirs in Greece, Zambia, and India. The last culminated in a destructive earthquake at the town of Koynanagar in 1967, although the Koyna Dam and Reservoir were not seriously damaged.

Changes in overlying and subsurface pressures

A series of earthquakes began suddenly in 1962 in the vicinity of a deep well near Denver, into which liquid wastes were being pumped. The shocks clearly were associated with the pumping, allowing for a time lag. The largest shocks caused some damage in Denver, where perceptible earthquakes otherwise have been infrequent. After pumping terminated, these earthquakes continued, and one of the last events was the largest of the whole series. These later events also were farther from the well. It is inferred that the pumping disturbed the equilibrium of the subsurface layers of rock and thus triggered the earthquake process.

DISTRIBUTION OF EARTHQUAKES

Geographic and geologic occurrence. Detailed knowledge of the geography of earthquakes depends on sensitive seismographs, which record all earthquakes of consequence, even at the greatest distances. Data from one well-equipped station would suffice for complete global listing of major events. With hundreds of stations now in continuous operation, cataloging extends to events of lesser magnitudes.

An early, unexpected finding was that the majority of earthquake sources occur beneath the sea. The greatest submarine seismicity is usually near continental coasts or island chains, so that the larger events often have strong effects on land. Nevertheless, the large earthquakes that attract public attention constitute only a minor fraction of those identified by seismographic recordings.

Historical listing is likely to give a disproportionate impression of the seismicity of Mediterranean Europe, northern Africa, and the Near East, because centres of population there include many examples of weak and ancient types of construction. These tend to fail under comparatively moderate shaking, often resulting in appalling loss of life (for example: 12,000 killed at Agadir, Morocco, in 1960). Even seismographic catalogs may lead to an erroneous conclusion simply because there are a relatively large number of good recording stations in Europe and because there is exceptional interest in earthquakes occurring there. This causes many small events to be studied and included in published lists, whereas larger and more numerous earthquakes in remote parts of the world may not be sufficiently recorded because their locations are not precisely known. This unevenness in treatment is gradually disappearing as more and more stations are established.

Major earthquake activity is primarily confined to comparatively narrow belts; this is particularly true beneath the sea. On the continents there are some broad areas, as in eastern Asia, where earthquakes occur along a sheaf of roughly parallel fault zones. Small earthquakes are more generally distributed, and there are few large areas without them.

Seismicity in the Pacific and Alpine belts and in the oceanic ridges

Most earthquakes are shallow in the seismological sense; they originate within the rocks of the Earth's crust, above the Mohorovičić Discontinuity, which separates the crust and mantle. In certain regions, however, particularly around the Pacific, there are deep-focus earthquakes that occur at depths of more than 700 kilometres (450 miles) below the surface; their sources are chiefly within comparatively thin slablike zones that are inclined at angles of about 45°. Usually the inclination is toward the continents, but generally it is toward the concave side of island arcs. For this reason, the dip is toward the Pacific in the New Hebrides and some other areas.

Shallow seismicity is associated mainly with: (1) the circum-Pacific belt of active structures, particularly the island arcs; this includes various branches, one extending into Indonesia and another around the Caribbean; (2) the trans-Asiatic belt, extending from Mediterranean Europe to the Pacific, with the relatively narrow Alpine belt along its southern margin; (3) the midoceanic ridges, now considered to form a connected worldwide rift system.

Earthquakes are nearly absent from much of the ocean basins, notably between the midoceanic ridges and the zones off the continental coasts; the Hawaiian rift structure is a conspicuous exception. They are also relatively sparse in the central, stable areas of the continents—the oldlands or shields, such as the Canadian Shield, the Brazilian Shield, and the Baltic Shield. The absence of earthquakes on the Antarctic continent is well confirmed, except for small ones near active volcanoes and a few that appear to be icequakes rather than ordinary earthquakes. These result from the fracturing of large ice masses.

Major seismicity in the circum-Pacific and Alpine belts is regionally associated with the youngest mountains, whereas minor earthquakes, and occasional large ones, occur in the areas of older mountain building, such as the Appalachian region of the United States, the southernmost part of Africa, and southeastern Australia.

Occasional large earthquakes occur in unusual locations. In 1811 and 1812, large earthquakes originated in the Mississippi Valley, centring not far from New Madrid, Missouri; they were comparable to the largest earthquakes of record in the western United States.

Tectonic theories must account for the orderly association of earthquakes and other geophysical phenomena with active island arcs of the Pacific type. In such an arc, there is a succession of definite zones of which the outermost is an oceanic trough or deep. Major shallow earthquakes originate chiefly under the slope that descends from the continent or large islands into this deep. Along the same zone appear remarkable deficiencies of gravity; small, nonvolcanic islands also occur here. The next arcuate zone, between the outermost zone and the continent, typically consists of large islands, such as those of Japan or the East Indies, with a central alignment of active volcanoes. Earthquakes occur at depths of 100 to 150 kilometres (60 to 90 miles) beneath these islands. Another interior zone exists in some arcs, including older and largely extinct volcanoes, beneath which earthquakes occur at depths of 200 kilometres (120 miles) or more. Still farther within the concave side of the arcuate structure, earthquakes occur at even greater depth, ranging from 300 to 700 kilometres (190 to 440 miles). Plots of the epicentres (points on the earth's surface directly above the focuses of earthquakes) against depth for shallow and deep earthquakes reveal the inclination (45°) of the active surface previously mentioned.

Earthquakes at great depth were known only from the Pacific region (including Indonesia) until 1934, when a large one originated at a depth of 650 kilometres (400 miles) under Spain. Earthquakes at depths of as much as 400 kilometres (250 miles) occur under the Tyrrhenian Sea west of Italy, and earthquakes at depths of 150 kilometres (90 miles), some rather large, have occurred repeatedly under Romania. A most remarkable repeating source is located at a depth of about 220 kilometres (140

miles) under the Hindu Kush mountain range, on the border between Pakistan and Afghanistan; associated with it are other sources at slightly shallower depth. These are displaced somewhat in the geographical sense, and the entire group defines a remarkable deep zone.

Earthquakes associated with block faulting

Not all large-scale earthquake activity occurs in arcuate structures, even in the Pacific belt. The major exception is a large part of the Pacific coast of North America, from northern Mexico to southeastern Alaska. Here deep earthquakes are absent, and volcanism is minor or extinct. The local earthquakes are associated with block displacements involving faults of considerable length, the outstanding one being the San Andreas Fault in California. A similar but smaller zone of block faulting in New Zealand includes the main seismic areas of both large islands; in the North Island it impinges on a zone that includes active volcanoes and earthquakes at depths to 200 kilometres (120 miles), occasionally more, and still farther north is the typical arc of the Kermadec Islands. Much of the activity of the trans-Asiatic belt, especially north of the Alpine belt, appears to involve block faulting.

Deep-focus earthquakes. Deep-focus earthquakes are those originating in the Earth's mantle—below the Mohorovičić Discontinuity, which occurs at a depth of about 35 kilometres (22 miles) beneath the continents and about 10 kilometres (6 miles) beneath the ocean floors. Shallow earthquakes originate in the Earth's crust. Focal depths of earthquakes are not always precisely determined, and the distinction between shallow and deep shocks is sometimes uncertain. An additional distinction between intermediate earthquakes, at depths ranging to 300 kilometres (190 miles), and deep-focus earthquakes in the restricted sense, from 300 kilometres to more than 700 kilometres, also is made, however. No earthquakes have definitely been assigned to depths exceeding 750 kilometres. The intermediate and deep groups show a difference in geographical distribution, which is very clear if the level of separation is taken slightly above or below the 300-kilometre level in some areas.

Investigation of intermediate and deep earthquakes depends almost entirely on seismographic recording. Even large events at great depth ordinarily produce only moderate shaking at the surface. Certain large earthquakes that occurred in 1881 in the Aegean and eastern Mediterranean region were correctly recognized as originating at relatively great depth, although some of these may have been within the crust.

Existence of deep-focus earthquakes was first emphasized about 1922, however, when the time of arrival of the primary wave group (first arrivals) was recorded. The wave groups from all earthquakes start from an earthquake focus at the same instant and proceed to the recording site. Because waves travel faster in the mantle than in the crust, however, there is a greater difference between the arrival times of primary and secondary groups of waves when the focus is deep—within the mantle (Figure 2). This was the case in Japan in 1929 when

From *The Earth Sciences* (1963) by Arthur N. Strahler, Fig. 20.11, p. 376, Harper & Row Publishers, Inc. (after L. Don Leet, 1950, *Earth Waves*, Cambridge, Mass., Harvard University Press)

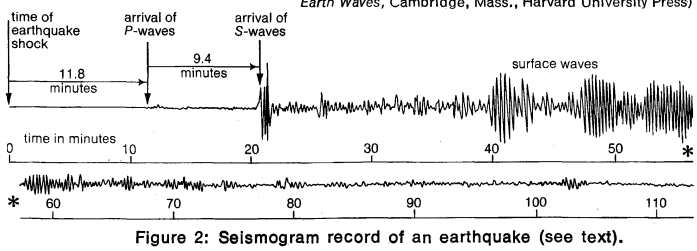


Figure 2: Seismogram record of an earthquake (see text).

the study of seismograms of earthquakes suggested that some quake events were originating at great depth.

Another distinguishing feature of seismograms of deep earthquakes is the small amplitude (wave height) of surface waves—those travelling near the Earth's surface—relative to other wave types. This is because the surface disturbance decreases rapidly as the depth of the source increases. Precise study has revealed the effect of increasing focal depth on the times of arrival of all the various waves at distant stations.

Particular significance attaches to the recording of waves of a special type. A wave can be reflected at the Earth's surface and then arrive at a given station in three ways; *i.e.*, there are three possible points of reflection. One point is roughly one-half the distance between epicentre and station; a second point is on the far side of the Earth, in which case the wave passes through the core twice; the third is a steep path from a deep source to the surface, with reflection relatively near the epicentre, toward the distant station along a route close to that of the direct path from the source (Figure 3).

From *The Earth Sciences* (1963) by Arthur N. Strahler, Fig. 20.15, p. 378, Harper & Row Publishers, Inc. (after B. Gutenberg, 1951, *Internal Constitution of the Earth*, New York, Dover Publications)

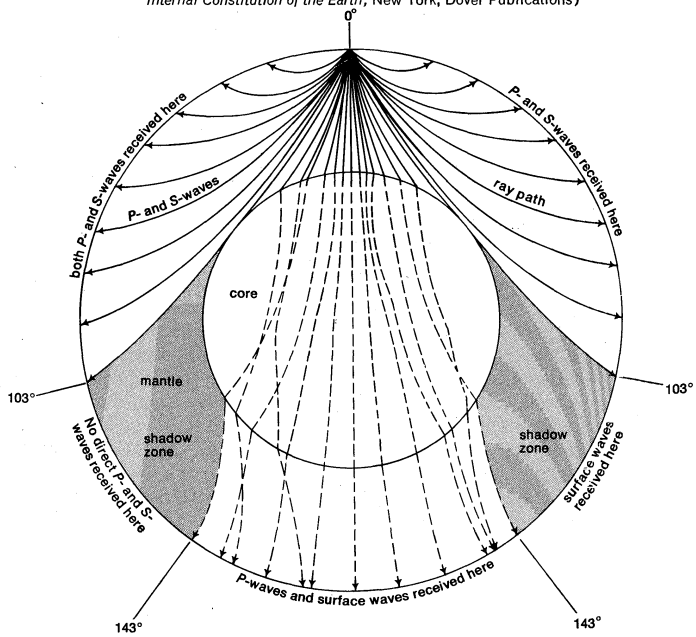


Figure 3: Possible ray paths from single earthquake source (see text).

Evidence for deep-focus earthquakes was received rather skeptically at first, because it conflicted with then-current ideas about the state of material at such depths, where strain might be expected to be relieved by plastic flow rather than by fracture. When the evidence became overwhelming, sudden changes of state were suggested. Because a large proportion of shearing fracture at the source is indicated, an appeal was made to the time element; some materials yield gradually under slowly changing load and yet fracture under sudden strain.

The "new global tectonics" attributes deep-focus earthquakes to fracture within slabs of crustal material that are being dragged down into the depths by convection in the mantle. These slabs form the active surfaces, dipping toward the concave side of Pacific-type arcs, on which the deep focuses are located. In most regions these zones of deep focuses are very thin. Fracturing, with resulting earthquakes, is considered to occur there much as for shallow earthquakes in the crust but with mechanical modifications appropriate to the greater pressure and temperature.

MAGNITUDE, MOTION, AND ENERGY OF EARTHQUAKES

Seismic intensity and magnitude scales. In 1935 an instrumentally founded magnitude scale for earthquakes was introduced. The objective was to eliminate gross errors in earthquake statistics caused by confusion of the terms intensity and magnitude. Essentially, there was confusion of the local effects of an earthquake with its dimensions and extent as a physical event. This was caused by three factors that are insufficiently understood: (1) a rapid decrease in the intensity of shaking with distance from the earthquake centre; (2) local enhancement of shaking on unstable ground, regardless of

distance from the source; (3) an exaggerated impression of violence derived from serious damage or total collapse of weak construction following only moderate shaking.

Mercalli intensity scale

In seismology, intensity is a relatively old, established term; it refers to the degree of shaking at a specified place. Usually it is expressed by means of an adopted arbitrary scale, which for many years was the Rossi–Forel scale of ten degrees. This has been gradually replaced by successive versions of the Mercalli scale of 12 degrees. The successive degrees do not derive from physical measurement of any kind; they merely represent ratings assigned by experienced workers from observations or reports of earthquake effects. Accordingly, intensity is usually stated in Roman numerals (I, II, . . ., XII), rather than in ordinary numbers, which might be interpreted to represent measurement.

The Mercalli scale levels are represented by rather long descriptions of effects. A few significant points are as follows: intensity I is ordinarily not perceptible to persons; II is felt by few; IV is sufficient to rattle windows and doors; VII may cause damage, even seriously, but only to inferior construction; VIII and IX are destructive; X is usually catastrophic in a centre of population; and XII is reserved for effects of extreme violence, which are only rarely seen.

There is a belief that the levels of the Mercalli scale can be correlated with the acceleration of the ground during earthquake shaking, thus indicating the internal forces acting within buildings. Any such correlation is extremely inaccurate and uncertain. A common rule of thumb associates intensity VII with acceleration of 0.1 gal (one gal equals one centimetre per second per second); this is often applied by engineers in designing structures for resistance to earthquakes.

An intensity scale, applied with judgment, affords a limited means to compare earthquakes with each other. If local intensities of two earthquakes, *A* and *B*, are entered on a pair of maps, it may be observed that event *A* developed higher intensity near its epicentre than did event *B*, and that the limits of successive lower intensities are wider for *A* than for *B*. The area affected, either in terms of damage or of perceptible shaking, is larger in the first instance. Earthquake *A* may then be said to be of greater magnitude than *B*, independently of instrumental recording or of any particular magnitude scale. Nevertheless, if both affect the same general region but have different epicentres, some localities may be shaken with greater intensity in event *B*, generally because they are closer to its epicentre than to that of *A*. Moreover, estimates of intensity may be distorted by exaggerated damage in large centres of population located on unstable ground.

Such comparisons and, consequently, estimates of magnitude in the general sense, are difficult to make for earthquakes originating offshore or in unpopulated areas. Accordingly, the term magnitude was introduced together with a formal scale (the Richter scale) based on instrumental recordings. The results cannot claim a high degree of precision, but experience has shown that they are generally more satisfactory than those based only on noninstrumental data.

Full understanding of the relation between intensity and magnitude requires a good foundation in fundamental physics, because the relationship involves a general property of radiated energy. It is possible, for example, to contrast the local intensity of illumination with the actual luminosity of a lamp, or the strength of a radio signal on a given receiver with the power rating of the station in kilowatts. In all such cases, local intensities are affected not only by distance, but also by conditions at the point of observation and along the path to that point from the source.

The Richter magnitude scale

The first magnitude scale was devised for use in southern California, where there was a group of recording stations with identical seismographs. These stations permitted the location of epicentres for earthquakes originating in this area, so that it was possible to study the decrease in recorded ground motion with increasing distance from the epicentre. By plotting the amplitude read from each seismogram against distance from the source, a sort of standard attenuation curve could be set up. The observed range of amplitudes was so large that the amplitude was plotted logarithmically. On the logarithmic plot, the curves representing different earthquakes were seen to be roughly parallel, which would imply that the amplitudes recorded for two given earthquakes are in constant ratio at corresponding distances. The logarithm of this ratio was defined as the difference in magnitude between two events, M_1 and M_2, which may be written: $M_1 - M_2 = \log A_1 - \log A_2$, in which M_1 and M_2 are the magnitudes involved, and A_1 and A_2 are the corresponding amplitudes. It remained to fix the zero level of magnitude. This was arbitrarily chosen to fit some of the smallest true earthquakes recorded by good seismographs. Magnitude 3 was fixed as a maximum trace amplitude of one millimetre on the recording sheet of a torsion seismometer, with a free pendulum period of 0.8 second, static magnification of 2,800, and damping of 0.8 critical (see SEISMOGRAPH).

Later developments have largely removed the restriction of magnitude to the data derived from a particular type of seismograph.

The magnitude scale is not itself an instrument but a procedure for using instruments. A common error is to suppose that it is "a scale of ten." Although the logarithmic base 10 is used, there is no fixed upper or lower limit. As a matter of observation, no earthquake has been confidently assigned a magnitude higher than 8.9. Special instruments have recorded very small earthquakes, smaller than those of zero magnitude, so that negative magnitudes apply to these "microearthquakes."

Extension of magnitudes to earthquakes recorded at great epicentral distances (teleseisms) was based not on the recorded body waves, which constitute the maximum of the seismogram at short distances, but on the large surface waves, with periods of oscillation near 20 seconds, which appear as the maximum on seismograms of shallow teleseisms as written by the seismographs most in use. Restriction to particular instruments was overcome by referring the scale to the actual amplitude of the ground oscillation, as calculated from the seismogram readings and from the known characteristics of the seismograph used. The scale was adjusted as far as possible to agree with the California local scale. The resulting tabulations could be mainly represented by a simple formula. A later modification of this formula, adopted at an international conference in 1967, equates the magnitude to the logarithm of the amplitude and the distance involved in the expression: $M = \log A + 1.66 \log \Delta + 2.0$, in which A is the maximum amplitude of the ground for surface waves with periods of about 20 seconds, and Δ (delta) is the distance from epicentre to station, expressed in degrees of arc of a great circle over the Earth's surface (one degree being about 111 kilometres or 69 miles). The formula fails when Δ exceeds about 150°. At each recording station, there should be a small local correction, expressing the normal deviation from the mean of all stations; at Pasadena, California, for example, this correction is −0.1 magnitude. Other special corrections can be applied where amplitudes are found to be regularly high or low along certain paths.

Use of body waves to determine magnitude

Tables have been developed for finding magnitude from deep-origin or body waves. This required beginning with shallow earthquakes whose magnitudes had been found from surface waves; it was then extended to deep-focus earthquakes. For body waves, it is necessary to use the amplitude divided by the corresponding measured period of oscillation, T.

For theoretical reasons, the magnitude determined from body waves has been preferred by some authorities to that determined from surface waves; but many questions have arisen in general practice, and the surface-wave magnitude provides a more satisfactory standard. There are three principal difficulties:

(1) The tables yield body-wave magnitudes, m, which agree with those from surface waves, M, in the range 6.7 to 6.9, but deviate from them for larger and smaller values; the relation between m and M can be expressed as $m = 2.5 + 0.63\,M$.

(2) For deep-focus earthquakes, the tabulations involving m did not allow for loss of energy in the low-velocity zone of the mantle; the effect is to overestimate earthquakes originating below that zone.

(3) Many workers fail to consider the largest waves, but take only the first few oscillations, which usually are smaller; the magnitude then reported applies to these earlier events, with a ludicrous underestimate of the main shock (magnitudes near 5 have often been given for earthquakes or magnitude obviously 7 or over).

Magnitude, energy, and nuclear explosions

Various equations intended to relate magnitude, M, to total radiated energy, E, have been published. The one most commonly employed equates the logarithm of energy to the magnitude in the form: $\log E = 11.4 + 1.5\,M$, in which E is energy in ergs and M is magnitude. This is at best only approximate; the relation between E and M appears to be variable. The number M is probably more closely related to the maximum rate of radiation of energy, which is power, measured in kilowatts or horsepower.

Many errors have resulted from applying the earthquake-magnitude scale to seismograms of nuclear explosions. The fundamental tables, constructed for relating ground motion to distance, do not properly apply to events originating near the Earth's surface. Moreover, the process by which nuclear energy is converted into seismic waves differs greatly from the mechanism of natural earthquakes. A separate magnitude scale can be developed for comparing the seismic effects of explosions, however. Magnitudes thus found are not necessarily directly related to the yield, which is the total energy released, commonly stated in kilotons. An appreciable fraction of the yield energy is temporarily stored, elastically surrounding the explosion centre, and does not contribute to the seismograms. Some of this energy is later released during collapse events, sending out seismic waves often recorded by stations some hundreds of miles distant.

Finally, it is important to note that the magnitude scale does not supersede intensity scales; their purposes are complementary. Unfortunately, no one has yet succeeded in constructing a satisfactory intensity scale based on physical measurements.

Seismograms and their interpretation. A typical seismogram is complex; it consists of a series of characteristically successive curves that represent the arrivals of seismic waves at the site of the recorder. The curves begin sharply or gradually, each continuing for a small part of the duration of the entire recorded disturbance. It was initially supposed that this complexity arose from a succession of events at the source; but later it became clear that most of the arriving signals are the result of a single event of comparatively brief duration and that the successive arrivals are caused by waves that started at the same time but that travelled over differing paths with various speeds.

P, *S*, and *L* waves

Analysis of seismograms began by empirically separating the recorded motions into two groups. Of these, one is a train of relatively large, slow oscillations, which formed the most conspicuous part of the records of normal large earthquakes as recorded by the instruments first in general use; this group came to be called the "principal earthquake." With improved recording, these oscillations were observed to be normally preceded by some minutes by smaller and more rapid vibrations, at first called the "preliminary tremor." The latter could often be subdivided into two succeeding groups, termed the first and second preliminary tremor. These two groups were designated by *P* (initial of *primae*) and *S* (*secundae*). Current writers sometimes call these "primary" and "secondary" waves, which is slightly misleading, because nothing more than the succession in time is implied.

The large waves of the "principal earthquake" were symbolized by *L* (for *longae*) because of their long periods of oscillation. They represent energy travelling around the earth, chiefly near its surface, whereas *P* and *S* are "body waves," propagated by the interior of the earth. Because the body waves have a shorter path and generally travel with higher velocity than surface waves, they often reach the recording station many minutes earlier. The velocity of *P* waves in several earth materials is shown in Figure 4.

P and *S* correspond to the two principal types of elastic waves that can exist in a solid body. In the *P* type, the oscillation of individual particles is along the ray, in the direction of propagation, and perpendicular to the wave front, just as for sound waves in air. The *P* waves are, in fact, essentially sound waves in the Earth, although their frequency of vibration is mostly well below the range for audible sound. Oscillation of *S* waves is along the wave front or perpendicular to the direction of propagation. Whereas *P* waves are longitudinal in vibration, *S* waves are transverse; this gives *S* waves some geometrical properties like those of light waves, so that they show such effects as polarization.

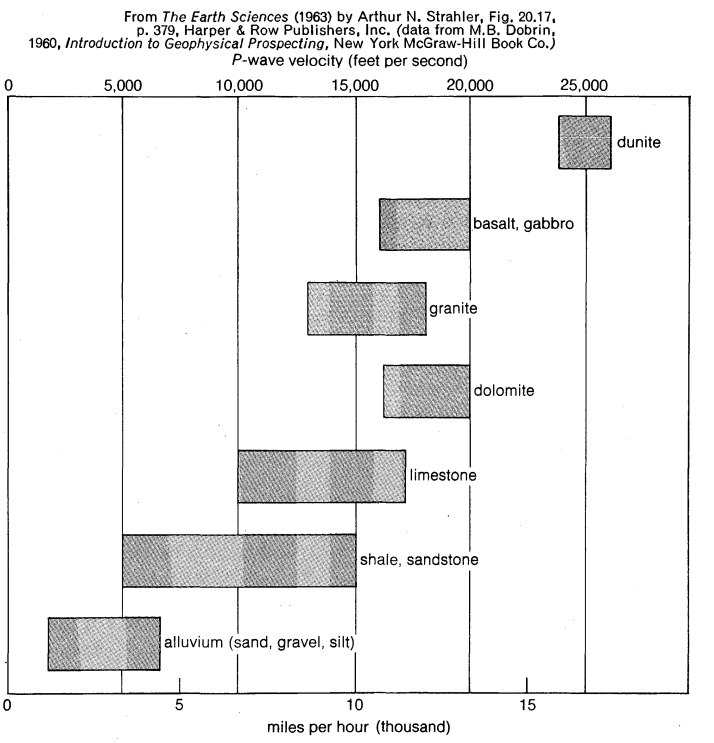


Figure 4: Velocity of propagation of *P* waves in various earth materials.

P waves are accompanied by compression and rarefaction, like ordinary sound waves; but *S* waves produce local shearing and deformation without change of volume. Because liquids are incapable of supporting shear, *S* waves can pass only through solids; in this manner, it can be concluded that the outer part of the earth's core is liquid, because *S* waves do not pass through it.

Wave motion in the core and mantle

The different recorded groups of body waves are designated by letter symbols that represent successive segments of the path traversed. For convenience, waves of longitudinal or *P* type within the core (or specifically in its outer portion) are assigned the letter *K*. Each such wave that has traversed both mantle and core is lettered *PKP*, showing that it begins and ends as *P* in the mantle. Longitudinal waves in the inner core are separated by the letter *I;* thus a wave passing through the whole earth along a diameter is called *PKIKP*.

Whereas seismic waves in the Earth are everywhere of the two main types, *P* (with *K* and *I*) and *S* (plus a hypothetical, *J*, for possible transverse waves in the inner core), analysis of seismograms confirms many possibilities because of reflection and refraction, mainly at the lower surface of the crust or at the interface between the core and mantle. Careful and difficult analysis also is being directed toward identifying reflections and refractions at the boundary that exists between the inner and outer cores.

The number of observable waves traversing the mantle alone is augmented by transformation from *P* to *S*, or the reverse, upon reflection. A single reflection at the surface

may result in *PP*, *PS*, *SP*, or *SS*, the letters being in the order in which the two parts of the path are traversed. Observable double reflections are *PPP*, *PPS*, *PSP*, *PSS*, *SPP*, *SPS*, *SSP*, *SSS*. Some of these (such as *PPS*, *PSP*, *SPP*) have identical travel times for a surface source but separate times for earthquakes originating at appreciable depth.

Reflections from the outside of the core are often observed; to distinguish these, an index, *c*, is inserted, so that the single reflections are *PcP*, *PcS*, *ScP*, and *ScS*.

Reflections from inside the core boundary are observed with sensitive recorders; the most common are *PKKP* and *SKKP*, in which repetition of *K* indicates the reflection. *PKP* may be reflected at the surface, so that it is possible to observe *PKPPKP*, and even *PKPPKPPKP*, as well as *SKPPKP*.

The most important refracted body wave is *PKP*. The drop in velocity at the surface of the core may greatly deflect this ray; there is a focussing effect (technically a caustic) at angular distances of about 142° from source to station. *SKP* begins as a transverse wave but becomes longitudinal in the core and thereafter; it has a similar focus or caustic at about 135°. *SKS*, which is transverse in the mantle and longitudinal in the core, is conspicuous on seismograms at distances from 85° to 100°, where it precedes *S* and confuses observation of that wave.

Surface waves are also of several kinds. These are not as clearly separable as the types of body waves; their generally greater wave lengths lead to complications not indicated by the simplified theories first applied. Ideally, there are two main types, associated with the personal names of two scientists who investigated these waves, Rayleigh and Love. These are related somewhat as *P* to *S* waves. Oscillation of the theoretically ideal Rayleigh wave should take place in a vertical plane through the source and the point of observation, with both vertical and horizontal displacements; those of the ideal Love wave should be horizontal, transverse to the direction of propagation.

Rayleigh waves and Love waves are instances of channel or boundary waves, associated with horizontal propagation in a limited range of levels, which may be near the surface of the Earth or along some deeper horizon, like low-velocity level in the mantle. Certain channel waves produced by earthquakes in the waters of the ocean, impinging on the continental coasts, produce recordings of a peculiar type called *T* waves.

A typical seismogram

The following is an abbreviated account of the main features of the seismogram of a typical shallow earthquake, preceding the arrival of surface waves:

At relatively short distances, up to about 1,000 kilometres (9°) from the epicentre, the record shows principally the *P* and *S* groups, separated by an interval ranging from a few seconds up to about 100 seconds. Within the *P* and *S* groups there is complexity because of the overlap of successive arrivals, caused by discontinuities at the base or structures within the Earth's crust. Beyond 1,000 kilometres, complexity originates chiefly from various internal reflections and refractions.

Distances are now given in degrees of arc from epicentre to stations over the Earth's surface, one degree being about 111 kilometres or 69 miles:

9°–85°. Principal wave groups are *P* and *S*, but at about 25°–35° there are often sharp arrivals of the core reflections *PcP*, *ScP*, *ScS*. Over continental paths *PP* may be large and conspicuous.

85°–105°. In this range *P* is clear, but identifying *S* often is difficult because of *SKS* immediately preceding and *PS* immediately following.

105°–142°. Here *P* is usually lost in the shadow of the core, which shuts off the rays as they dip deeper into the earth. In this shadow zone seismographs show *P* waves diffracted around the core boundary; also, there are occasional sharp recordings of *PKP* waves that have been widely deflected by the boundary of the inner core. *SKP*, and shortly after it, *PKS*, arrive with large amplitudes near 135°. *PKKP* is most evident near 120°. *PP* and *PS* are large and sometimes are mistaken for *P* and *S*, implying much too short a distance.

142°–180°. Near 142°, *PKP* arrives with strikingly large amplitude. It is followed, at an interval increasing with distance, by *PKP* that has travelled around through the core and covered a range of central angle of more than 180°. There are few conspicuous waves following *PKP*, though *SKKS* is sometimes large enough to be taken for *S*. Reflections such as *SS* and *SSS* may appear as large waves; finally, the surface waves begin, often more than an hour after the initial *PKP*, so that the recording looks like that of two separate disturbances.

BIBLIOGRAPHY. The nature of seismic waves and their measurement and interpretation are topics that are treated in all texts on seismology. Among such works the interested reader should consult KEITH E. BULLEN, *Seismology* (1954), and *An Introduction to the Theory of Seismology*, 3rd ed. (1963); MILTON B. DOBRIN, *Introduction to Geophysical Prospecting*, 2nd ed. (1960), which gives an elementary account of gravity, magnetics, and other geophysical methods, in addition to seismology; and CHARLES F. RICHTER, *Elementary Seismology* (1958). The relationship of earthquakes to the Earth's structure and composition is treated in the related articles cited in the introduction to this article. Worthy of note as references in this vein are JOHN H. HODGSON, *Earthquakes and Earth Structure* (1964); BENO GUTENBERG and CHARLES F. RICHTER, *Seismicity of the Earth and Associated Phenomena*, 2nd ed. (1965); and the collected geophysical papers in P.J. HART (ed.), *The Earth's Crust And Upper Mantle* (1969).

(C.F.R.)

Earth Sciences

The systematized knowledge of the solid Earth and the fluids within and around it constitutes the Earth sciences. Three main families of sciences are involved; namely, the geological sciences, which deal with the rocky crust of the Earth and its interior; the hydrologic sciences, which are concerned with all the waters of the Earth; and the atmospheric sciences, which encompass the physical and chemical properties of the atmosphere as well as air and its motions.

This article is concerned with the characteristics and contributions of these families of Earth sciences and with their relationships to each other and to other disciplines as perceived from a historical viewpoint. Accordingly, the principal theme of this article is the development of ideas and concepts that are central to the Earth sciences. For further coverage of the scope and methods of the many disciplines involved, see GEOLOGICAL SCIENCES; HYDROLOGIC SCIENCES; and ATMOSPHERIC SCIENCES. The actual subject matter or content of any particular discipline is presented in the many articles on the Earth and its development—*e.g.*, MINERALS; OCEANS AND SEAS; WINDS AND STORMS.

This article is divided into the following sections:

I. Nature and origins of the Earth sciences
  Physical, historical, and applied aspects
  Foundations in myth and legend
  Central ideas and observations
II. From antiquity to the 16th century
  Geological sciences
  Hydrologic and atmospheric sciences
III. The 16th, 17th, and 18th centuries
  Geological sciences
  Hydrologic sciences
  Atmospheric sciences
IV. The 19th century
  Geological sciences
  Hydrologic sciences
  Atmospheric sciences
V. The 20th century
  Geological sciences
  Hydrologic sciences
  Atmospheric sciences
  Emerging trends

I. Nature and origins of the Earth sciences

PHYSICAL, HISTORICAL, AND APPLIED ASPECTS

Each of the three families of Earth sciences is conventionally divided into disciplines and subdisciplines according to its particular subject matter. More than 100 special fields of Earth science have already been named, and the definition and classification of all these specialties lie beyond the scope of this article. For convenience in

discussion the special subject fields may be grouped as physical sciences, historical sciences, and applied sciences. Those fields dedicated primarily to understanding the dynamical, physical, or chemical properties of natural phenomena under their purview will be called physical sciences. Those concerned with the sequential changes that have taken place on the Earth will be called historical. Fields dealing with the application of scientific information and historical inferences to the solution of practical problems will be called applied sciences; the latter can also be termed technologies.

Threefold groupings in hydrologic, atmospheric, and geological sciences

To illustrate this threefold grouping, meteorology is the science of the atmosphere; paleoclimatology, the reconstruction of ancient weather conditions over long average time periods, is its historical counterpart; and weather forecasting is one of the practical applications of this science. Hydrology is the science of the waters that occur on the Earth. Physical oceanography, limnology, potamology, and hydrogeology are subdivisions dealing, respectively, with oceans, lakes, rivers, and water contained in the voids in rocks. Glaciology is a special field concerned with the frozen water of glaciers. Paleohydrology is a historical branch, dealing generally with the origin, chemical evolution, and migration of the Earth's water through time; it includes special fields such as paleolimnology, the reconstruction of conditions that obtained in ancient lakes. Among the applications of hydrology, prospecting for and conservation of groundwater comprise two prominent subjects.

In geology, the scientific disciplines are often grouped under the heading of physical geology, which includes the sciences of minerals (mineralogy), rocks (petrography and petrology), structures in the solid Earth (structural geology), sedimentary strata and their fossil content (stratigraphy and paleontology, respectively), and landforms (geomorphology). The history of the development of these natural entities is broadly comprehended under the heading of historical geology. Applied geology, or economic geology as it is often called, includes mining geology, engineering geology, and petroleum geology.

Scope of geophysics and geochemistry

Geophysics and geochemistry are major branches of Earth science that overlap many of the special fields named above. Broadly speaking, geophysics is concerned with the physical properties and dynamics of the Earth in all its parts. Among the phenomena investigated by geophysicists are the Earth's fields of gravity and magnetism, the thermal properties of the Earth, the motions of fluids within and surrounding the solid Earth, and the release of mechanical and thermal energy in earthquakes and volcanoes. Geochemistry deals with the composition of the planet and with the chemical changes that occur in nature. Although both of these geosciences are concerned with changes affecting the Earth at present, they also have the aim of discovering the chains of events that led up to present conditions. Much effort is currently directed toward reconstructing the thermal history of the Earth, for example, which is envisioned as a heat machine with a finite store of energy ultimately to be dissipated. Paleomagnetics is a flourishing new field dealing with past changes in the Earth's magnetic field. Working in laboratories where the behaviour of Earth substances subjected to high temperatures and pressures simulates changes presumed to obtain at great depth, the geochemist seeks to understand how ancient rocks in the crust of the Earth were formed. The practical applications of these two major geosciences are also numerous. Geophysical prospecting, using instruments designed to measure differences in the magnetic or gravimetric fields over small areas or to measure variations in the rate of propagation of elastic waves through different kinds of rock layers, has successfully guided exploration for a variety of things, including the cannon of sunken Spanish galleons and vast subterranean supplies of petroleum and natural gas. Variations from place to place in the chemical composition of plants and of the soils in which they grow have also been used in prospecting for mineral deposits. In some instances, variations in the concentration of certain elements in the plants eaten or the water drunk by animals may be correlated with regional differences in the incidence of specific diseases and malfunctions. In the emerging subdiscipline of medical geology, geochemistry promises to play an important role.

FOUNDATIONS IN MYTH AND LEGEND

The countless special questions for which Earth scientists seek answers reduce to a few general queries, such as the following: Of what is the Earth made, and how are its parts connected? How old is it, and how did it come to its present condition? What causes the changes of landscape that go on all the time? What is man's place in the order of nature? What of the Earth is useful or otherwise valuable? These are among the perennial questions for which answers were attempted long before the Earth sciences existed, and the prescientific answers often took the form of creation stories and myths.

Stories of the creation

The creation story that undoubtedly has played the most conspicuous part in the debate about the Earth from which the historical sciences have grown is the one told in the first chapter of Genesis. Biblical scholars trace elements of the Mosaic account back to the Babylonian creation epic (*Enuma elish*), which was current in the 22nd century BC. As known from clay tablets found in the ruins of Ashurbanipal's library at Nineveh, the Babylonian epic interprets the creation of the world in terms of a conflict between deities. In the beginning, Tiamat, a monstrous goddess of darkness, ruled over a chaotic world. Marduk, supreme god of the Babylonians, led a host of other deities from their home in the midst of heaven to wage war against Tiamat. Marduk slew Tiamat and cut her body into two parts. One part he raised up to form the sky, a firm barrier that then separated the "upper waters" from the waters of the Earth. Then Marduk created the luminous bodies of the heavens, arranging the stars in patterns to resemble the images of the great gods. The Moon god he next created and appointed to watch over the night and keep track of the days. Like the Babylonian, the Hebraic account begins with a world in chaos and darkness. Light was created from the darkness. A firmament, or sky, was created to separate the upper from the lower waters, allowing rain to fall only when the windows of heaven were opened. The lower waters gathered into one place; dry land emerged, and vegetable life grew on it. The luminous bodies of heaven were created for the ordering of time as days and nights, seasons and years. In the last stage, animal life was placed on the Earth, and finally man was created.

The Mosaic account of creation has been identified as a work written during the Babylonian Exile in defense of monotheism against the Babylonian view that the world originated through a conflict of deities. Granting the motivation, the creation story is still cast in the form of Earth history, in which the successive events are attributed to supernatural causes. The ideas that chaos preceded order, that plants preceded animals, and that man appeared late on the scene can also be fitted into purely naturalistic versions of cosmogony and geology. Thus the creation story came to be accepted as an embodiment of scientific as well as theological truth and is still so interpreted in some quarters today.

Poseidon as the cause of earthquakes

Less pretentious in scope and intent than the creation stories are the myths that relate specific natural phenomena to the work of supernatural agencies. Spectacular events, such as earthquakes and volcanic eruptions, have evoked images of powerful and capricious gods and goddesses. Thus the early Greeks attributed earthquakes to the angry outbursts of the sea god Poseidon (the Roman Neptune). With his three-pronged spear he could control the waves of the sea, shatter rocks, and shake the shores. One Earth-shaking episode has Poseidon pursuing Polybotes, an upstart giant who defied the gods. When the chase led past Cos, the sea god tore off a piece of the island and hurled it at Polybotes, burying him beneath it. Poseidon, not content with his ocean domain, contested with other gods for possession of the land. He was also the ruler of springs and turned them on and off at will. Poseidon was worshipped along the shores and at inland temples in regions affected by earthquakes.

Pele, the volcano goddess

The old Hawaiians believed that the active volcano Kilauea, on the island of Hawaii, is the home of a family of fire deities headed by the goddess Pele. From her original home in the western Pacific, Pele was driven by her sister and implacable enemy, Na-maka-o-ka-hai, goddess of the sea, who persistently attacked Pele's home with floods and high tides. Pele fled first to the northwestern end of the Hawaiian Islands, where she attempted to build a new home on the lowlands of Kauai. With her magical digging tool, the *pa-oa*, she struck down into the Earth to make fire pits in which she could take refuge. But water quickly quenched the fires she kindled. So Pele went southeastward from island to island trying to dig other fire pits, but these too were so near the sea that water rushed in and put out the fire in great explosions of steam. Finally Pele found a place on Kilauea where she could make an enduring fire pit and live in safety with friendly members of her family. Na-maka-o-ka-hai, seeing the great fiery clouds rising from the pit of Halemaumau, gave up the chase. There Pele lives today. A petulant goddess, Pele sometimes creates new land with lava flows; at other times she sends down streams of lava to destroy the dwellings of her kinsmen. Whenever she stamps her feet in anger the Earth quakes.

The stories of Poseidon and Pele are examples of myths that relate natural occurrences in a rational way, up to the point at which they invoke supernatural forces as ultimate causes. Both myths relate the retreat of cliffy shores to attacks by waves. The Poseidon myth rightly connects earthquakes with local changes in the flow of streams and springs. The Pele myth properly associates earthquakes with volcanic outbursts and the construction of ash cones to the meeting of fire and water. Moreover, the Hawaiians seemed to have grasped that the deeply dissected Kauai, Pele's early home, is older than Hawaii, where Kilauea is still in the constructional phase of its development. The mythic element enters where the storyteller deifies the forces that move the waves, shake the land, and start the volcanic fires.

CENTRAL IDEAS AND OBSERVATIONS

Earth scientists contend with two difficulties inherent in their subject. Most of the Earth is inaccessible for direct observation. And the history of the Earth is not given but must be inferred from present configurations of matter, present transformations of energy, and from the changes observed to have taken place on Earth during the brief span of recorded history. Despite these difficulties, three major ideas have been shaped that profoundly modified ancient and traditional views of nature. The first is that the Earth is mobile, always changing in its configurations. The second is that the world is almost inconceivably ancient. The third is that prehistoric life, as known from fossils, has changed with time and that the usual outcome of these changes has been the extinction of species.

Cartographers and mariners: Earth configuration

The image of the accessible world first began to emerge from the efforts of cartographers and mariners. Scientific cartography of the ancient world reached its peak in the work of Ptolemy (Claudius Ptolemaeus; flourished *c.* AD 120–170). Using the astronomical and mathematical principles then available, Ptolemy produced a map covering about 180° of longitude and 80° of latitude centring near the Persian Gulf. Soon afterward, religious cosmography took over in the West, and until the 15th century the world was commonly drawn as a circular disk, oftentimes with Jerusalem at the centre and the Garden of Eden and Paradise also indicated. Interest in cartography revived with the great voyages of exploration beginning in the 15th century. Mercator's chart of the world, produced in 1569, shows fairly well the congruency of the opposing coasts of South America and Africa, a symmetry that later led to the theory of continental drift, the idea that the continents are separating as the ocean basins expand through time. Thereafter maps have steadily improved and have been used at bases for regional analysis of all kinds of geologic, hydrologic, and atmospheric data.

Forebears of modern oceanographers, the mariners added much from their personal experiences to knowledge of the waves, currents, tides, and variations in the temperature and saltiness of seawater. Their success at navigation depended largely on the reliability and capability of their instruments. When Magellan set out on his voyage in 1519 he had at hand a terrestrial globe, theodolite and quadrants, compasses and hourglasses. Then and long afterward one of the hazards to navigation lay in faulty determination of longitude, whose calculation requires knowledge of the precise time of observation. Following a series of ship disasters around 1690, the Board of Longitude was set up in England, and encouragement was given to the development of better timekeeping devices. With the improved chronometers developed by 1764, Captain James Cook set out exploring the South Pacific in 1768, voyages that mark the beginning of modern navigation. Cook's soundings to 200 fathoms (1,200 feet) and observations of winds, currents, and temperature of seawater have since been praised for their accuracy.

The obvious mobility of the atmosphere and hydrosphere only accentuates the seeming stability of the solid Earth. Ancient observers knew, of course, about the dislocations in the ground that sometimes attend earthquakes. But the first evidence for profound changes in the position of land and sea was provided by marine fossils contained in the rocks of high mountains. This evidence, however, was not generally accepted at face value until after the 17th century. The motive force for elevation of the continents itself remained mysterious until late in the 18th century, when the solid Earth was envisioned as a machine driven by internal thermal energy. The source of this energy was assumed to be solely residual from the Earth's original molten state until a century later, when it was discovered that heat is generated by radioactivity.

The antiquity of the Earth has long been a subject of debate argued both from scientific and theological grounds. Until well into the 18th century the short Christian chronicle held sway; but geologists of the 19th compiled such lengthy sequences of terrestrial events that to some of them geological time seemed virtually limitless. Geochemical dating of radioactive minerals has finally shown that the Earth must be at least 4,600,000,000 years old.

Before the middle of the 19th century the order of succession of different kinds of fossils in rock strata was fairly well-known. Moreover, it was recognized that nearly all species described by paleontologists are extinct. An explanation of the developmental trends observable in the fossil record was not forthcoming, however, until the appearance on the scene of the evolutionary biologists led by Charles Darwin.

Contributions from chemistry, physics, and technology

Knowledge of the composition and dynamical characteristics of the Earth has come mostly from applications of chemistry and physics after the latter part of the 17th century and from technology in the form of improved scientific instruments. Chemical oceanography began with investigations by Robert Boyle around 1670, and atmospheric chemistry developed about 100 years later in the work of Carl Scheele and Henry Cavendish. Chemical classifications of minerals were standard after 1850, and the later development of physical chemistry made possible modern experimental mineralogy and petrology. Sir Isaac Newton's application of the law of gravitation to the behaviour of the tides in 1684 was a landmark in the history of geophysics, and it led to Pierre-Simon Laplace's dynamical theory of tides in 1775. Knowledge of what is supposed to exist far below the Earth's surface rests on physical principles relating to gravity, magnetism, thermodynamics, and the propagation of energy in elastic media.

The organization of scientific societies has given impetus to the development of science generally. Perhaps not coincidentally, the 17th century, which marks the beginning of scientific societies in Italy, France, and England, was also the time when a host of new instruments were developed and improved for research. In addition to the telescopes and microscopes that extended the range of human vision, there were pendulum clocks, barome-

ters, and thermometers, which sharpened observations and experiments by yielding quantitative data. These were the forerunners of the myriad instruments subsequently developed to quantify information about the Earth and thus make data amenable to mathematical analysis and statistical study.

The progress of science is constrained by whatever society will tolerate or support, and the utilitarian aspects of science have ranked high in the popular view since the 17th century. Atmospheric science has always been overshadowed by its practical application in forecasting weather. At the beginning of the 19th century, when European travel was by sailing ship, horseback, or coach, a great interest in the weather led to establishment of meteorological departments. But interest lagged for a time with the coming of the steamer and railway. The greatest geological effort of the 20th century has been the exploration for fossil fuels, and no doubt the present acceleration of interest in oceanography is prompted by hopes of developing additional stores of food and minerals. This is not to deprecate applied science. Petroleum geology, for example, has paid handsome scientific dividends, among other matters having added the third dimension of depth to the geological map in many remote and relatively inaccessible parts of the world. Techniques of drilling developed in search of oil have been applied to coring the ocean's floor for purely scientific purposes, and exploration geophysics has added much to knowledge of the Earth's interior.

II. From antiquity to the 16th century

GEOLOGICAL SCIENCES

Classification of rocks and minerals by Theophrastus

Knowledge of Earth composition and structure. The oldest known treatise on rocks and minerals is the *De lapidibus* of Theophrastus (*c.* 372–287 BC). Written probably in the early years of the 3rd century, this work remained the best study of mineral substances for almost 2,000 years. Although reference is made to some 70 different materials, the work is more an effort at classification than systematic description. Theophrastus recognized two main categories: stones and earths. Stones are classified according to their colour, tenacity, texture, density, fusibility, inflammability, solubility in water, magnetic properties, utility, and value. He identified the essential characteristics of building stones and mentioned localities where different kinds of building stones are found. Both organic and inorganic substances were included with the stones. Coal and asphalt, which will burn, he contrasted with substances such as red garnet, which are seemingly unaffected by fire. The organic amber is mentioned in context with the inorganic magnetite (iron oxide), because both have the ability to attract foreign objects: straws and bits of wood by rubbed amber and iron by the lodestone. The treatise ends with a brief account of "earths," which include substances then in common use as pigments, cosmetics, and insecticides. Red ochre, yellow ochre, vertigris, lapis lazuli, realgar, orpiment, and cinnabar are listed among the pigments. White lead, produced by the action of acetic acid vapour on lead, was used both as a cosmetic and as a pigment. Boiled bituminous clay and sand is reported as effective against the predations of woodworms on vines. In discoursing on stones, Theophrastus casually included the artificial along with the natural. His category *gypsos*, for example, includes the manufactured plaster of paris and quicklime as well as ordinary gypsum.

Chinese writings on minerals and ore deposits

In early Chinese writings on mineralogy, stones and rocks were distinguished from metals and alloys, and further distinctions were made on the basis of colour and other physical properties. Systematic treatments of mineral substances are found in pharmacopoeias, in works on alchemy, and in treatises describing different kinds of rocks used as inkstones (for grinding blocks of carbon-black ink with water). In the pharmacopoeias the minerals were sometimes ranked according to their imputed therapeutic value, the number of mineral substances in works of this kind ranging from 46 in writings from near the beginning of the Christian Era to 215 in a list from the early 12th century. The speculations of Cheng Ssu-hsiao (died AD 1332) on the origin of ore deposits were more advanced than those of his contemporaries in Europe. In brief, his theory was that ore is deposited from groundwater circulating in subsurface fissures. He envisioned the rocks below the surface as traversed by a vast network of cracks filled with fluids. It was, he wrote, as though the Earth were "hanging and floating on a sea boundlessly vast." Ore was thought to be formed by evaporation or precipitation from the fluids underground.

Speculations on earthquakes and volcanic eruptions

Ancient accounts of earthquakes and volcanic eruptions are sometimes valuable as historical records but tell little about the causes of these events. Aristotle held that volcanic explosions and earthquakes alike are caused by spasmodic motions of hot winds that move underground and occasionally burst forth in volcanic activity attended by Earth tremors. Strabo (64 BC–*c.* AD 21) subscribed to the same idea and went on to suggest that volcanoes act as safety valves to vent the explosive force of winds pent up within the Earth. Classical and medieval ideas on earthquakes and volcanoes were brought together in William Caxton's *Mirrour of the World* (1480), an influential work and one of the earliest encyclopaedias in the English language. Earthquakes are here again related to movements of subterranean fluids. Streams of water in the Earth compress the air in hidden caverns. If the roofs of the caverns are weak, they rupture, causing cities and castles to fall into the chasms opened thereby; if strong, they merely tremble and shake from the heaving by the wind below. Volcanic action follows if the outburst of wind and water from the depths is accompanied by fire and brimstone from hell.

Chinese invention of the seismograph

The Chinese have the double distinction of keeping the most faithful records of their earthquakes and of inventing the first instrument capable of detecting them. Records of the dates on which major quakes rocked China go back to 780 BC. In order to detect quakes at a distance, the mathematician, astronomer, and geographer Chang Heng (AD 78–139) invented an instrument that has been called the first seismograph. This consisted of a large bronze vessel containing a heavy pendulum. Arms radiating in eight directions of the compass were levered in the walls of the vessel so that they could be moved by the first swing of the pendulum. When the pendulum was set in motion by earthquake waves, one or another of the arms would be moved, releasing a bronze ball that would fall into a cup attached to the bottom of the vessel. With this release the pendulum was immobilized. The position of the ball that had fallen permitted an estimation of the direction from which the quake had come.

Knowledge of Earth history. The occurrence of seashells embedded in the hard rocks of high mountains aroused the curiosity of early naturalists and eventually set off a controversy on the origin of fossils that continued through the 17th century. Xenophanes of Colophon (flourished *c.* 560 BC) was credited by later writers with observing that seashells occur "in the midst of earth and in mountains," that the prints of fish and of seals are to be found in rocks quarried at Syracuse, Sicily, and that there are all kinds of marine animals in the rocks of Malta. He is said to have believed that these relics originated during a catastrophic event that caused the Earth to be mixed with the sea and then to settle out, burying the organisms in the drying mud. For these views Xenophanes is sometimes called the father of paleontology. Theophrastus is the author of a work on "fishes remaining in a dry condition," a description of fossils found in rocks to the south of the Black Sea. Herodotus and Strabo both accepted the organic nature of fossils and used these in turn as evidence of past changes in sea level. Throughout the Middle Ages, however, fanciful ideas on the nature of fossils were more popular among European authors, or, if the organic hypothesis was entertained, the fossils were usually treated as victims of Noah's Flood.

Fossil evidence of changes in sea level and climate

Petrified wood was described by Chinese scholars as early as the 9th century AD, and around 1080 Shen Kua cited fossilized plants as evidence for change in climate. Stands of petrified bamboo had been uncovered in alluvium near Yen-chou. Shen Kua marveled at this, be-

cause bamboo does not now grow in this area. He suggested that "in very ancient times the climate was different so that the place was low, damp, gloomy, and suitable for bamboos." Other kinds of fossils that attracted the attention of early Chinese writers include spiriferoid brachiopods ("stone swallows"), cephalopods, crabs, and the bones and teeth of reptiles, birds, and mammals ("dragon's bones and dragon teeth"). Although these objects were commonly collected simply as curiosities or for medicinal purposes, Shen Kua (*c.* 1070) recognized marine invertebrate fossils for what they are and for what they imply historically. Observing seashells in strata of the T'ai-hang Shan-mo range, he concluded that this region, though now far from the sea, must once have been a shore. What is today a continent, he observed, must have been of mud and other sediment that was once below the water. The work of the Huang Ho (Yellow River) and the other great rivers of China in carrying sediment to the sea and depositing it there shows how the whole continent must have been laid down.

Knowledge of landforms and of land–sea relations. Changes in the landscape and in the position of land and sea related to erosion and deposition by streams were recognized by a number of the early writers. Herodotus (*c.* 484–426 BC) correctly concluded that the northward bulge of Egypt into the Mediterranean is caused by the deposition of mud carried by the Nile. He records that a sounding line let down in 11 fathoms (66 feet) of water from a ship still a day's sail from the delta brought up mud like that deposited by the Nile during its annual floods. The flatland of the delta and the narrowing floodplain to the south, he reasoned, are "the gift of the Nile," which has filled an ancient gulf with its sediment. He cites marine fossils in strata of northern Egypt as evidence for the former existence of the gulf. Herodotus also cites other areas where smaller rivers draining into the sea have built new land as alluvial plains and, in some instances, have tied former islands to the mainland. Strabo generalized the view of Herodotus on the work of streams in stating that

Formation of river deltas and alluvial plains

> all streams imitate the Nile in that they keep converting the channel just in front of them to land . . .; to a less extent those that do not bring down much mud but to a greater extent those that flow for a great distance through a country with a soft soil and have many torrents as tributaries.

Strabo was willing to believe that streams draining Asia Minor may in time build deltas connecting Cyprus with the mainland.

The early Chinese writers were not outdone by the Romans and Greeks in their appreciation of changes wrought by erosion. In the *Chin shu* ("History of the Chin Dynasty") it is said of Tu Yü (AD 222–284) that when he ordered monumental stelae (huge rock slabs) to be carved with the records of his successes, he had one buried at the foot of a mountain and the other erected on top. He predicted that in time they would likely change their relative positions, because the high hills will become valleys and the deep valleys will become hills.

The concept of cyclicity

Aristotle guessed that changes in the position of land and sea might be cyclical in character, thus reflecting some sort of natural order. If the rivers of a moist region should build deltas at their mouths, he reasoned, seawater would be displaced and the level of the sea would rise to cover some adjacent dry region. A reversal of climatic conditions might cause the sea to return to the area from which it had previously been displaced and retreat from the area previously inundated. The idea of a cyclical interchange between land and sea was elaborated in the *Discourses of the Brothers of Purity*, a classic Arabic work written between AD 941 and 982 by an anonymous group of scholars at Basra (Iraq). The rocks of the lands disintegrate and rivers carry their wastage to the sea, where waves and currents spread it over the sea floor. There the layers of sediment accumulate one above the other, harden, and, in the course of time, rise from the bottom of the sea to form new continents. Then the process of disintegration and levelling begins over again. Cycles of this kind are also mentioned in the Chinese literature. In the 12th century Chu Hsi envisioned that episodes of uplift of mountains and drowning of lowlands were sudden and catastrophic. According to Joseph Needham, the leading Western scholar on the history of Chinese science and technology, Chinese ideas of periodic cataclysms in which the world has been destroyed and created anew are of Indian origin, probably brought to China by the Buddhists.

HYDROLOGIC AND ATMOSPHERIC SCIENCES

The only substance known to the ancient philosophers in its solid, liquid, and gaseous states, water is prominently featured in early theories about the origin and operations of the Earth. Thales of Miletus (*c.* 624–*c.* 545 BC) is credited with a belief that water is the essential substance of the Earth. Anaximander of Miletus (*c.* 610–545 BC), although unwilling to concede that water or any other tangible substance could be considered the ultimate source of all things earthly, held that water was probably the source of life. Animals, he proposed, originated in water and later migrated to the land, where they shed their outer coverings (husks or shells) and otherwise adapted to meet the new conditions. In the system proposed by Empedocles of Agrigentum (*c.* 490–430 BC) water shared the primacy Thales had given it with three other elements: fire, air, and earth. The doctrine of the four earthly elements was later embodied in the universal system of Aristotle (384–322 BC) and thereby influenced Western scientific thought until late in the 17th century.

Knowledge of the hydrologic cycle. The idea that the waters of the Earth undergo cyclical motions, changing from seawater to vapour to precipitation and then flowing back to the ocean, is probably older than any of the surviving texts that hint at or frame it explicitly. Xenophanes held that the source of precipitation is to be sought in the sea, which is "the begetter of clouds and winds and rivers." Aristotle was more specific when he wrote in his *Meteorologica* that by the Sun's heat

> the finest and sweetest water is every day carried up and is dissolved into vapor and rises to the upper region, where it is condensed again by the cold and so returns to the earth. . . . So the sea will never dry up; for before that can happen the water that has gone up beforehand will return to it.

The idea of the hydrological cycle developed independently in China as early as the 4th century BC and was explicitly stated in the following passage from *Lü-shih Ch'un Ch'iu* ("The Spring and Autumn Annals of Mr. Lü"), written in the 3rd century BC:

> The waters flow eastward from their sources, resting neither by day nor by night. Down they come inexhaustibly, yet the deeps are never full. The small [streams] become large and the heavy [waters in the sea] become light [and mount to the clouds].

Origin of rivers, springs, and precipitation

A circulatory system of a different kind, involving movements of water on a large scale within the Earth, was envisioned by Plato (*c.* 428–348/347 BC). In one of his two explanations for the origin of rivers and springs, he described the Earth as perforated by passages connecting with Tartarus, a vast subterranean reservoir. Upsurges in Tartarus force water through the connecting cavities and out into seas, lakes, rivers, and springs. The waters of streams ultimately drain back into Tartarus, thus completing the cycle. In Caxton's *Mirrour* the circulation of water in the hidden passages of the Earth is compared with the circulation of blood.

> Just as the blood of a man runs through the veins of the body and issues forth at a particular place, so the water runs through the veins of the earth and comes to the surface in springs and fountains.

A coherent theory of precipitation is found in the writings of Aristotle. Moisture on the Earth is changed to airy vapour by heat from above. Because it is the nature of heat to rise, the heat in the vapour carries it aloft. When the heat begins to leave the vapour, the vapour turns to water. The formation of water from air produces clouds. Heat remaining in the clouds is further opposed by the cold inherent in the water and is driven away. The cold presses the particles of the cloud closer together, restoring in them the true nature of the element water.

Water naturally moves downward, and so it falls from the cloud as raindrops. Mist is what is left behind after rain has fallen. Snow falls from clouds that have frozen.

In Aristotle's system, the four earthly elements were not stable but could change into one another. If air can change to water in the sky, it should also be able to change into water underground. Streams head in mountains, which act like sponges soaking up air. This air, losing its heat and taking up cold, turns to water, which flows out as springs to nourish streams in their headwaters. Streams, of course, get a part of their water from rain, but since perennial streams flow on sunny as well as rainy days, they must have more reliable sources in the Earth.

The origin of the Nile. Of all the rivers known to the ancients, the Nile was most puzzling with regard to its sources of water. Not only does this river maintain its course up the length of Egypt through a virtually rainless desert, but it rises regularly in flood once each year. Along the southern boundary of Egypt the annual rise begins in June and reaches a maximum in early September. Speculations on the queer behaviour of the Nile were many, varied, and mostly wrong. Thales suggested that the strong winds that blow southward over the delta in summertime hold back the flow of the river and cause the waters to rise upstream in flood. A later variant on this theme was proposed by the Venerable Bede (AD 672/673–735), who thought that these same winds produced waves in the Mediterranean, which in turn build sandbars across the mouths of the distributaries, thus backing up the waters to cause floods. Oenopides of Chios (flourished *c.* 475 BC) thought that heat stored in the ground during the winter dries up the underground veins of water so that the river shrinks. In the summer the heat disappears, and water flows up into the river, causing floods. In the view of Diogenes of Apollonia (flourished *c.* 435 BC), the Sun controls the regimen of the stream. During the hot summers the Sun evaporates water from the surface, drawing up new sources of water from underground passages that connect ultimately with the sea. This action was compared with the movement to the wick of fluid in a lamp. The idea that the Nile waters connect with the sea is an old one, tracing back to the geographic concepts of Hecataeus of Miletus (*c.* 520 BC). Hecataeus conceived the lands of the world to be circular in plan and surrounded by a great swift stream, Oceanus. Of all the rivers, the Nile alone draws its headwaters from Oceanus, whose surgings are reflected in its annual rise and ebb. Ephorus, a Greek historian of the 4th century BC, favoured a solar theory, which envisioned the Sun's "sweating out" water from pores in the Earth during the summer. Other more reasonable explanations related the discharge of the Nile to precipitation in the headwater regions, as snow (Anaxagoras of Clazomenae, *c.* 500–428 BC), or from rain that filled lakes supposed to have fed the river (Democritus of Abdera, *c.* 460–*c.* 357 BC). Eratosthenes (*c.* 276–194 BC), who had prepared a map of the Nile Valley southward to the latitude of modern Khartoum, anticipated the correct answer when he reported that heavy rains had been observed to fall in the upper reaches of the river and that these were sufficient to account for the flooding.

Knowledge of the tides. The tides of the Mediterranean, being inconspicuous in most places, attracted little notice from the Greek and Roman naturalists. Herodotus noted the tides of the Red Sea, and from the voyage of Pytheas (born *c.* 300 BC) came the report of the high spring tides in Pentland Firth between Scotland and the Orkney Islands. Also in the early 4th century, sailors with the expedition of Alexander the Great experienced the tides at the mouth of the Indus. Poseidonius (135–50 BC) correlated variations in the tides with phases of the Moon. Throughout the medieval period in Europe tides were commonly attributed to lunar influence.

Observations of the tidal bore near Hangchow

By contrast, the tides along the eastern shores of Asia generally have a considerable range and were subjects of close observation and much speculation among the Chinese. In particular, the tidal bore on the Ch'ien-t'ang Chiang (Ch'ien-t'ang River) near Hangchow attracted early attention; with its front ranging up to 12 feet in height, this bore is one of the largest in the world. As early as the 2nd century BC, the Chinese had recognized a connection between tides and tidal bores and the lunar cycle. Wang Ch'ung, writing in the 1st century AD, declared that

> the natural processes of Heaven and Earth have remained the same from the most ancient of times. . . . When the rivers fall into the sea there is nothing but rapid motion, but when the sea enters the rivers the waters begin to roar and foam, doubtless because their channels are small, shallow, and narrow . . . the rise of the wave follows the waxing and waning moon.

Tables recording the height of the tides were systematically kept in China at least as far back as the 9th century AD, some four centuries in advance of the earliest European tables.

Prospecting for groundwater. Although the origin of the water in the Earth that seeps or springs from the ground was long the subject of much fanciful speculation, the arts of finding and managing groundwater were already highly developed in the 8th century BC. The construction of long, hand-dug underground aqueducts (*qanats*) in Armenia and Persia represents one of the great hydrological achievements of the ancient world. *Qanats* derive their water from permeable deposits of gravel and sand washed down by streams from mountain ranges into drier lowlands along their borders. Seepages in these alluvial fans (*q.v.*) or patterns of vegetation suggesting proximity to sources of underground water guided the well diggers in their prospecting. Whenever water was found in sufficient quantity, a tunnel was then begun at the downslope location at which delivery of water was desired. The tunnel was then extended toward the head well until the water-bearing beds were penetrated, whereupon the water would flow by gravity to a point of discharge for use in irrigation or for household purposes. These tunnels are quite commonly ten to 16 kilometres (six to ten miles) long and range in depth up to nearly 120 metres (400 feet) below the surface. After some 3,000 years *qanats* are still a major source of water in Iran, where there are more than 270 kilometres (170 miles) of these tunnels in operation.

Construction of *qanats*

In the 1st century BC, Vitruvius (Pollio Vitruvius), a Roman architect and engineer, described methods of prospecting for groundwater in his *De architectura libri decem*. To locate places where water wells should be dug, Vitruvius recommended looking for spots where mist rises from the Earth in early morning. Concentrations of plants of kinds known elsewhere to inhabit marshy situations are also guides to groundwater at shallow depths. More significantly, Vitruvius had learned to associate different quantities and qualities of groundwater with different kinds of rocks and topographic situations. Underground supplies of water are more likely to be found in compact gravel and sand than in loose gravel or clay. Foothills and lowland belts along the borders of mountains are favourable situations for prospecting, because rain falling in the uplands percolates through openings in the rocks and moves underground toward the piedmont belts. Compared with water drawn from clay and loose gravel, water from flinty rocks and from the foothill areas is colder and more wholesome.

III. The 16th, 17th, and 18th centuries

GEOLOGICAL SCIENCES

Ore deposits and mineralogy. A common belief among alchemists of the 16th and 17th centuries held that metalliferous deposits were generated by heat emanating from the centre of the Earth but activated by the heavenly bodies. The Sun influenced the production of gold, which therefore should be most abundant in the hot tropics. In the same mysterious way the Moon generated deposits of silver, Mercury quicksilver, Venus copper, Mars iron, Jupiter tin, and Saturn lead. The origin of the subterranean fire itself was a subject of disagreement. Johann Rudolf Glauber, writing in 1652, proposed that the fire was generated by energies originating in the stars. More than a century earlier Georgius Agricola had offered a

more mundane explanation, attributing the interior heat of the Earth to the burning of coal, bitumin, or sulphur.

Agricola's *De natura fossilium*

Georgius Agricola has with much justification been called the father of mineralogy. Trained first in philology and then in medicine, he became interested in minerals while practicing in the Bohemian mining centre of Joachimsthal. Of his seven geological books, *De natura fossilium* (1546) contains his major contributions to mineralogy and, in fact, has been called the first textbook on that subject. In Agricola's time and well into the 19th century, "fossil" was a term that could be applied to any object dug from the Earth. Thus Agricola's classification of fossils provided pigeon holes for organic remains, such as ammonites, and for rocks of various kinds in addition to minerals. The classification is based on observable physical properties of natural substances, without reference to magical properties then so commonly attributed to various minerals, rocks, and fossils. Individual kinds of minerals, their associations and manners of occurrence, are described in detail, many for the first time.

With the birth of analytical chemistry toward the latter part of the 18th century, the classification of minerals on the basis of their composition at last became possible. Abraham Gottlob Werner (1750–1817) was one of those who favoured a chemical classification in preference to a "natural history" classification based on external appearances. His last of several classifications, published shortly after his death, recognized 317 different substances ordered in four classes. This was still a mixed classification: two of his four classes reflected chemical composition, but his "earthy minerals" and "combustible minerals" reflected other properties reminiscent of the older natural histories.

Paleontology and stratigraphy. During the 17th century the guiding principles of paleontology and historical geology began to emerge in the work of a few individuals. Nicolaus Steno, a Danish scientist and theologian, presented carefully reasoned arguments favouring the organic origin of what are now called fossils. Also, he elucidated three principles that made possible the reconstruction of certain kinds of geological events in a chronological order. In his *Canis carcariae dissectum caput* (1667) he concluded that large tongue-shaped objects found in the strata of Malta were the teeth of great sharks, whose remains were buried beneath the sea floor and later raised out of the water to their present situations. This excursion into paleontology led Steno to confront a broader question. How can one solid body, such as a shark's tooth, become embedded in another solid body, such as a layer of rock? He published his answers in 1669 in a hastily written paper entitled (in translation) "A Preliminary Discourse Concerning a Solid Body Enclosed by Processes of Nature Within a Solid." Steno cited evidence to show that when the hard parts of an organism are covered with sediment it is they and not the aggregates of sediment that are firm. Consolidation of the sediment into rock may come later, and, if so, the original solid fossil will become encased in solid rock. He recognized that sediments settle from fluids layer by layer to form strata that are originally continuous and nearly horizontal. In abbreviated wording, his principle of superposition of strata states that in a sequence of strata, as originally laid down, any stratum is younger than the one on which it rests and older than the one that rests upon it. The principle of initial continuity of strata states that wherever the bared sides of strata are seen, one may infer that their original extensions have been removed (as by erosion) or dislocated (by faulting). The principle of initial horizontality of strata states that where strata are now tilted at high angles one may infer that they have been rotated from their original near-horizontal positions. Applying these principles to the geology of Tuscany, where Steno worked under the patronage of the Grand Duke Ferdinand II, Steno concluded that the sedimentary rocks, geological structures, and physiography of that region record six stages of development. He illustrated this history with geological cross sections, thought to be the first ever published.

Steno's principles of superposition, continuity, and horizontality of strata

In 1667 and 1668, Robert Hooke read papers before the Royal Society in which he expressed many of the ideas contained in Steno's works. Hooke forcefully argued for the organic nature of fossils, which he affirmed are as essential to the study of the Earth's history as coins and urns are to the study of ancient civilizations. Elevation of beds containing marine fossils to mountainous heights he attributed to the work of earthquakes. Streams attacking these elevated tracts wear down the hills, fill depressions with sediment, and thus level out irregularities of the landscape. Unfortunately Hooke's addresses were not published until 1705, by which time most of his ideas had been incorporated in the writings of others.

Earth history according to Genesis, Werner, and James Hutton. Theorists of the late 17th and 18th centuries invented elaborate geological histories, most of which have proved worthless. Thomas Burnet's popular *Sacred History of the Earth* (1681) fitted a state of earthly paradise and Noah's Flood into the geological chronicle and went on to predict a chaotic episode from which the Golden Age will emerge. Georges-Louis Leclerc, comte de Buffon, author of *Époques de la nature* and the encyclopedic *Histoire naturelle*, divided the history of the Earth into seven epochs corresponding to the seven days of Genesis. Assuming that the Earth was originally molten, Buffon calculated that 74,832 years were required for cooling to its present condition. His "epochs" therefore ranged up to 35,000 years in length—too long for the theologians of the Sorbonne, who once forced Buffon to retract his views on the antiquity of the Earth. The two major theories of the 18th century, however, were the Neptunian and the Plutonian.

Neptunists and the universal ocean

The Neptunists, led by Werner and his students, maintained that the Earth was originally covered by a turbid ocean. The first sediments deposited over the irregular floor of this universal ocean formed the granite and other crystalline rocks that Werner called Primitive. Then, as the ocean began to subside, the Transitional Rocks and Stratified Rocks were laid down in succession. Later the Alluvium was washed down to cover parts of the continents. The Volcanic Rocks were the youngest; Neptunists took small account of volcanism and thought that lava was formed by the burning of coal deposits underground.

James Hutton, leader of the Plutonists, viewed the Earth as a dynamic body that functions as a heat machine. Streams wear down the continents and deposit their waste in the sea. Subterranean heat causes the outer part of the Earth to expand in places, uplifting the compacted marine sediments to form new continents. The history of the Earth is a repetition of these cycles of denudation and rejuvenation of the lands. Hutton recognized that granite is an intrusive igneous rock and not a primitive sediment as the Neptunists claimed. Intrusive sills and dikes of igneous rock provide evidence for the driving force of subterranean heat. Great angular unconformities separating sedimentary sequences, such as the one exposed at Siccar Point near his farm in Scotland, Hutton viewed as evidence for past cycles of sedimentation, uplift, and erosion. These events must have required vast amounts of time, but "time is to nature endless and as nothing." In the historical record recorded in the rocks, he found "no vestige of a beginning, no prospect of an end." Hutton's *Theory of the Earth*, published as an essay in 1788, was expanded to a two-volume work in 1795. John Playfair, a professor of natural philosophy at the University of Edinburgh, defended Hutton against the counterattacks of the Neptunists, and his *Illustrations of the Huttonian Theory* (1802) is the clearest contemporary account of Plutonist theory.

HYDROLOGIC SCIENCES

The idea that there is a circulatory system within the Earth, by which seawater is conveyed to mountain tops and there discharged, persisted until early in the 18th century. Two questions left unresolved by this theory were acknowledged even by its advocates. How is seawater forced uphill; how is the salt lost in the process?

The rise of subterranean water. René Descartes supposed that the seawater diffused through subterranean

Ideas on the origin of springs

channels into large caverns below the tops of mountains. There, warmed by the internal heat of the Earth, the water was distilled, the distillate condensing on the roofs of the caverns and pouring out to the surface as springs. The Jesuit philosopher Athanasius Kircher in his *Mundus subterraneus* (1664) suggested that the tides pump seawater through hidden channels to points of outlet at springs. The water of hot springs he assumed to have been warmed in passing through conduits near caverns filled with fire. To explain the rise of subterranean water beneath mountains, Robert Plot, first professor of chemistry at Oxford, appealed to the pressure of air, which forces water up the insides of mountains just as it forces quicksilver up the tube of the barometer. The idea of a great subterranean sea, "a mighty collection of water inclosed in the bowels of the earth," connecting with the ocean and supplying it with water together with all springs and rivers, was resurrected in 1695 in John Woodward's *Essay towards a Natural History of the Earth and Terrestrial Bodies.* Against the proposition that water springs from mountain tops, so often stated in theories of internal circulation, John Keill, an Oxford mathematician, observed that such is not the experience of those who have climbed mountains. Keill cited as expert witness Edward Lhuyd, a Welsh paleontologist and keeper of the Ashmolean Museum, who asserted that "throughout all his travels over Wales, he could observe no such thing as a running spring on the top of a mountain."

Relation of precipitation and surface runoff

Coexisting with the idea of a hydrological cycle propelled from within the Earth was the concept of an external cycle by which the subsurface and surface waters of the Earth are derived from precipitation of moisture evaporated from the sea. Leonardo da Vinci held that both types of cycles operated simultaneously. The French Huguenot Bernard Palissy, famous for his invention of a technique for making enamelled pottery, stoutly maintained, to the contrary, that rainfall is the sole source of rivers and springs. In his *Discours admirables* (1580) he described how rainwater falling on mountains enters cracks in the ground and flows down along these until, diverted by some obstruction, it flows out upon the surface as springs. Perennial streams, he reasoned, must be fed by a host of little springs. Palissy scorned the idea that seawater courses in veins to the tops of mountains. For this to be true, sea level would have to be higher than mountain tops, mountain springs would all be salty and would fluctuate in volume with the rise and fall of tides, and the conduits leading from sea to mountains would have to be watertight—an obvious impossibility.

Explanation of artesian flow

In his *Discours* Palissy suggested that water would rise above the level at which it was first encountered in a well provided the source of the groundwater came from a place higher than the bottom of the well. This is an early reference to conditions essential to the occurrence of artesian water, a popular subject among Italian hydrologists of the 17th and 18th centuries. Bernardino Ramazzini was one of those who wrote about the famous artesian wells and springs around the city of Modena in the Po Valley. To account for the force that caused these waters to gush forth, he assumed that a great natural cistern lay nearby in the "bowels of the Apennines." Water flowed from this cistern down through an underground channel beneath the Po Valley, so that wells dug into this channel would allow the water to rise and spill out their tops. How the water got into the cistern in the first place posed a problem. Ramazzini thought it might be supplied by the sea. Antonio Vallisneri, president of the University of Padua, came close to the right answer when he concluded that meltwater from mountain snowfields sinks into the ground to form subterranean streams that pass below Modena. In wells drilled to intersect these streams the water must necessarily rise because it is under great pressure. Issues raised in the discussions about the hydrologic cycle and the origin of water in springs and streams were finally settled by measurements of precipitation, discharge from drainage basins, and evaporation of seawater.

In the latter part of the 17th century, Pierre Perrault and Edmé Mariotte conducted hydrological investigations in the basin of the Seine River that established that the local annual precipitation was more than ample to account for the annual runoff. Perrault's studies covered only a small drainage basin above Ayney le Duc, Mariotte's the entire basin above the confluence of the Marne. Perrault's calculations showed that only about one-sixth of the water that falls as rain and snow during the year is accounted for by stream discharge. Mariotte's figures showed that the annual volume of water flowing in the Seine past a point near Pont Royal is less than one-sixth the volume of water annually precipitated in the drainage basin upstream. Extrapolating from his findings, Perrault deemed it likely that "the waters of rains and snows are sufficient to cause the flow of all the rivers of the world."

Halley's explanation of ocean salinity

Evaporation from the sea. The question remained as to whether the amount of water evaporated from the sea is sufficient to account for the precipitation that feeds the streams. Edmund Halley and his associates at Gresham College measured the rate of evaporation from pans of water exposed to the air during hot summer days. Assuming that this same rate would obtain for the Mediterranean, Halley calculated that some 5,280,000,000 tons of water are evaporated from this sea during a summer day. Assuming further that each of the nine major rivers flowing into the Mediterranean has a daily discharge ten times that of the Thames, he calculated that a daily inflow of fresh water back into that sea would be 1,827,000,000 tons, only slightly more than a third of the amount lost by evaporation. Halley went on to explain what happens to the remainder. A part falls back into the sea as rain before it reaches land. Another part is taken up by plants, later to be "exhaled in vapour by the action of the sun" and then carried elsewhere by the winds. "Yet," he concluded, "after several vicissitudes of rising in vapour and falling in rain or dews, each particle of the water is at length returned to the sea from whence it came."

In the course of the hydrological cycle, Halley reasoned, the rivers constantly bring salt into the sea in solution, but the salt is left behind when seawater evaporates to replenish the streams with rainwater. Thus the sea must be growing steadily saltier. Knowing the salinity of the sea and the annual input of new salt from streams, it should be possible to estimate the age of the Earth, he suggested.

ATMOSPHERIC SCIENCES

Water vapour in the atmosphere. After 1760 the analytical chemists at last demonstrated that water and air are not the same substance in different guises. Long before this development, however, investigators had begun to draw a distinction between water vapour and air. Otto von Guericke, who invented the air pump around 1650, produced artificial clouds by releasing air from one flask into another one from which the air had been evacuated. A fog then formed in the unevacuated flask. Guericke concluded that air cannot be turned into water, though moisture can enter the air and later be condensed into water.

Guericke's experiments, however, did not answer the question as to how water enters the atmosphere as vapour. Descartes, in *Les Météores* (1637), envisioned water as composed of minute particles that were elongate, smooth, and separated by a highly rarified "subtle matter." When the Sun warms water, the heat agitates this subtle matter and the water particles fly into the air just as dust is raised by the kick of a boot. These clouds of water particles raised into the air constitute water vapour, which is thus made of the same particles as water but separated by much more of the subtle matter. Other investigators associated vapour with fire. The priest Urbano d'Aviso suggested in 1666 that water vapour consists of tiny bubbles that are filled with fire and ascend through the air like balloons. The idea that fire, conceived as a material substance, is mainly responsible for evaporation persisted until near the end of the 18th century. A third view of evaporation held that water is dis-

solved in the atmosphere to produce vapour just as salt is dissolved in water. Charles Le Roy, who defended this position in 1752, went on to affirm that temperature determines how much water air can dissolve, just as it determines how much salt water can hold in solution.

Cloud formation and motion

The same uncertainty as to how water gets into the air surrounded the question as to how it remains suspended as clouds. A popular view in the 18th century was that clouds are made of countless tiny bubbles that float in air just as highly vesicular bits of pumice float in water. Guericke had suggested that the fine particles in his artificial clouds were bubbles. Other observers professed to have seen, with the aid of microscopic devices, bubble-shaped particles of water vapour rising from warm water or hot coffee.

Pressure, temperature, and atmospheric circulation. If clouds are essentially multicompartmented balloons, their motions could be explained by the movements of winds blowing upon them. Descartes suggested that the winds might blow upward as well as laterally, causing the clouds to rise or at least preventing them from descending. In 1749, Benjamin Franklin explained updrafts of air as due to local heating of the atmosphere by the Sun. Sixteen years later, the mathematical physicist Johann Heinrich Lambert described the conditions necessary for the initiation of convection currents in the atmosphere. Lambert reasoned that the warm air rising by local heating flows laterally aloft into bordering areas of cooler air, increasing the downward pressure of these and causing their lower layers to flow into the area of ascending currents, thus producing circulation.

Importance of the barometer

The fact that Lambert could appeal to changes in air pressure to explain circulation reflects an important change from the view still current in the late 16th century that air is weightless. This misconception was corrected after 1643 with the invention of the mercury barometer. Almost immediately it was discovered that the height of the barometer varied with the weather, usually standing at its highest during spells of clear weather and falling to the lowest marks on rainy days. This was exactly the opposite of the results anticipated, for common sense would seem to require that clear air is lighter than murky air.

Toward the end of the 18th century it was beginning to be understood that variations in the barometer must somehow be related to the general motion and circulation of the atmosphere. That these variations could not be due solely to changes in humidity was the conclusion of Horace Bénédict de Saussure in his *Essais sur l'hygrométrie* (1783). From experiments with changes of water vapour and pressure in air enclosed in a glass globe, Saussure calculated that if the entire atmosphere should change from a state of full saturation to one of complete dryness, a change of only a half inch in the mercury would result. He concluded that changes in temperature must be immediately responsible for variations of the barometer and that these in turn must be related to the movement of air from one place to another.

The idea that clouds lose their rain by mechanical effects of colliding with hills was an imperfect explanation of a phenomenon so common that it must have been noted by observers of all times. Orographic precipitation is the name now given to the rain and snow precipitated from moist air blowing in from the sea and forced to rise above uplands with attendant cooling. In 1693 Halley described the characteristics of orographic precipitation and went on to draw a providential moral on the purpose of the hills.

> This, if we may allow final causes, seems to be the design of the hills, that their ridges being placed through the midst of the continents, might serve . . . to distil fresh water for the use of man and beast, and their heights to give a descent to those streams to run gently, like so many veins of the macrocosm, to be more beneficial to the creation. . . .

IV. The 19th century

GEOLOGICAL SCIENCES

Crystallography and the classification of minerals and rocks. René-Just Häuy, whose treatises on mineralogy and crystallography appeared in 1801 and 1822, respectively, has been credited with advancing mineralogy to the status of a science and with establishing the science of crystallography. From his studies of the geometrical relationships between planes of cleavage, he concluded that the ultimate particles forming a given species of mineral all have the same shape and that variations in crystal habit only reflect differences in the ways identical molecules are put together. Meanwhile, Jöns Jacob Berzelius was improving methods for chemical analysis of minerals, and in 1814 he published a system of mineralogy offering a comprehensive classification of minerals based on their chemistry. Berzelius recognized silica as an acid and introduced into mineralogy the important group of minerals known as silicates. At midcentury James Dwight Dana's *System of Mineralogy* appeared in a third edition reorganized around a chemical classification, which thereafter became standard for mineralogic handbooks.

Optical studies of rocks

The development of the polarizing microscope and the technique for grinding sections of rocks so thin as to be virtually transparent came in 1827 from studies of fossilized wood by William Nicol. In 1849 Clifton Sorby showed that minerals viewed in thin section could be identified by their optical properties, and soon afterward improved classifications of rocks were made on the basis of their mineralogic composition. Ferdinand Zirkel's *Mikroscopische Beschaffenheit der Mineralien und Gesteine* (1873) contains one of the first mineralogic classifications of rocks and marks the emergence of microscopical petrography as an established branch of science.

William Smith and faunal succession. In 1683 the zoologist Martin Lister proposed to the Royal Society that a new sort of map be drawn showing the areal distribution of the different kinds of British "soiles" (vegetable soils and underlying bedrock). The work proposed by Lister was not accomplished until 132 years later, when William Smith published his *Geologic Map of England and Wales with part of Scotland* (1815). A self-educated surveyor and engineer, Smith had the habit of collecting fossils and making careful note of the strata that contained them wherever his professional travels took him. He discovered that the different stratified formations in England contain distinctive assemblages of fossils. His map, reproduced on a scale of five miles to the inch, showed 20 different rock units, to which Smith applied local names in common use—*e.g.*, London Clay and Purbeck Beds. In 1816 Smith published a companion work, *Strata Identified by Organized Fossils*, in which the organic remains characteristic of each of his rock units were illustrated. His generalization that each formation is "possessed of properties peculiar to itself [and] has the same organized fossils throughout its course" is the first clear statement of the principle of faunal sequence, which is the basis for worldwide correlation of fossiliferous strata into a coherent system. Smith thus demonstrated two kinds of order in nature: order in the spatial arrangement of rock units and order in the succession of ancient forms of life. In his own words:

> Fossils have been studied as great curiosities, collected with great pains, treasured with great care and at great expense, and shown and admired with as much pleasure as a child's rattle or a hobby horse is shown and admired by himself and his play fellows, because it is pretty; and this has been done by thousands who have not paid the least regard to that wonderful order and regularity which Nature has disposed of these singular productions, and assigned to each class its peculiar stratum.

Other paleontologists, William Buckland, for example, were interested more in what fossils can tell about ancient life than in their usefulness for correlating strata. This difference in interest divides the stratigraphical paleontologists from the paleobiologists until the present. Before the end of the 19th century, systematic studies had been made on the main groups of fossil animals and plants. The results are summarized in Karl A. von Zittel's *Handbuch der Palaeontologie*, compiled over 17 years beginning in 1876.

Smith's principle of faunal sequence was another way of saying that there are discontinuities in the sequences of fossilized plants and animals. These discontinuities

Cuvier's catastrophic viewpoint

were interpreted in two ways: as indicators of episodic destruction of life or as evidence for the incompleteness of the fossil record. Baron Georges Cuvier was one of the more distinguished members of a large group of naturalists who believed that paleontological discontinuities bore witness to sudden and widespread catastrophes. Cuvier's skill at comparative anatomy enabled him to reconstruct from fragmentary remains the skeletons of large vertebrate animals found at different levels in the Cenozoic sequence of northern France. From these studies he discovered that the fossils in all but the youngest deposits belong to species now extinct. Moreover, these extinct species have definite ranges up and down in the stratigraphic column. Cuvier inferred that the successive extinctions were the result of convulsions that caused the strata of the continents to be dislocated and folded and the seas to sweep across the continents and just as suddenly subside.

Charles Lyell and uniformitarianism. In opposition to the catastrophist school of thought, Charles Lyell proposed a uniformitarian interpretation of geologic history, first expounded in his *Principles of Geology* (3 vols., 1830–33). His system was based on two propositions: the causes of geologic change now operating include all the causes that have acted from the earliest time to which we can look back; and these causes have always operated at the same average levels of energy as at present. These two propositions add up to a "steady-state" theory of the Earth, whose history is without definite directional trends either physically or organically. There has been no progressive cooling of the Earth through time. Changes in climate have fluctuated around a mean, reflecting changes in the position of land and sea. Progress through time in the organic world is likewise an illusion, the effect of an imperfect paleontological record. If mammalian fossils seem to be missing from the older strata, this is probably due to the fact that ancient mammals were terrestrial, whereas these strata are predominantly marine. Given the vast duration of geological time, a host of small changes can produce spectacular results: a long series of quakes with upward dislocations of only a few feet each could in time produce an Andean chain, and given enough time a small stream could cut a Grand Canyon. Appealing to convulsions of nature to explain past changes is a procedure "calculated to foster indolence, and blunt the keen edge of curiosity." The main part of the *Principles*, however, was devoted less to theory than to procedures for inferring events from rocks; and for Lyell's clear exposition of methodology his work was highly regarded throughout its many editions, long after the author himself had abandoned anti-progressivist views on the development of life.

The steady-state view of Earth history

Louis Agassiz and the Ice Age. Huge boulders of granite resting upon limestone of the Jura Mountains were subjects of controversy during the 18th and early 19th centuries. Saussure described these in 1779 and called them erratics. He concluded that they had been swept to their present positions by torrents of water, even though their nearest possible sources lay in croppings of granite 100 kilometres (60 miles) away. Saussure's interpretation was in accord with the tenets of diluvial geologists, who interpreted erratics and sheets of unstratified sediment (till or drift) spread over the northerly parts of Europe and North America as the work of the Deluge.

Argument for large-scale glaciation of Europe and Asia

In 1837, the Swiss zoologist and paleontologist Louis Agassiz delivered a startling presidential address before the Helvetian Society, proposing that, during a geologically recent stage of refrigeration, glacial ice had covered Eurasia from the North Pole to the shores of the Mediterranean and Caspian. The onset of glacial climate must have been sudden, because mammoths in Siberia were frozen in their tracks. Wherever erratics, till, and striated pavements of rock occur throughout the world, these provide sure evidence of this recent catastrophe. The reception accorded this address was glacial, too, and Alexander von Humboldt advised Agassiz to return to his fossil fishes. Instead, he began intensive field studies and in 1840 published his *Études sur les glaciers*, a monumental and superbly illustrated work, clearly demonstrating that Alpine glaciers had been far more extensive in the past than now. That same year he visited the British Isles in the company of Buckland and extended the glacial doctrine to Scotland, northern England, and Ireland. In 1846 he carried his campaign to North America and there found much additional evidence for an Ice Age.

Geological time and the age of the Earth. By the middle of the century the fossiliferous strata of Europe had been grouped into systems arrayed in chronologic order. The stratigraphic column, a mental composite of these systems, was pieced together from exposures in different regions by application of the principles of superposition and faunal sequence. Time elapsed during the formation of a system became known as a period, and the periods in turn were grouped into eras: the Paleozoic (Cambrian through Permian periods), Mesozoic (Triassic, Jurassic, and Cretaceous periods), and Cenozoic (Tertiary and Quaternary periods).

Charles Darwin's *Origin of Species* (1859) offered a theoretical explanation for the empirical principle of faunal sequence. The fossils of the successive systems are different not only because parts of the stratigraphic record are missing but also because most species have lost in their struggles for survival and also because those that do survive evolve into new forms with the passage of time. On the tree of life there are many dead branches and branches fallen to the ground surround its trunk, but all life is interrelated, grown from the same stock. Darwin borrowed two ideas from Lyell and the uniformitarians: the idea that geologic time is virtually without limit and the idea that a sequence of minute changes integrated over long periods of time produce remarkable changes in natural entities.

Limitations based on the Earth's heat production

Both the evolutionists and the historical geologists were embarrassed when, beginning in 1864, William Thomson (later Lord Kelvin) attacked the steady-state theory of the Earth and placed numerical strictures on the length of geological time. The Earth might function as a heat machine, but it could not also be a perpetual motion machine. Assuming that the Earth was originally molten, Thomson calculated that not less than 20,000,000 and not more than 400,000,000 years could have passed since the Earth first became a solid body. Other physicists of note put even narrower limits on the Earth's age ranging down to 15,000,000 or 20,000,000 years. All these calculations, however, were based on the common assumption, not always explicitly stated, that the Earth's substance is inert and hence incapable of generating new heat through time. Shortly before the end of the century this assumption was negated by the discovery of radioactive elements that disintegrate spontaneously and release heat to the Earth in the process.

Concepts of landform evolution. The scientific exploration of the American West following the end of the Civil War yielded much new information on the sculpture of the landscape by streams. John Wesley Powell in his reports on the Colorado River and Uinta Mountains (1875, 1876) explained among other matters how streams may come to flow across mountain ranges rather than detour around them. The Green River does not follow some structural crack in its gorge across the Uinta Mountains; instead it has cut its canyon as the mountain range was slowly bowed up, somewhat as a rotary saw cuts a slot in the log pressed against it. "The mountains," Powell wrote, "are carved by the clouds—patient artists who take what time may be necessary for their work." Or, streams cutting down through surfaces of unconformity could become superimposed across the structural grain of the strata beneath and cut across the belts of harder rock later brought into relief by erosion. Given enough time, streams will erode their drainage basins to plains approaching sea level as a base. Grove Karl Gilbert's *Report on the Geology of the Henry Mountains* (1877) offered a detailed analysis of fluvial processes. According to Gilbert (1843–1918) all streams work toward a graded condition, a state of dynamic equilibrium that is attained when the net effect of the flowing water is neither erosion of the bed nor deposition of sediment, when the landscape reflects a balance between the resistance of the

Gilbert's statement of dynamic equilibrium

rocks to erosion and the processes that are operative upon them. After 1884 William Morris Davis developed the concept of the geographical cycle, during which elevated regions were supposed to pass through successive stages of dissection and denudation characterized as youthful, mature, and old. Youthful landscapes are said to have broad divides and narrow valleys. With further denudation the original surface on which the streams began their work is reduced to ridgetops. Finally in the stage of old age, the region is reduced to a nearly featureless plain near sea level or its inland projection. Uplift of the region in any stage of this evolution will activate a new cycle. Davis' views dominated geomorphic thought until well into the 20th century, when quantitative approaches resulted in the rediscovery of Gilbert's ideas.

Gravity, isostasy, and the Earth's figure. Discoveries of regional anomalies in the Earth's gravity led to the realization that high mountain ranges have underlying deficiencies in mass about equal to the apparent surface loads represented by the mountains themselves. In the 18th century Pierre Bouguer had observed that the deflections of the pendulum in Peru are much less than they should be if the Andes represent a load perched on top of the Earth's crust. Similar anomalies were later found to obtain along the Himalayan front. To explain these anomalies it was necessary to assume that beneath some depth within the Earth pressures are hydrostatic (equal on all sides). If excess loads are placed upon the crust, as by addition of a continental ice cap, the crust will sink to compensate for the additional mass and will rise again when the load is removed. The tendency toward general equilibrium maintained through vertical movements of the Earth's outer layers was called isostasy in 1899 by Clarence Edward Dutton.

Concept of the geosyncline and mountain building

Evidence for substantial vertical movements of the crust was supplied by studies of regional stratigraphy. In an address published in 1883, James Hall demonstrated that Paleozoic rocks of the folded Appalachians were several times as thick as sequences of the same age in the plateaus and plains to the west. He concluded that the folded strata in the mountains must have accumulated in a linear submarine trough that filled with sediment as it subsided. Downward crustal flexures of this magnitude came to be called geosynclines.

HYDROLOGIC SCIENCES

Darcy's law and groundwater flow. Quantitative studies of the movement of water in streams and aquifers led to the formulation of mathematical statements relating discharge to other factors. Henri Darcy was the first to state clearly a law describing the flow of groundwater. Darcy's experiments, reported in 1856, were based on the ideas that an aquifer is analogous to a main line connecting two reservoirs at different levels and that an artesian well is like a pipe drawing water from a main line under pressure. His investigations of flow through stratified beds of sand led him to conclude that the rate of flow is directly proportional to the energy loss and inversely proportional to the length of the path of flow. A coefficient whose value depends on the characteristics of the sand was also entered as a factor in Darcy's law. Jules-Juvénal Dupuit extended Darcy's work and developed equations for underground flow toward a well, for the recharge of aquifers, and for the discharge of artesian wells. Philip Forchheimer, an Austrian hydrologist, introduced the theory of functions of a complex variable to analyze the flow by gravity of underground water toward wells and developed equations for determining the critical distance between a river and a well beyond which water from the river will not move into the well. In the United States, Charles Sumner Slichter, unaware of Forchheimer's work, duplicated much of it. He developed an elaborate expression for the movement of water through a vertical column of soil and developed techniques for measuring the rate of flow of groundwater by injecting electrolytes into the ground at one point and detecting their arrival at another by use of an ammeter.

Surface water discharge. A complicated empirical formula for the discharge of streams resulted from the studies of Andrew Atkinson Humphreys and Henry Larcom Abbot in the course of the Mississippi Delta Survey of 1851–60. Their formula contained no term for roughness of channel and on this and other grounds was later found to be inapplicable to the rapidly flowing streams of mountainous regions. In 1869 Emile-Oscar Ganguillet and Rudolph Kutter developed a more generally applicable discharge equation following their studies of flow in Swiss mountain streams. Toward the end of the century, systematic studies of the discharge of streams had become common. In the United States the Geological Survey, following its establishment in 1879, became the principal agency for collecting and publishing data on discharge, and by 1906 stream gauging had become nationwide.

Foundations of oceanography. In 1807 Thomas Jefferson ordered the establishment of the U.S. Coast Survey (later Coast and Geodetic Survey and now the National Ocean Survey). Modelled after British and French agencies that had grown up in the 1700s, the National Oceanographic Survey was charged with the responsibilities of hydrographic and geodetic surveying, studies of tides, and preparation of charts. Beginning in 1842 the U.S. Navy began expansive oceanographic operations through its office of charts and instruments. Lieut. Matthew Fontaine Maury, officer in charge of operations, began to correlate observations of winds and currents made by ships of the Navy and promoted international cooperation in gathering meteorologic and hydrologic data at sea. In 1847 Maury compiled the first wind and current charts for the North Atlantic and in 1854 issued the first depth map to 4,000 fathoms (7,300 metres [24,000 feet]). His *Physical Geography of the Sea* (1855) is generally considered the first oceanographic textbook. Aside from Maury's scientific contributions, his charts greatly reduced the cost and time of travel across the Atlantic.

Maury's compilation of oceanographic data

The voyage of the "Beagle" (1831–36) is well remembered for the biological and geological contributions of its young naturalist, Charles Darwin. From his observations in the South Pacific, Darwin formulated a theory for the origin of coral reefs, which with minor modifications has stood the test of time. Darwin viewed the fringing reefs, barrier reefs, and atolls as successive stages in a developmental sequence. The volcanic islands around which the reef-building organisms are attached slowly sink, but at the same time the shallow-water organisms that form the reefs build their colonies upward so as to remain in the sunlit layers of water. With submergence of the island, what began as a fringing reef girdling a landmass at last becomes an atoll enclosing a lagoon. The development of these reefs thus records a contest between organisms and the blind forces of nature—an idea Darwin would use later in his theory of organic evolution.

Laying telegraphic cables across the Atlantic called for investigations of the configuration of the ocean floor, of the currents that sweep the bottom, and of the benthonic animals that might damage the cables. The explorations of the British ships "Lightning" and "Porcupine" in 1868 and 1869 turned up surprising oceanographic information. In waters west of the British Isles masses of bottom water adjacent to each other were found to have distinctly different temperatures. Living crinoids, a group of echinoderms thought to have been long extinct, were brought to the surface. Following closely upon these voyages, the "Challenger" was authorized to determine "the conditions of the Deep Sea throughout the Great Ocean Basins."

The "Challenger" expedition (1872–76)

The "Challenger" left port in December of 1872 and returned in May 1876, after logging 68,890 nautical miles. Under the direction of Wyville Thomson, Scottish professor of natural history, she occupied 350 stations scattered over all oceans except the Arctic. The work involved in analyzing the information gathered during the expedition was completed by Thomson's shipmate Sir John Murray, and the results filled 50 large volumes. Hundreds of new species of marine organisms were described, including new forms of life from deep waters.

The temperature of water at the bottom of the oceans was found to be nearly constant below the 2,000-fathom (12,000-foot) level, averaging about 2.5° C (36.5° F) in the North Atlantic and 2° C (35° F) in the North Pacific. Soundings showed wide variations in depths of water, and from the dredgings of the bottom came new types of sediment—red clay, and oozes made predominantly of the minute skeletons of foraminifera, radiolarians, or diatoms. Improved charts of the principal surface currents were produced, and the precise location of many oceanic islands was determined for the first time. Seventy-seven samples of seawater were taken at different stations from depths ranging downward to nearly a mile. Wilhelm Dittmar conducted quantitative determinations of the seven major constituents (other than the hydrogen and oxygen of the water itself); namely, sodium, calcium, magnesium, potassium, chloride, bromide, and sulfate. Surprisingly, the percentages of these components turned out to be nearly the same in all samples.

Tidal analysis and prediction

Efforts to analyze the rise and fall of the tides in mathematical terms reflecting the relative and constantly changing positions of Earth, Moon, and Sun, and thus to predict the tides at particular localities, has never been entirely successful, because of local variations in configuration of shore and sea floor. Nevertheless, harmonic tidal analysis gives essential first approximations that are essential to tidal prediction. In 1884 a mechanical analogue tidal prediction device was invented by William Ferrel of the U.S. Coast and Geodetic Survey, and improved models were used until 1965, when the work of the analogue machines was taken over by electronic computers.

ATMOSPHERIC SCIENCES

Composition of the atmosphere. Studies of barometric pressure by John Dalton led him to conclude that evaporation and condensation of vapour do not involve chemical transformations. Water vapour, he stated, "always exists as a fluid *sui generis*, diffused among the rest of the aerial fluids." If so, the introduction of vapour into the air by evaporation must change the average specific gravity of the air column and, without altering the height of that column, will change the reading of the barometer. Rudolf Clausius in 1857 clarified the mechanics of evaporation in his kinetic theory of gases. Evaporation occurs when more molecules of a liquid are leaving its surface than returning to it, and the higher the temperature the more of these escaped molecules will be in space at any one time.

Sampling the atmosphere by balloon

Following the invention of the hot-air balloon by the Montgolfier brothers in 1783, balloonists produced some useful information on the composition and movements of the atmosphere. In 1804 the celebrated French chemist Louis-Joseph Gay-Lussac ascended to 6,636 metres (21,766 feet), took samples of air, and later determined that the rarefied air at that altitude contained the same percentage of oxygen (21.49 percent) as the air on the ground. Austrian meteorologist Julius von Hann, working with data from balloon ascents and climbing in the Alps and Himalayas, concluded in 1874 that about 90 percent of all the water vapour in the atmosphere is concentrated below 6,000 metres (20,000 feet)—from which it follows that high mountains can be barriers against the transport of water vapour. In 1843 balloonist John Wise inadvertently demonstrated the violent motions of convectional updrafts in thunderstorms when he floated into such a storm near Harrisburg, Pennsylvania. He was lifted up and thrust down eight or ten times, and when he finally landed his beard bristled with hoarfrost and his equipment was covered with ice and snow.

Understanding of clouds, fog, and dew. Most of the names given to clouds (cirrus, cumulus, stratus, nimbus, and their combinations) were coined in 1803 by Luke Howard. Howard's effort was not simply taxonomic: he recognized that clouds reflect in their shapes and changing forms "the general causes which effect all the variations of the atmosphere, just as the countenance of a person reflects the state of his mind or body."

After Guericke's experiments, it was widely believed that water vapour condenses into cloud as soon as the air containing it cools to the dew point. That this is not necessarily so was proved by Paul-Jean Coulier from experiments reported in 1875. Coulier found that the sudden expansion of air in glass flasks failed to produce an artificial cloud if the air in the system was filtered through cotton wool. He concluded that dust in the air was essential to the formation of cloud in the flask. John Aitken carried the Coulier experiments further to show that the fewer the dust particles in the system the coarser grained the fog. These and subsequent investigations have shown that the dust particles, or nuclei of condensation, are mostly invisible sea salts and products of combustion. Even in unsaturated air water may form on these nuclei to produce haze.

Discovery of condensation nuclei

Beginning around the second quarter of the century, efforts were made to classify precipitation in terms of the causes behind the lowering of temperature. In 1841 Elias Loomis recognized the following causes: warm air coming into contact with cold earth or water, responsible for fog; mixing of warm and cold currents, which commonly results in light rains; and sudden transport of air into high regions, as by flow up a mountain slope or by warm currents riding over an opposing current of cold air, which may produce heavy rains. The formation of dew Loomis attributed to radiation, but in this he had been anticipated by demonstrations conducted after 1812 by William Charles Wells. Among other matters Wells had disposed of the idea that dew on plants is transpired solely by the plants themselves, pointing out that dew forms on dead vegetation as well as on live.

Observation and study of storms. Storms, particularly tropical revolving storms, were subjects of much interest.

Recognition of the revolving nature of storms

As early as 1697 some of the more spectacular features of revolving storms were recorded in William Dampier's *New Voyage Round the World.* On July 4, 1687, Dampier's ship survived the passage of what he called a "tuffoon" off the coast of China. The captain's vivid account of this experience clearly describes the calm central eye of the storm and the passage of winds from opposite directions as the storm moved past. Heinrich Wilhelm Dove in 1828 recognized that tropical revolving storms are travelling systems with strong winds moving counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere. The whirlwind character of these storms was independently established by William C. Redfield in the case of the September hurricane that struck New England in 1821. He noted that in central Connecticut the trees had been toppled toward the northwest, whereas some 80 kilometres (50 miles) westward they had fallen in the opposite direction. Redfield identified the belt between the equator and the tropics as the region in which hurricanes are generated, and he recognized how the tracks of these storms tend to veer eastward when they enter the belt of westerly winds at about latitude 30° N. In 1849, Sir William Reid studied the revolving storms that occur south of the Equator in the Indian Ocean and confirmed that they have reversed rotations and curvatures of path compared with those of the Northern Hemisphere. Captain Henry Piddington subsequently investigated revolving storms affecting the Bay of Bengal and Arabian Sea, and in 1855 he named these cyclones (from the Greek word for the coil of a snake) in his *Sailor's Horn-book for the Laws of Storms in all Parts of the World.*

Consideration of energy sources

Beginning in 1835 James Pollard Espy began extensive studies of storms from which he developed a theory to explain their sources of energy. Radially convergent winds, he believed, cause the air to rise in their area of collision. Upward movement of moist air is attended by condensation and precipitation. Latent heat released through the change of vapour to cloud or water causes further expansion and rising of the air. The higher the moist air rises the more the equilibrium of the system is disturbed, and this equilibrium cannot be restored until moist air at the surface ceases to flow toward the ascending column. Though Espy greatly overestimated the energies inherent in rising air, he called attention to one of the forces that drives storms.

That radially convergent winds are not necessary to the rising of large air masses was demonstrated by Loomis in the case of a great storm that passed across the northeastern United States in December of 1836. From his studies of wind patterns, changes of temperature, and changes in barometric pressure, he concluded that a cold northwest wind had displaced a wind blowing from the southeast by flowing under it. The southeast wind made its escape by ascending from the surface of the Earth. Loomis had clearly recognized what today would be called a frontal surface.

Weather and climate. Modern meteorology began when the daily weather map was developed as a device for analysis and forecasting, and the instrument that made this kind of map possible was the electromagnetic telegraph. In the United States the first telegraph line was strung in 1844 between Washington, D.C., and Baltimore. Concurrently with the expansion of telegraphic networks, physicist Joseph Henry arranged for telegraph companies to have meteorological instruments in exchange for current data on weather telegraphed to the Smithsonian Institution. Some 500 stations had joined this cooperative effort by 1860. The Civil War temporarily prevented further expansion, but, meanwhile, a disaster of a different order had accelerated development of synoptic meteorology in Europe. On November 14, 1854, an unexpected storm wrecked British and French warships off Balaklava, now Ukrainian S.S.R. Had word of the approaching storm been telegraphed to this port in the Black Sea, the ships might have been saved. This mischance led in 1856 to the establishment of a national storm-warning service in France. In 1863 the Paris Observatory began publishing the first weather maps in modern format.

The first national weather service in the United States began operations in 1871 as an agency of the Department of War, with responsibility assigned to the Signal Service. The initial objective was to provide storm warnings for the Gulf and Atlantic coasts and the Great Lakes. In 1877 forecasts of temperature changes and precipitation averaged 74 percent in accuracy, as compared with 79 percent for cold-wave warnings. After 1878 daily weather maps were published.

Synoptic meteorology made possible the tracking of storm systems over wide areas. In 1868 Alexander Buchan published a map showing the travels of a cyclonic depression across North America, the Atlantic, and on into northern Europe. In the judgment of Sir Napier Shaw, Buchan's study marks the entry of modern meteorology, with "the weather map as its main feature and forecasting its avowed object."

Mapping regional variations

In addition to weather maps, a variety of other kinds of maps showing regional variations in the components of weather and climate were produced. In 1817, Alexander von Humboldt published a map showing the distribution of mean annual temperatures over the greater part of the Northern Hemisphere. Humboldt was the first to use isothermal lines in mapping temperature. Dove produced maps in 1848 showing mean monthly temperatures and in 1852 introduced further refinements showing thermal anomalies (local departures from mean temperatures). Buchan drew the first maps of mean monthly and annual pressure for the entire world. Published in 1869, these maps added much to knowledge of the general circulation of the atmosphere. Maury's charts of winds and currents (1848–60) were the first to use observations of winds made at sea. In 1886 Léon Teisserenc de Bort published maps showing mean annual cloudiness over the Earth for each month and the year. The first world map of precipitation showing mean annual precipitation by isohyets was the work of Loomis in 1882. This work was further refined in 1899 by Andrew John Herbertson's maps showing precipitation for each month of the year.

Thus although the 19th century was still in the age of meteorologic and climatologic exploration, broad syntheses of old information fairly well kept pace with acquisition of the new. For example, Julius Hann's massive *Handbuch der Klimatologie*, first issued in 1883, is mainly a compendium of works published in the *Meteorologische Zeitschrift*, of which Hann was long the editor. The *Handbuch* was kept current in revised editions until 1911, and this work is still sometimes called the most skillfully written account of world climate.

V. The 20th century

GEOLOGICAL SCIENCES

Radiometric dating. In 1905, shortly after the discovery of radioactivity, Bertram Boltwood suggested that lead is one of the disintegration products of uranium, in which case the older a uranium-bearing mineral the greater should be its proportional part of lead. Analyzing specimens whose relative geological ages were known, Boltwood found that the ratio of lead to uranium did indeed increase with age. After estimating the rate of this radioactive change he calculated that the absolute ages of his specimens ranged from 410,000,000 to 2,200,000,000 years. Though his figures were too high by about 20 percent, their order of magnitude was enough to dispose of the short scale of geological time proposed by Kelvin. Later studies have shown that radioactivity is a more complex process than Boltwood realized. At present the decay series generally used for dating minerals and rocks include isotopes in the following couplets of parent and daughter elements: thorium–lead, rubidium–strontium, potassium–argon, and carbon–nitrogen.

Advent of geochronology

Radiometric dating has confirmed the chronological order of the geological systems as first established solely on stratigraphic and structural principles and has set a lower limit on the age of the Earth. The oldest sedimentary rocks thus far discovered occur in the Swaziland sequence of South Africa, from which samples dated at 3,375,000,000 (± 10,000,000) years have been taken. Igneous rocks with ages in the range of 3,000,000,000 to 3,500,000,000 years have been found in South Africa, northwestern U.S.S.R., and Minnesota. By extrapolating backward in time to a situation when there was no lead that had been produced by radiogenic processes, a figure of about 4,600,000,000 years is obtained for the minimum age of the Earth. This figure is of the same order as ages obtained for certain meteorites and lunar rocks.

Experimental study of rocks. Experimental petrology began with the work of Jacobus Henricus van't Hoff, one of the founders of physical chemistry. Between 1896 and 1908 he elucidated the complex sequence of chemical reactions attending the precipitation of salts (evaporites) from the evaporation of seawater. Van't Hoff's aim was to explain the succession of mineral salts present in Permian rocks of Germany. His success at producing from aqueous solutions artificial minerals and rocks like those found in natural salt deposits stimulated studies of minerals crystallizing from silicate melts simulating the magmas from which igneous rocks have formed. Working at the Geophysical Laboratory of the Carnegie Institution of Washington, D.C., Norman L. Bowen conducted extensive phase-equilibrium studies of silicate systems, brought together in his *Evolution of the Igneous Rocks* (1928). Experimental petrology, both at the low-temperature range explored by van't Hoff and in the high ranges of temperature investigated by Bowen, continues to provide laboratory evidence for interpreting the chemical history of rocks.

A second branch of experimental science began in 1906 when Percy W. Bridgman developed a technique for subjecting rock samples to high pressures similar to those obtaining deep in the Earth. Studies of the behaviour of rocks in the laboratory have shown that their strength increases with confining pressure but decreases with rise in temperature. Down to depths of a few kilometres the strength of rocks would be expected to increase. At greater depths the temperature effect should become dominant, and response to stress should result in flow rather than fracture of rocks. In 1959 M.K. Hubbert and William W. Rubey demonstrated that fluids in the pores of rock may reduce internal friction and permit gliding over nearly horizontal planes of the large overthrust blocks associated with folded mountains.

Seismology and the structure of the Earth. Earthquake study was institutionalized in 1880 with the formation of

Discovery of the Earth's core and of the Moho

the Seismological Society of Japan under the leadership of John Milne. Milne and his associates invented the first accurate seismographs, including the instrument later widely known as the Milne seismograph. Seismology has revealed much about the structure of the Earth's core, mantle, and crust. R.D. Oldham's studies of earthquake records in 1906 led to the discovery of the Earth's core, which lies about 2,900 kilometres (1,800 miles) below the surface. From studies of the Croatian quake of October 8, 1909, Andrija Mohorovičić discovered the discontinuity (now called the Moho) that separates the crust from the underlying mantle. Since 1950 additional information on the crust has been obtained from analysis of artificial tremors produced by explosions, including underground nuclear blasts. These studies have shown that the Moho is apparently present under all the continents at a variable depth averaging around 35 kilometres (22 miles) and that the crust above it locally thickens beneath mountain ranges with folded structure. The pattern of elastic-wave velocities suggests that the upper part of the continental crust is granitic in composition and that the upper part of the oceanic crust is made of basalt-like substance.

Lunar studies. Analysis of materials and information gathered during the Apollo missions to the Moon is still under way, but results to date indicate that the surface materials are mostly of lunar origin with relatively small amounts of meteoritic material. The rocks are mainly igneous, similar to terrestrial basalt and dolerite but higher in iron and titanium and lower in alkaline content. Organic materials appear to be missing. Some rocks appear to have been fractured and otherwise modified by impact of meteorites, which were probably responsible for finely divided materials forming the regolith or "lunar soil." Radiometric dating shows that the rocks are different ages but all very old, ranging back about 4,000,000,000 years or more. Many of the craters appear to be definitely the result of impact of extralunar objects, whereas others represent former volcanism. The sinuous steep-sided valleys called rills may represent collapsed lava tubes. The origin of the flattened depressions called maria remains unresolved, but there is no evidence that their flat floors contain the sediments of former seas; rather they seem to have been formed by floods of lava.

Precambrian life. New forms of ancient life continue to be discovered, and their description and classification continues to swell the volume of a literature already so vast as to require a remarkable specialization among paleontologists. Paleoecology, the study of relationships between the plants and animals of the past and their presumed contemporary environments, has flourished since the late 1950s. New techniques for recovering and preparing fossils have given paleontologists better materials to work with, and the scanning electron microscope has revealed structures in fossils that were never before seen. The history of life before the Cambrian Period remains sketchy, but rodlike bacteria and both unicellular and multicellular forms suggesting blue-green and green algae have been found in the Fig Tree Series of South Africa in rocks dated at 3,100,000,000 years. The Gunflint Chert of Ontario has also yielded a varied assemblage of plant microfossils believed to be near 2,000,000,000 years in age.

Study of fossil algae by electron microscopy

Continental drift, rock magnetism, and sea-floor spreading. In 1915 the German meteorologist Alfred Wegener proposed that throughout most of geological time there was only one continental mass, which he named Pangea. At some time during the Mesozoic, Pangea fragmented and the parts began to drift apart. Westward drift of the Americas opened the Atlantic Ocean, and the Indian block drifted across the Equator to join with Asia. In 1937 Alexander L. Du Toit modified Wegener's hypothesis by suggesting two primordial continents: Laurasia in the north and Gondwana in the south. Aside from the congruency of continental shelf margins across the Atlantic, proponents of continental drift have amassed impressive geological evidence to support their views. Similarities in fossil terrestrial organisms in pre-Cretaceous (older than 136,000,000 years ago) strata of Africa and South America and in pre-Jurassic rocks (older than 190,000,000 years ago) of Australia, India, Madagascar, and Africa are explained if these continents were formerly connected but difficult to account for otherwise. Fitting the Americas with the continents across the Atlantic brings together similar kinds of rocks and structures. Evidence of widespread glaciation during the Late Paleozoic is found in Antarctica, southern South America, southern Africa, India, and Australia. If these continents were formerly united around the south polar region, this glaciation becomes explicable as a unified sequence of events in time and space.

Interest in continental drift heightened during the 1950s as knowledge of the Earth's magnetic field during the geological past developed from the studies of Stanley Keith Runcorn, P.M.S. Blackett, and others. Ferromagnetic minerals such as magnetite acquire a permanent magnetization when they crystallize as components of igneous rock. The direction of their magnetization is the same as the direction of the Earth's magnetic field at the place and time of crystallization. Particles of magnetized minerals released from their parent igneous rocks by weathering may later realign themselves with the existing magnetic field at the time these particles are incorporated into sedimentary deposits. Studies of the remanent magnetism in suitable rocks of different ages from over the world indicate that the magnetic poles were in different places at different times. The polar wandering curves are different for the several continents, but in important instances these differences are reconciled on the assumption that continents now separated were formerly joined. The curves for Europe and North America, for example, are reconciled by the assumption that America has drifted about 30° westward relative to Europe since the Triassic Period (190,000,000 to 225,000,000 years ago).

Remanent magnetism and polar wandering

Increased knowledge of the configuration and composition of the ocean's floor has provided additional evidence for drift. A major result of the new bathymetry has been the delineation of the Mid-Ocean Ridge, a broad swell with a relief up to 2.4 kilometres (1.5 miles), which follows a sinuous course of nearly 60,000 kilometres (40,000 miles) along the bottoms of the major ocean basins. In the early 1960s Harry H. Hess and Robert S. Dietz proposed that new oceanic crust is being generated by igneous activity along the crest of the Mid-Ocean Ridge. Molten material rises from beneath the crests, cools, and is later pushed aside by new intrusions. Thus the ocean floor is pushed at right angles and in opposite directions away from the crests, and these moving plates carry not only the oceanic crust but the continents as well. The hypothesis of sea-floor spreading has gained strong independent support from studies of oceanic magnetic anomalies and deep-sea sediments. Seismic profiles and cores from boreholes in the sea floor reveal that the sediments covering the oceanic crust are in most places no more than a few hundred feet thick. No sedimentary rocks older than the Jurassic have yet been found on the ocean floor. These findings are incompatible with the venerable doctrine of the permanency of the continents and ocean basins. Putting all of the evidence together, it seems likely that the Americas were joined with Europe and Africa until about 200,000,000 years ago, when a rift split them apart along what is now the crest of the Mid-Atlantic Ridge. A movement of one to two centimetres a year would suffice to bring the continents to their present position.

Nature of marine sediments

HYDROLOGIC SCIENCES

Water resources and seawater chemistry. Quantitative studies of the distribution of water have revealed that an astonishingly small part of the Earth's water is contained in lakes and rivers. Ninety-seven percent of all the water is in the oceans; and, of the freshwater constituting the remainder, three-fourths is locked up in glacial ice and most of the rest is in the ground. Approximate figures are also now available for the amounts of water involved in the different stages of the hydrologic cycle. Of the 859 millimetres (33.8 inches) of annual global precipitation, 23 percent falls on the lands; but only about a third of the precipitation on the lands runs directly back to the

sea, the remainder being recycled through the atmosphere by evaporation and transpiration.

With refinements in analytical techniques, the chemistry of seawater has been shown to be more complex than previously realized. Of the naturally occurring elements, some 77 have already been detected in seawater or in the organisms that inhabit it, and it is thought to be only a matter of time until traces of the others are detected. Contrary to the idea widely circulated in the older literature of oceanography, that the relative proportions of the ocean's dissolved constituents are constant, investigations since 1962 have revealed statistically significant variations in the ratios of calcium and strontium to chlorinity. The role of organisms as influences on the composition of seawater has become better understood with advances in marine biology. It is now known that plants and animals may collect certain elements to concentrations as much as 100,000 times their normal amounts in seawater. Abnormally high concentrations of beryllium, scandium, chromium, and iodine have been found in algae; of copper and arsenic in both the soft and skeletal parts of invertebrate animals; and of zirconium and cerium in plankton.

Desalinization, tidal power, and minerals from the sea. For ages a source of food and common salt, the sea is increasingly becoming a source of water, chemicals, and energy. In 1967 Key West, Florida, became the first American city to be supplied solely by water from the sea, drawing its supplies from a plant that produces more than 2,000,000 gallons of refined water daily. Magnesia was extracted from the Mediterranean in the late 19th century; at present nearly all the magnesium metal used in the United States is mined from the sea at Freeport, Texas. Harnessing the energy of the tides was accomplished on a small scale in the tidal mills of England and Wales as early as the 12th century. Since that time many ambitious schemes for using tidal power have been devised, but the first major hydrographic project of this kind was not completed until 1967, when a dam and electrical generating equipment were installed across the Rance River in Brittany. The annual power output is about 500,000,000 kilowatt-hours. The sea floor and the strata below the continental shelves are also sources of mineral wealth, actual and potential. Concretions of manganese oxide, evidently formed in the process of subaqueous weathering of volcanic rocks, have been found in concentrations as high as 100,000 tons per square mile of sea floor. In addition to the manganese, these concretions contain copper, nickel, cobalt, zinc, and molybdenum. To date, oil and gas have been the most valuable products to be produced from beneath the sea. Already about 20 percent of the world's oil and gas comes from offshore drilling, and it is predicted that by the end of the century this figure will rise to more than 50 percent.

Economic potential of the sea floor

Ocean bathymetry. Modern bathymetric charts show that about 20 percent of the surfaces of the continents are submerged to form continental shelves. Altogether the shelves form an area about the size of Africa. Continental slopes, which slant down from the outer edges of the shelves to the abyssal plains of the sea floor, are nearly everywhere furrowed by submarine canyons. The depths to which these canyons have been cut below sea level seem to rule out the possibility that they are drowned valleys cut by ordinary streams. More likely the canyons were eroded by turbidity currents, dense mixtures of mud and water that originate as mud slides in the heads of the canyons and pour down their bottoms.

Profiling of the Pacific Basin prior to and during World War II resulted in the discovery of hundreds of isolated eminences rising 1,000 or more metres (about 3,000 feet) above the floor. Of particular interest were seamounts in the shape of truncated cones, whose flat tops rise to between a mile and a few hundred feet below the surface. Hess has interpreted the flat-topped seamounts (guyots) as volcanic mountains planed off by action of the waves before they subsided to their present depths. Subsequent drilling in guyots west of Hawaii confirmed this view; samples of rocks from the tops contained fossils of Cretaceous age representing reef-building organisms of the kind that inhabit shallow water.

Ocean circulation, currents, and waves. Early in the century, Vilhelm Bjerknes and Vagn Walfrid Ekman investigated the dynamics of ocean circulation and developed theoretical principles that have influenced subsequent studies of currents in the sea. Bjerknes showed that very small forces resulting from pressure differences caused by nonuniform density of seawater can initiate and maintain fluid motion. Ekman analyzed the influence of winds and the Earth's rotation on currents. He theorized that in a homogeneous medium the frictional effects of winds blowing across the surface would cause movement of successively lower layers of water, the deeper the currents so produced the less their velocity and the greater their deflection by the Coriolis effect (an apparent force due to the Earth's rotation that causes deflection of a moving body to the right in the Northern Hemisphere and to the left in the Southern Hemisphere), until at some critical depth an induced current would move in a direction opposite to that of the wind.

Results of many investigations suggest that the forces that drive the ocean currents originate at the interface between water and air. The direct transfer of momentum from the atmosphere to the sea is doubtless the most important driving force for currents in the upper parts of the ocean. Next in importance are differential heating, evaporation, and precipitation across the air–sea boundary, altering the density of seawater and thus initiating movement of water masses with different densities. Studies of the properties and motion of water at depth have shown that strong currents also exist in the deep sea and that distinct types of water travel far from their geographic sources. For example, the highly saline water of the Mediterranean that flows through the Strait of Gibraltar has been traced over a large part of the Atlantic, where it forms a deep-water stratum that is circulated far beyond that ocean in currents around Antarctica.

Improvements in devices for determining the motion of seawater in three dimensions have led in recent years to the discovery of new currents and to the disclosure of unexpected complexities in the circulation of the oceans generally. As recently as 1951 an immense countercurrent moving eastward across the Pacific was found below depths as shallow as 20 metres (65 feet), and in the following year an analogous equatorial undercurrent was discovered in the Atlantic. In 1957 a deep countercurrent was detected beneath the Gulf Stream with the aid of subsurface floats emitting acoustic signals. The Gulf Stream and other currents sometimes called "ocean rivers" have proved to consist of numerous disconnected filaments moving in complex patterns.

Countercurrents and the Gulf Stream

Surface waves of the ocean are also exceedingly complex, at most places and times reflecting the coexistence and interferences of several independent wave systems. During World War II, interest in forecasting wave characteristics was stimulated by the need for this critical information in the planning of amphibious operations. Harold Ulrick Sverdrup and Walter Heinrich Munk combined theory and empirical relationships in developing a method of forecasting "significant wave height"—the average height of the highest third of the waves in a wave train. Subsequently, this method was improved to permit wave forecasters to predict optimal routes for mariners. Forecasting of the most destructive of all waves, the tsunami, or "tidal waves," caused by submarine quakes and volcanic eruptions, is another recent development. Soon after 159 persons were killed in Hawaii by the tsunami of 1946, the U.S. Coast and Geodetic Survey established a seismic sea-wave warning system. Using a seismic network to locate epicentres of submarine quakes, the installation predicts the arrival of tsunami at points around the Pacific Basin often hours before the arrival of the waves.

Glacier motion and the high-latitude ice sheets. Beginning around 1948, principles and techniques developed in metallurgy and solid-state physics were brought to bear on the mechanics of glacial movements. Laboratory studies showed that glacial ice deforms like other crystal-

line solids (such as metals) at temperatures near the melting point. Continued stress produces permanent deformation. In addition to plastic deformation within a moving glacier, the glacier itself may slide over its bed by mechanisms involving pressure melting and refreezing and accelerated plastic flow around obstacles. The causes underlying changes in rate of glacial movement, in particular those spectacular accelerations called surges, will require further study. Surges involve massive transfer of ice from the upper to the lower parts of glaciers at rates as much as 20 metres a day, in comparison with normal advances of a few metres a year.

Analysis of cores from the Greenland and Antarctic ice sheets

As a result of numerous scientific expeditions into Greenland and Antarctica, the dimensions of the remaining great ice sheets are fairly well-known from gravimetric and seismic surveys. In parts of both continents it has been determined that the base of the ice is below sea level, probably due at least in part to subsidence of the crust under the weight of the caps. In 1966 a borehole was drilled 1,390 metres (4,560 feet) to bedrock on the North Greenland ice sheet, and two years later a similar boring of 2,162 metres (7,091 feet) was cut through the Antarctic ice at Byrd Station. From the study of annual incremental layers and analyses of oxygen isotopes, the bottom layers of ice cored in Greenland were estimated to be more than 150,000 years old, compared with 100,000 years for the Antarctic core. With the advent of geochemical dating of rocks it has become evident that the Ice Age, which in the earlier part of the century was considered to have transpired during the Quaternary Period, actually began much earlier. In Antarctica, for example, potassium–argon age determinations of lava overlying glaciated surfaces and sedimentary deposits of glacial origin show that glaciers existed on this continent at least 10,000,000 years ago. The causes of continental glaciation, whether in recent or ancient times, are still not understood.

ATMOSPHERIC SCIENCES

Probes, satellites, and data transmission. Kites equipped with meteorgraphs were used as atmospheric probes in the late 1890s, and in 1907 the U.S. Weather Bureau recorded the ascent of a kite to 7,044 metres (23,111 feet) above Mt. Weather, Virginia. Kites were later replaced by balloons, to which recording instruments with parachutes were attached. Experimentation with balloons up to altitudes of about 31 kilometres (19 miles) showed that columns of warm air may rise more than a mile above the surface of the Earth and that the lower atmosphere is often stratified, with winds in the different layers blowing in different directions. In the 1930s airplanes began to be used for observations of the weather, and the years since 1945 have seen the development of rockets and weather satellites. TIROS (Television Infra-Red Observation Satellite), the world's first all-weather satellite, was launched in 1960, and in 1964 the Nimbus Satellite of the United States National Aeronautics and Space Administration (NASA) was rocketed into near-polar orbit with cameras continuously trained on the Earth.

In the 1920s the radio replaced the telegraph and telephone as the principal instrument for transmitting weather data. By 1936 the radio meteorgraph (radiosonde) was developed, with capabilities of sending signals on relative humidity, temperature, and barometric pressure from unmanned balloons. Radar and television are later developments that give visual images of cloud formations. The television equipment on Nimbus sends down more than 1,000 pictures a day in a spectacular round-the-clock surveillance of changing cloud patterns over the world.

Bjerknes' polar front theory of cyclones

Weather forecasting. Efforts at incorporating numerical data on weather into mathematical formulas that could then be used for forecasting were initiated early in the century at the Norwegian Geophysical Institute. Vilhelm Bjerknes and his associates at Bergen succeeded in devising equations relating the measurable components of weather, but their complexity precluded the rapid solutions needed for forecasting. Out of their efforts, however, came the polar front theory for the origin of cyclones and the now-familiar names of cold front, warm front, and stationary front for the leading edges of air masses. The polar front theory envisioned cyclones as spinning off the undulatory boundary separating predominantly west-moving cold polar air from prevailingly east-moving masses of warmer moist air in middle latitudes. In 1935 the U.S. Weather Bureau organized a section for air-mass analysis, and the following year the familiar lines for fronts began to appear on its weather maps.

In 1922 the British mathematician Lewis Fry Richardson demonstrated that the complex equations of the Norwegian school could be reduced to long series of simple arithmetic operations. With no more than the desk calculators and slide rules then available to the arithmeticians, however, the solution of a problem in procedure only raised a new one in manpower. In 1946 John von Neumann and his fellow workers at The Institute for Advanced Study, Princeton, New Jersey, began work on an electronic device to do the computation faster than the weather developed. Four years later the von Neumann group could claim that, given adequate data, their computer could actually forecast the weather as well as a weatherman. To say this was not to claim perfection. In his *Book of Weather Signs*, written around 300 BC, Theophrastus noted some 200 portents of the weather, and a few of these, such as red sunrises as portents of rain, are accurate about 70 percent of the time. In 1887 official forecasts in the United States were about 74 percent accurate. Later this figure rose to about 80 percent for forecasts 12 to 18 hours in advance, and the application of computer technology has added about 5 percent to this figure.

Cloud physics. Studies of cloud physics have shown that the nuclei around which water condenses vary widely in their degree of concentration and areal distribution, ranging from a hundred or so per cubic inch over the oceans to more than 65,000,000 in the polluted air of some cities. The droplets that condense on these foreign particles may be as small as 0.001 centimetre (0.0004 inch) in diameter. Raindrops apparently may form directly from the coalescence of these droplets, as in the case of tropical rains, or in the temperate zones through the intermediary of ice crystals. According to the theory of Tor Bergson and Walter Findeisen, vapour freezing on ice crystals in the clouds enlarges the crystals until they fall. What finally hits the ground depends on the temperature of air below the cloud—if below freezing, snow; if above, rain.

The Van Allen radiation belts

Properties and structure of the atmosphere. Less than a year after the Space Age began with the launching of the U.S.S.R. Sputnik I in 1957, the U.S. satellite Explorer I was sent into orbit with a Geiger counter for measuring the intensity of cosmic radiation at different levels above the ground. At altitudes around 1,000 kilometres (600 miles) this instrument ceased to function due to saturation by charged particles. This and subsequent investigations showed that a zone of radiation encircles the world between about latitude 75° N and 75° S, with maximum intensities at 5,000 and 16,000 kilometres (3,000 and 10,000 miles). Named after the American physicist James Van Allen, a leading investigator of this portion of the Earth's magnetosphere, these zones are responsive to events taking place on the Sun. The solar wind, a stream of atomic particles emanating from the Sun in all directions, seems to be responsible for the electrons entrapped in the Van Allen region as well as for the teardrop shape of the magnetosphere as a whole, with its tail pointing always away from the Sun. Following solar flares, the Van Allen region may become overloaded with electrons, which then cascade downward into the atmosphere, producing auroras.

In 1898 Teisserenc de Bort, studying variations of temperature at high altitudes with the aid of balloons, discovered that at elevations around seven miles the figure for average decrease of temperature with height (about 5.5° C per 1,000 metres [3° F per 1,000 feet] of ascent) dropped and the value remained nearly constant at

around −55° C (−67° F). De Bort named the atmospheric zones below and above this temperature boundary the troposphere and stratosphere. Subsequent explorations by sounding rockets and satellites have added much information about the physical and chemical characteristics of the atmosphere. Although the chemical composition of the troposphere has been found to be nearly constant, a layer of ozone (triatomic oxygen, O_3) has been found concentrated in the stratosphere at an altitude of about 26 kilometres (16 miles). This layer shields the Earth from much of the ultraviolet radiation from space that otherwise would be harmful to life. The ionosphere, which is the region above the stratosphere between about 55 and 600 kilometres (35 and 400 miles), contains the electrically conducting layers capable of reflecting radio waves. Like the two spheres of air below it, the ionosphere consists mainly of molecular nitrogen, but it also contains significant percentages of atomic nitrogen, oxygen, and hydrogen. The exosphere, uppermost envelope of the atmosphere, extending outward from about 600 kilometres (400 miles), consists mainly of hydrogen above 2,400 kilometres (1,500 miles); its boundary with the interplanetary medium is indefinite.

Discovery of the jet streams

Toward the end of World War II the B-29 Superfortress came into use as the first large aircraft to cruise regularly at 10,000 metres (30,000 feet). Heading westward from bases in the Pacific, these planes sometimes encountered unexpected head winds that slowed their flight by as much as 300 kilometres (200 miles) per hour. The jet streams, as these high-altitude winds were named, have been found to encircle the Earth following wavy courses and moving from west to east at velocities ranging upward to 500 kilometres (300 miles) per hour. Aircraft have also proved useful in studies of the structure and dynamics of tropical hurricanes. Following the destruction wrought to the Atlantic Coast of the United States in 1955 by hurricanes Connie and Diane, a national centre was established in Florida with the missions of locating and tracking and hopefully of learning how to predict the paths of hurricanes and to dissipate their energy.

Weather modification. The modification of weather has long been a dream of those who suffer from droughts or endure the ravages of storms. Many of the early schemes for making rain had no more basis in science than the rainmaking ceremonies of aboriginal peoples. As late as the 1890s experiments were conducted in the United States to the end of producing rain by setting off charges of dynamite lofted by balloons or kites. Aside from the total destruction of the balloons and kites no positive results were reported. More promising were the cloud-seeding experiments of the 1940s, in which silver iodide was released into clouds as smoke or solid carbon dioxide broadcast into clouds from airplanes. The results are still ambiguous, though it seems that clouds about to give up their moisture from natural causes may do so more readily after seeding. The inadvertent weather modification that has followed industrialization and the building of large cities has, however, already produced measurable changes in local climate and may someday produce effects more widespread. The introduction of some 12,000,000,000 tons of carbon dioxide into the atmosphere each year from the burning of fuels may in time raise the Earth's average temperature. Cities affect the flow of wind, warm the atmosphere over them, and send pollutants into the sky. Updrafts and an abundance of condensation nuclei may increase rainfall and winter fog and reduce sunshine and daylight.

Classification of climate. Renewed interest in the classification of world climates was manifest first in the work of Albrecht Penck (1858–1945), who in 1910 proposed a system of classification based on the rate of precipitation. His humid climates were those in which precipitation exceeds evaporation, the opposite of arid climates. Nival climates are characterized by more snowfall than can be lost through ablation—*i.e.*, the combined processes by which wastage of ice or snow occurs. In 1918 Wladimir Köppen (1846–1940) incorporated some of Penck's ideas in a revision of his own earlier classification. Four of his five major classes were based on seasonal variations in temperature (winterless climates, summerless climates, climates with only mild winters, and climates with severe winters but distinct summers); the fifth, dry climates, was based on precipitation. Classifications developed by Charles W. Thornthwaite, beginning in 1931, re-emphasize the importance of precipitation and relegate temperatures to secondary rank. In differentiating climatic types, he used ratios between precipitation and temperature and between precipitation and evaporation as measures of the "efficiency of temperature" and the "effectiveness of precipitation." The difficulty inherent in classification of climates is essentially the same as that impeding the progress toward a satisfactory classification of soils: too many independent climatic variables in addition to the dependent variables of plant physiology, rock weathering, and economic activity.

Climate systems of Köppen and Thornthwaite

EMERGING TRENDS

As the Earth sciences have grown in manpower and other resources, they have exhibited trends toward increasing specialization on the part of their practitioners, more team research, more general employment of quantitative approaches to problems, and increasing involvement in social issues.

With growth in numbers of research workers, the sheer bulk of their writings has itself become an obstacle to the communication of findings. A recent survey of the literature of the geological sciences alone indicates that upwards of 30,000 titles are published annually, not counting any in Oriental languages. Bibliographic and indexing services are no longer able to keep pace with the presses.

Increasing specialization. The high degree of specialization in the fields of modern science has often been diagnosed as a cause of intellectual tunnel vision. Generalists are few in a population characterized by such experts as the boundary-layer meteorologist, the biochemical oceanographer, and the paleomagneticist. But this diversity of specialists only reflects the complexity of their subjects. For example, the history of ancient man has recently been found to span more than 2,000,000 years, and, to help elucidate the story of human development, the archaeologist must enlist the aid of various specialists in the Earth sciences. At important archaeological sites assistance is often required of the stratigrapher, sedimentologist, vertebrate paleontologist, invertebrate paleontologist, paleobotanist, and soil scientist. If materials suitable for geochemical dating are present, the geochronologist will also make his contribution. The overall aims in such an investigation are to reconstruct as completely as possible the physical and biological environment in which ancient humans lived and to place their occupation of the site in a framework of time approximated in years.

Interdisciplinary problems

Team research. Team research has become commonplace in projects crossing the major disciplinary boundaries. It is now accepted that understanding of the motions of the oceans and atmosphere will require integrated studies by hydrologic and atmospheric scientists; and the tropical belts have been selected as the logical places for such joint investigations to begin. The warm oceans are major sources of the energy that powers the global circulation of water and air, but much remains to be learned about the mechanisms by which solar energy absorbed in the ocean is released to do its work. The Barbados Oceanographic and Meteorological Experiment (BOMEX), conducted in the summer of 1969 by scientists from Canada and the United States, marked the first large-scale study of energy transfer in the tropics. The region studied was square in cross section, 500 kilometres (300 miles) to the side, extending downward into the ocean for 500 metres (1,600 feet) and upward into the atmosphere for five kilometres. Preliminary analyses of the vast amount of numerical data collected show that the rate of evaporation (over six millimetres per day) was higher than expected. Moreover, the time of greatest evaporation was not around noon, when the input of heat into the ocean was greatest, but near sunset. The rate of water vapour transport, accomplished through large ed-

dies elongated downwind, differed significantly from the rate of heat transport, which was controlled by smaller eddies. An even more ambitious study to be sponsored by nine countries under the Global Atmospheric Research Program is scheduled for 1974. The belt to be investigated extends on either side of the Equator from Central America to the eastern edge of Africa. The objective is a better understanding of atmospheric processes as a basis for constructing more accurate numerical models of the atmosphere.

Use of electronic computers

Quantitative approach to problems. The conduct of experiments with numerical models of natural phenomena has been made possible by electronic computers and is but one example of the trend toward quantification in the Earth sciences generally. Numerical modelling of the weather has always been intimately associated with computers. Current operational models compute wind velocities and temperatures at six levels within a horizontal area of 300 kilometres (200 miles), but all the factors needed to forecast local weather situations cannot yet be included directly. Research models even more elaborate tax the capabilities of the largest and fastest computers. A better understanding of meteorological processes, more detailed observational information on weather, and still faster computers will be needed before accurate five-day forecasts become possible.

BOMEX is on the small end of the "big science" that requires not only teams of specialists but large investments in equipment and services. To conduct the field study off Barbados, it was necessary to employ some 28 aircraft, 12 ships, satellites and buoys, hundreds of radiosonde balloons, and floating-laboratory instrument platforms (which in this age of acronyms are known as "flips"). At the other end of the spectrum in big science is the ongoing program of lunar and planetary exploration. All mammoth projects in Earth and planetary science require public funding, often from more than one nation.

Involvement in social goals. The atomic bombing of Hiroshima in 1945 first made the global public aware of science. Since that event there has been increasing support of scientific research and development, but a growing sentiment requires that this support should be directed toward constructive ends. Ecology, environment, relevance, and pollution are watchwords symbolizing the new drift in the Earth sciences toward humanitarian goals. Some observers fear that political insistence on instant application of new findings to betterment of the human condition will dry up support for basic research.

Underground waste disposal and earthquake attenuation

This need not be the case. Denver, Colorado, was shaken by unaccustomed earthquakes in 1966. These coincided in time with efforts of the U.S. Army to dispose of nerve gas products by pumping them in fluid form and under high pressure into deep wells nearby. Seismologists suspected a causal connection. Experiments later conducted in fluid injection of oil wells in California confirmed that the frequency of quakes there seemed to vary with the hydrostatic pressure at depth. If so, the possibility is opened for postponing quakes by pumping fluids from the fault zones along which they occur and of inducing more frequent but less destructive quakes by pumping fluids into these same zones. Earthquake prediction and control would surely promote the general welfare; and the research needed to provide the techniques and put them to practice would just as surely strengthen the theoretical bases of seismology and structural geology.

BIBLIOGRAPHY. F.D. ADAMS, *The Birth and Development of the Geological Sciences* (1938), the best general account for the years prior to 1830; ARISTOTLE, *Meteorologica*, vol. 3 in *The Works of Aristotle Translated into English*, ed. by W.D. ROSS (1931), classical theories of astronomy and Earth science; A.K. BISWAS, *History of Hydrology* (1970), a factual chronicle of developments since the earliest times; R.J. CHORLEY, A.J. DUNN, and R.P. BECKINSALE, *The History of the Study of Landforms*, vol. 1, *Geomorphology Before Davis* (1964), an expansive account covering developments to around the end of the 19th century; CHARLES DAVISON, *The Founders of Seismology* (1927), a general history of seismology beginning with the mid-18th century; W.A. DEER, "Trends in Petrology," *Mem. Proc. Manch. Lit. Phil. Soc.*, 94:63–92 (1952–53), a description of the interaction of Vulcanist and Plutonist viewpoints in the development of petrology since the late 18th century; C.C. GILLISPIE, *Genesis and Geology: A Study in the Relations of Scientific Thought, Natural Theology, and Social Opinion in Great Britain, 1790–1850* (1951), analyses of the impact of developments in geology upon Christian beliefs in the decades before Darwin (extensive bibliography); A.R. and M.B. HALL (eds.), *The Correspondence of Henry Oldenburg*, 8 vol. (1965–71), displays diversity of views on Earth science current in Europe around the third quarter of the 17th century; C.P. IDYLL (ed.), *Exploring the Ocean World: A History of Oceanography* (1969), a symposium treating each of the several branches of oceanography in historical format; JOHN LIEGHLY, "Climatology Since the Year 1800," *Trans. Am. Geophys. Un.* 30:658–672 (1949), a review article with lengthy bibliography; EDWARD MACCURDY, *The Notebooks of Leonardo da Vinci*, vol. 1, pp. 333–359 (1938), contains passages that display Leonardo's remarkable insights into Earth history; G.P. MERRILL, *The First One Hundred Years of American Geology* (1924, reprinted 1969), covers in detail the period 1785–1879; W.E.K. MIDDLETON, *A History of the Theories of Rain and Other Forms of Precipitation* (1966), an explanation of advances in meteorology in context with developments in other branches of science beginning with the 17th century; JOSEPH NEEDHAM, *Science and Civilization in China*, vol. 3, *Mathematics and the Sciences of the Heavens and the Earth* (1959), contains a comprehensive and elaborately illustrated account of the history of Earth science in China to around AD 1500; RUSHDI SAID, "Geology in Tenth Century Arabic Literature," *Am. J. Sci.*, 248:63–66 (1950), a summary of geological views held by the Brothers of Purity; GEORGE SARTON, *A History of Science*, vol. 1, *Ancient Science Through the Golden Age of Greece* (1952), vol. 2, *Hellenistic Science and Culture in the Last Three Centuries B.C.* (1959, both vol. reprinted 1970), a general account of ancient science and technology in the West, amply documented; *The Appreciation of Ancient and Medieval Science During the Renaissance, 1450–1600* (1955), contrasts the imitative and retrospective with the creative trends in Renaissance science; GUSTAV SCHERZ (ed.), *Steno, Geological Papers*, trans. by A.J. POLLOCK (1969), translations of all known geological writings of Nicolaus Steno, with critical notes and commentary; CECIL SCHNEER, "The Rise of Historical Geology in the Seventeenth Century," *Isis*, 45:256–268 (1954), an analysis of the points at issue in the fossil controversy; and (ed.), *Toward a History of Geology* (1969), 25 essays on the history of geological thought, mainly of the 18th and 19th centuries; NAPIER SHAW, *Manual of Meteorology*, vol. 1, *Meteorology in History* (1942), a rambling but literate and entertaining history of meteorology from the earliest to modern times; EVELYN STOKES, "Fifteenth Century Earth Science," *Earth Sci. J.*, 1:130–148 (1967), an analysis of classical and medieval views of nature, especially those reflected in Caxton's *Mirrour of the World;* THEOPHRASTUS, *On Stones*, Greek text with translation and commentary by E.R. CALEY and J.F.C. RICHARDS (1956), a scholarly explication of the oldest known treatise on minerals and rocks; P.D. THOMPSON *et al.*, *Weather* (1965), an introduction to meteorology with much historical material, well-illustrated; STEPHEN TOULMIN and JUNE GOODFIELD, *The Discovery of Time* (1965), traces the history of the idea of geological time; H.F. TOZER, *A History of Ancient Geography*, 2nd ed. (1935), the discovery of the world related to exploration and the construction of maps and charts; W.D. WESTERVELT (comp. and trans.), *Hawaiian Legends of Volcanoes* (1964), Pele is the principal character in mythic accounts of volcanism; WILLIAM WHEWELL, *History of the Inductive Sciences from the Earliest to the Present Time*, 3rd ed., 2 vol. (1872; reprinted in 3 vol., 1967), vol. 2 contains an analysis of uniformitarian and catastrophist views of Earth history; K.A. VON ZITTEL, *Geschichte der Geologie und Paläontologie bis Ende des 19. Jahrhunderts* (1899; Eng. trans., *History of Geology and Palaeontology to the End of the Nineteenth Century*, 1901), best for its history of paleontology.

(C.C.A.)

East Africa, History of

In the following historical account, East Africa may be taken to comprise the geographical area covered by the modern states of Kenya, the Somali Democratic Republic, Tanzania, and Uganda.

This article is divided into the following sections:

I. The coast until 1856
  Azania
  The Shirazi migration
  The Portuguese invasion
  The Omani ascendancy
II. The interior before the colonial era
  The Stone Age

The spread of ironworking and the Bantu migrations
Migrations of Nilotic and Cushitic peoples
Developments in the 19th century
Political activities
Arabs and East African trade
III. The colonial era
The arrival of the Europeans
Colonial rule
The new order
"White man's country"
The colonial economies
Opening a breach: Somalia
The crises of the 1950s
The approach of independence
IV. Independence and after
The beginnings
The Somali Republic
Economic cooperation
The Zanzibar revolution and the Tanganyika mutiny
Ujamaa in Tanzania
Political unrest in Uganda
Kenyatta's Kenya

I. The coast until 1856

Evidence of the existence of microlithic industries—very small tools typical of the Late Stone Age (*c.* 500 BC)—have been found at various points along the East African coast. Such industries appear to have been established by small hunting communities similar to those that survive to this day in other parts of eastern and southern Africa. No coastal settlement that predates the Islāmic period (from *c.* 700) has as yet been found, and, if it were not for some historical evidence to the contrary, it would be tempting to link the advent of iron with that of the Muslims.

The earliest written accounts of the East African coast occur in the *Periplus Maris Erythraei*—apparently written by a Greek merchant living in Egypt in the second half of the 1st century AD—and in Ptolemy's *Guide to Geography*, the East African section of which, in its extant form, probably represents a compilation of geographical knowledge available at Byzantium *c.* 400. The *Periplus* describes in some detail the shore of what was to become Somaliland. It records the existence of six or so places where some kind of commerce was being organized and notes that regular customs dues were collected from both sides of what was later to be called Cape Guardafui. Ships, it goes on, sailed from the coast to western India to bring back cotton cloth, grain, oil, sugar, and ghee, while others moved down the Red Sea to the East African coast bringing cloaks, tunics, copper, and tin. Aromatic gums, tortoiseshell, ivory, and slaves were traded in return.

AZANIA

But, because of its offshore islands, better landing places, and wetter climate, it was the part of the East African coast farther to the south that Arab traders seem to have preferred. They sailed there with the northeast monsoon, returning home in the summer with the southwest. They dubbed the part of the coast to which they sailed Azania, or the Land of Zanj—by which they meant the land of the blacks and by which they knew it till the 10th century. South of Sarapion, Nikon, the Pyralaae Islands, and the island of Diorux (about whose precise locations only speculation seems possible), the chief town was Rhapta, which may lie buried in the Rufiji Delta. Here the situation differed somewhat from that in the north, and, though tortoiseshell and rhinoceros horn were exported from there—as were quantities of ivory and coconut oil—no mention is made of slaves. Rhapta's main imports were metal weapons and iron tools—suggesting that iron smelting was not yet known. Mafia island, which lies out to sea here, could perhaps be Menouthias, the only island named in both the *Periplus* and by Ptolemy, although this could also be either Pemba or Zanzibar (perhaps there has been a conflation of all three in the one name).

Azania, it was said, was always ultimately subject to the most important power in Arabia—Charibael (or Qaribil) of Ḥimyar, in southwestern Arabia, is specifically mentioned—but was more usually directly subordinate to the people of Mocha in the Yemen. There was some mutual linguistic understanding and some intermarriage, but there was nothing to suggest that the local inhabitants were as yet Bantu.

There is little firm information concerning the period till the 8th century. Greek and Roman coins have been found, and there are some accounts of overseas migrations to the coast. No settlements from this period have been found.

Arab geographical designations in the 9th century

A new period opened, it seems, in the 9th century. The first identifiable building sites are dated from this time, and, according to Arab geographers, the East African coast was then generally thought of as being divided into four: 1) Berber, which ran down to the Webi Shebeli; 2) Zanj proper; 3) the land of Sofala, whence gold was beginning to be shipped by about the 10th century; and 4) a vaguely described land of Waq waq, beyond. The only island that is mentioned is Qanbalu, which appears to have been what is now called Pemba. Speakers of Bantu language now seem to have reached the coast. They had rulers, priests, warrior bands, and cattle, which they harnessed like horses. Muslims, too, are reported at Merka and Manda, on Qanbalu, and later at Zanzibar. Though there is some suggestion that in the 10th century the Muslims had not yet begun to move farther south and that thus they were not to be found at Brava, Malindi, or Mombasa, on Qanbalu they soon became rulers of a pagan population, whose language they adopted. Moreover, at Zanzibar an extant Kūfic inscription (the only one) recording the construction of a mosque by Shaykh as-Sayyid Abū'Imrān Mūsā ibn al-Ḥasan ibn Muḥammad in 1107 confirms that by this time substantial Muslim settlements had been established.

The main coastal settlements were sited on islands, largely, no doubt, because of the greater security these provided against attacks from the mainland; and their populations seem mostly to have been made up from migrants from the Persian Gulf—some from the great port of Sīrāf, others from near Bahrain—though conceivably some too came from Daybul, at the mouth of the Indus River, in northwestern India. They exported ivory (some of it went as far as China) and also tortoiseshell, ambergris, and leopard skins. Such trade goods as they obtained from the interior were apparently bought by barter at the coast.

Some of the ruins at Kilwa, on the southern Tanzanian coast, probably date from the 9th or perhaps from the 8th century. They have revealed an extensive pre-Muslim settlement standing on the edge of what was the finest harbour on the coast. Though there is little evidence to suggest that its inhabitants had any buildings to begin with, wattle-and-daub dwellings appeared in due course, and by the 10th century short lengths of coral masonry wall were being built. The inhabitants, whose main local currency was cowrie shells, traded with the peoples of the Persian Gulf and, by the early 11th century, had first come under Muslim influence. By 1300, like the inhabitants of neighbouring Mafia, they were living in Muslim towns, the rulers of which were Shī'ites. Although no houses were being built of coral, stone mosques were being constructed.

External trade was increasing: glass beads were being imported from India and porcelains, transshipped either in India or in the Persian Gulf, were arriving from China. There appears also to have been a rather extensive trade with the island of Madagascar.

The most important site of this period yet to have been found is at Manda, near Lamu, on the Kenyan coast. Apparently established in the 9th century, it is distinguished for its seawalls of coral blocks, each of which weighs up to a ton. Though the majority of its houses were of wattle and daub, there were also some of stone. Trade, which seems to have been by barter, was considerable, with the main export probably of ivory. Manda had close trading connections with the Persian Gulf—with Sīrāf in particular. It imported large quantities of Islāmic pottery and, in the 9th and 10th centuries, Chinese porcelain. There is evidence of a considerable

iron-smelting industry at Manda and of a lesser one at Kilwa.

THE SHIRAZI MIGRATION

For much of the 13th century the most important coastal town was Mogadishu (modern Mogadiscio), a mercantile city to which new migrants came from the Persian Gulf and southern Arabia. Of these, the most important were called Shirazi, who, in the second half of the 12th century, had migrated southward to the Lamu islands, to Pemba, to Mafia, to the Comoro Islands, and to Kilwa, where by the end of the 12th century they had established a dynasty. Whether they were actually Persian in origin is somewhat doubtful. Though much troubled by wars, by the latter part of the 13th century they had made Kilwa second in importance only to Mogadishu. When the Kilwa throne was seized by Abū al-Mawāhib, major new developments ensued: Kilwa captured Mogadishu's erstwhile monopoly of the gold trade with Sofala and exchanged cloth—much of it made at Kilwa—and glass beads for gold; and with the great wealth that resulted new pottery styles were developed, a marked increase in the import of Chinese porcelain occurred, and stone houses, which had hitherto been rare, became common. (The great palace of Husuni Kubwa, with well over 100 rooms, was built at this time and had the distinction of being the largest single building in all sub-Saharan Africa. Husuni Ndogo, with its massive enclosure walls, was probably built at this time, too, as were the extensions to the great mosque at Kilwa. The architectural inspiration of these buildings was Arab; their craftsmanship was of a high standard; and the grammar of their inscriptions was impeccable.) Kilwa declined in the late 14th century and revived in the first half of the 15th but then—partly because of internal dynastic conflict but also partly because of diminishing profits from the gold trade—declined again thereafter.

The Shirazi dynasty

Elsewhere, especially on the Kenyan coastline, the first half of the 15th century seems to have been a period of much prosperity. Whether at Gedi (south of Malindi) or at Songo Mnara (south of Kilwa), architectural styles were relatively uniform. Single-story stone houses, mostly of coral, were common. Each coastal settlement had a stone mosque, which, typically, centred upon a roofed rectangular hall divided by masonry pillars. Chinese imports arrived in ever larger quantities, and there are signs that eating bowls were beginning to come into more common use.

When Ibn Baṭṭūṭah, the Arab traveller, visited the coast in the early 14th century, he described Mogadishu as a large town with many rich merchants and a highly developed court life revolving around a ruler. By the 15th century, Mombasa was a very substantial town, as was Pate. The ruling classes of these towns were Muslims of mixed Arab and African descent who were mostly involved in trade; beneath them were African labourers who were often slaves and a transient Arab population.

The impetus in this society was Islāmic rather than African. It was bound by sea to the distant Islāmic world, whence immigrants still arrived to settle on the East African coast, to intermarry with local people, and to adopt the Swahili language. The impact of these settlements was limited, while their influence upon the East African interior was nonexistent.

During the 15th century, Shirazi families continued to rule in Malindi, Mombasa, and Kilwa and at many lesser places along the coast. They also dominated Zanzibar and Pemba. The Nabahani, who were of Omani origin, ruled at Pate and were well represented in Pemba as well. Coastal society derived a certain unity by its participation in a single trading network, by a common adherence to Islām, and by the ties of blood and marriage among its leading families. Politically, however, its city-states were largely independent, acknowledging no foreign control, and their limited resources confined their political activities to East Africa and to a variety of local rivalries—Zanzibar and Pemba, for example, appear frequently to have been divided between several local rulers. Mombasa occupied the premier position on this part of the coast, although its control over the area immediately to the north was disputed by its main rival, Malindi. Close connections seem to have existed between Mombasa and a number of places to the south. Its Shirazi rulers were able to mobilize military support from some of the inland peoples, and as a result of the place it had won in the trade of the northwestern Indian Ocean they had turned Mombasa into a prosperous town. Its population of about 10,000 compared with only 4,000 at Kilwa.

The rise of Mombasa

THE PORTUGUESE INVASION

This was the situation on the East African coast when Portuguese ships under Vasco da Gama arrived in 1498. The manifestly superior military and naval technology of the Portuguese and the greater unity of their command enabled them, in the years that lay ahead, to mount assaults upon the ill-defended East African city-states. Though the Portuguese became skilled at playing one small state against another, their global enterprise was such that they did not immediately impose direct rule.

The coastal cities were quite unprepared for the Portuguese. As early as 1502 the *shaykh* at Kilwa was obliged to agree to a tribute to the Portuguese, as the ruler of Zanzibar was later. A large Portuguese fleet established a Portuguese factory (trading post) at Sofala in 1505, and shortly afterward the Portuguese sacked both Kilwa and Mombasa. They sacked Brava in 1506 and forced Lamu and Pate to submit, and only Mogadishu escaped unscathed. Within eight years of their arrival they had managed to dominate the coast and the trade routes that led from there to India.

Their effect on Kilwa was serious from the first. A Portuguese fort was built, and for a time a garrison was maintained. More serious, however, was the Portuguese assumption of Kilwa's previous control over the Sofala gold trade. A more terrible fate struck the city in 1587 when the Simba, a marauding warrior people who had migrated from the Zambezi Valley, massacred its population.

Malindi, to the north, moved perhaps by its rivalry with Mombasa, proved to be the most reliable of Portugal's allies.

In retaliation against a revolt, the Portuguese sacked Mombasa for a second time in 1528, and from the mid-16th century Jesuit and Dominican missionaries arrived from Portugal and Augustinian canons set up houses at Faza, Pate, Mombasa, and Zanzibar. Converts were not, however, numerous.

Toward the end of the 16th century, first in 1585 and again in 1588, Turkish expeditions descending the northern coast with promises of assistance against the Portuguese encouraged the coast north of Pemba to revolt. A Portuguese fleet from Goa, which had been summoned by Malindi, secured a temporary submission in 1587, but a second fleet had then to be dispatched—more particularly against Mombasa. In 1589, in association with a force of marauding Simba, who had now moved northward, the Portuguese sacked Mombasa for a third time. Soon after, when a major expedition sent from Mombasa against Malindi was surprised and overwhelmed by the Segeju tribe, which had been moving down the coast, Mombasa passed much more firmly under Portuguese control. This was helped by the death of Mombasa's last Shirazi ruler, Shah ibn Mishhan, who, in leaving no clear successor, gave the Portuguese the opportunity to install Shaykh Aḥmad of Malindi in his place. In 1593, with an architect from Italy in charge and with masons from India to assist them, the Portuguese set about building their great Ft. Jesus at Mombasa. In the following year it was occupied by a garrison of 100 men.

Turkish attacks

With Mombasa's downfall, the major hindrance to Portuguese power on the East African coast was overthrown. Henceforth, throughout the 17th century Mombasa became the main Portuguese base in East Africa under its own Portuguese "captain." (Ft. Jesus rarely seems to have had a garrison of more than a few hundred men, and there seems never to have been more than 1,000 Portuguese at any one time upon the whole

coastline.) Hitherto, although the Portuguese had periodically ravaged the coastal cities and had inflicted considerable damage upon them, they had never exercised any continuing authority. Concerned, however, at the possibility of further Turkish incursions, they turned their established system of alliances involving the intermittent collection of tribute into a firmer and more ongoing dominion. They installed garrisons elsewhere than at Mombasa and brought about the downfall of a number of Shirazi dynasties, and, although they did not exercise day-to-day control over local rulers, they did make them dependent on them for their position. (Local rulers were in particular required to pay regular tribute to the Portuguese king on pain of dethronement and even of death.)

Portugal's chief interests were not imperial but economic. With Mombasa in their grip, they controlled the commercial system of the western Indian Ocean. Customs houses were opened at Mombasa and Pate, and ironware, weapons, beads, jewelry, cotton, and silks were imported. The main exports were ivory, gold, ambergris, and coral. There was a flourishing local trade in timber, pitch, rice, and cereals but few signs of any considerable traffic in slaves. Individual Portuguese traders often developed excellent relations with Swahilis in the coastal cities.

The new *shaykh* of Mombasa became much disturbed at both the arrogance and the importunities of the Portuguese, about which, in 1614, he complained to the Portuguese viceroy at Goa, in India. Given little satisfaction, he returned to Mombasa, but then fled to the *nyika* ("wilderness"), where, at the instance of the Portuguese, he was murdered. Though his son Yūsuf, who was taken to Goa to have a Portuguese education, returned to Mombasa in 1630—a baptized Christian and married to a Portuguese wife—he too ran afoul of the Portuguese. On August 15, 1631, he murdered the Portuguese commandant at Mombasa, renounced Christianity, adopted Islām, and ordered the massacre of most of the Portuguese in the town. Although there were risings sympathetic to him at Tanga and Mangata, his call to revolt secured little support elsewhere along the coast. Zanzibar especially held to its Portuguese alliance. Nevertheless, with Ft. Jesus under his control, Yūsuf successfully resisted the first of the two Portuguese fleets sent against him from Goa; and it was only as a result of disaffection in his own ranks that he had to flee the city before the second Portuguese fleet arrived. In the next few years he was actively engaged in stirring up a succession of revolts against the Portuguese—including a rebellion at Pate in 1637—until in 1638 he died at Jidda while seeking Turkish support.

Yūsuf's rebellion

Though the Portuguese managed to ride out these local upheavals, their authority over a much wider area was already being undermined by the intensity of Dutch and English commercial competition elsewhere in the Indian Ocean and by the rise of new powers on the Persian Gulf. They lost Hormuz to the Persians in 1622 and Muscat to the Imām of Oman in 1650. Two years later the Omanis launched their first major intervention into East Africa's affairs when in response to a Mombasan appeal the Imām sent ships to Pate and Zanzibar and killed their Portuguese inhabitants. As a consequence, Pate became the centre of East Africa's resistance to Portuguese rule. At Pate's invitation the Imām returned in 1660 to attack both Faza and Mombasa. In 1669 he thrust southward and narrowly failed to capture Mozambique. In 1678, following further intrigues, the Portuguese attacked Pate and executed the rulers of Pate, Lamu, Siu, and Manda, but in January 1679, on the appearance of a fleet from Muscat, they hurriedly retreated far to the south. Pate rebelled once again, in 1686, but its then ruler was captured by the Portuguese and executed at Goa in December 1688. Eventually, however, in 1696, in alliance with Pate—as upon most previous occasions—the Imām of Oman sailed to East Africa with a fleet of over 3,000 men, to lay siege to Mombasa, and, although Ft. Jesus was reinforced, the great Portuguese stronghold finally fell to Sayf ibn Sulṭān in December 1698. A few years later Zanzibar, the last of Portugal's allies in East Africa, also fell to the Imām.

Defeat of the Portuguese

THE OMANI ASCENDANCY

There ensued, after the Omani victory, a century during which, despite a succession of Omani incursions, the East African coast remained very largely free from the dominance of any outside power. Oman itself suffered an invasion by the Persians and was long distracted by civil conflict. Its originally successful Yarubi dynasty lost prestige as a consequence, fell from power, and was then superseded by the Albusaidis, who very soon found themselves preoccupied by conflicts at home. Moves against them also originated along the East African coast.

In 1727 Pate joined with the Portuguese to expel the Omanis, especially from Mombasa, where in 1728–29 Portuguese authority was momentarily restored. But the Mombasans wanted as little to be controlled by Portugal as by Muscat and soon evicted the Portuguese once again. Thereafter, Kilwa, Zanzibar, Lamu, and Pate largely kept themselves free from both Omani and Portuguese control. Distracted though it was by protracted internecine quarrels, Pate was pre-eminent in the Lamu archipelago and like all the other coastal towns was ambitious to preserve its independence.

Even so, Mombasa, in quite new circumstances, in the 18th century reached the apogee of its power as an independent city-state. The architects of this achievement were the Mazrui, an Omani clan who had provided some of the Imām's governors to Mombasa but who, because they were opposed to the Albusaidis, did not long persist in their allegiance to Muscat. They owed their authority in Mombasa itself to an ability to hold the balance between the rival factions in the Swahili population and also to their ability peacefully to overcome all but one of their dynastic successions. In 1746 a Mazrui notable, Liwali ʿAlī ibn Athman al-Mazrui, overthrew an Omani force that had murdered his brother. Soon after he seized Pemba and, but for a family quarrel, might have won Zanzibar; his successor, Masʿūd ibn Nāṣir, initiated a pattern of cooperation with Pate, which survived even Pate's successive civil upheavals, maintained close links with inland Nyika, and established Mazrui predominance from Pangani to Malindi.

Both Mombasa and Pate were disastrously defeated by Lamu in the battle of Shela, *c.* 1810. Pate's pre-eminence in the Lamu islands was destroyed, Mombasa's authority on the coast was diminished, and the way was open to Muscat's great intrusion into East African affairs. Lamu appealed to Oman for a garrison to assist it, to which Sayyid Saʿīd of Muscat very soon responded.

The Albusaidis, who had captured Kilwa in 1785, had their main footing upon the coast in Zanzibar, which had long held to its association with them. Thanks to the city's growing success, from the end of the 18th century onward, in turning itself into the main entrepôt for the trade in the area south of Mombasa, Zanzibar soon rivalled Mombasa as the focal point for the whole coastline and as such was both developed and used by Sayyid Saʿīd as the base for his growing ambitions. Having won the succession to Muscat after an internecine struggle following his father's death in 1804, Sayyid Saʿīd spent much of the next two decades establishing his authority there. (In this he was assisted by the British, who were much concerned to safeguard their route to India, which ran close to Muscat on its way past the Persian Gulf.) He wrested Pemba from Mazrui control in 1822 and by 1824 had installed a Muscat garrison in Pate as well, thus bringing to an end the previous influence that the Mazrui had exercised.

The rise of Zanzibar

Sensing the increasing threat from Muscat, the Mazrui appealed to the British for assistance. Though their application was formally denied, a British naval officer, Capt. W.F. Owen, on his own initiative raised a British flag of protection over Mombasa in 1824. Since the British had no desire formally to extend their authority to East Africa at this time, let alone to break with

their ally Sayyid Sa'īd, it was hauled down in 1826. This gave Sayyid Sa'īd his opportunity, and two years later he mounted his first assault upon Mombasa. His first attack failed to establish his authority, as did two further assaults in 1829 and 1833, and it was only when he successfully intervened in a dynastic dispute among the Mazrui, which followed on the death of a *liwali* in 1835, that he was able in 1837 to fasten his control over Mombasa and to topple the Mazrui from their position. His dominion along the whole coastline thus became assured, and after over a century's interval the East African littoral once more found itself dominated by a single outside power. Though this outcome owed much to the inability of the coastal towns to unite against an invader, it owed much as well to the striking personality of Sa'īd himself, to his investment in a navy, to his force of Baluchi soldiers (with which he supplemented his Omani levies), and to the support he received from the British.

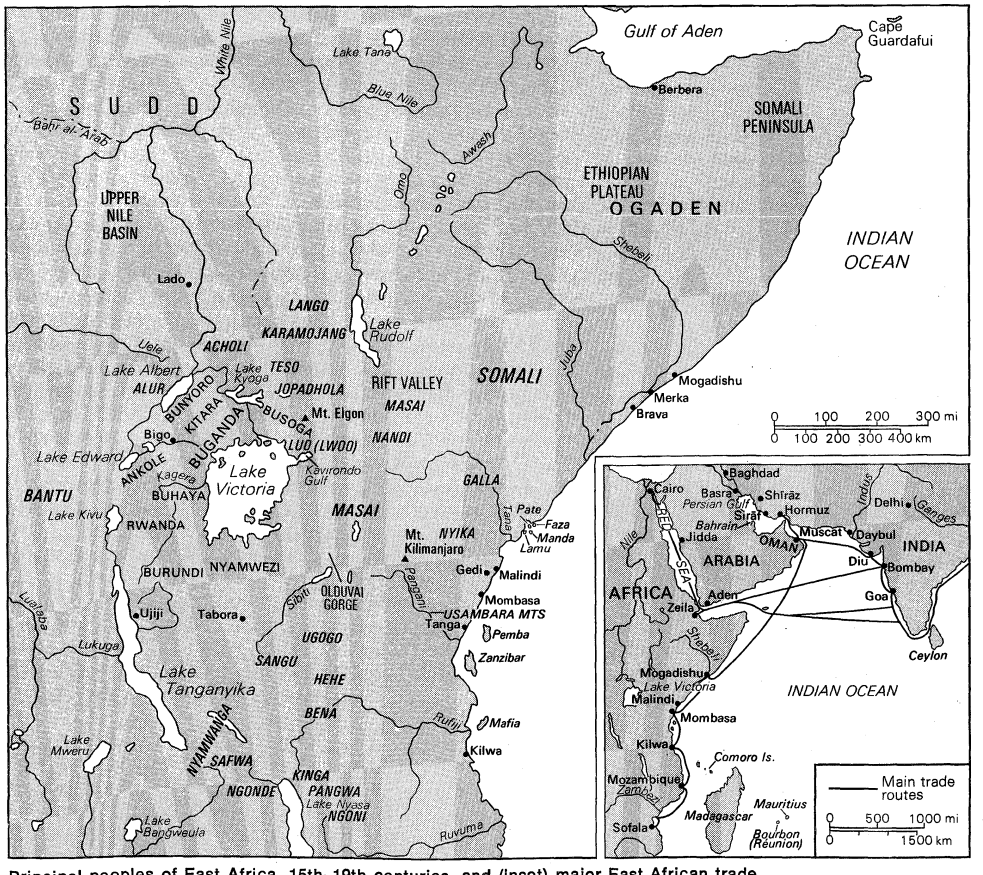
It also stemmed from his intimate association with the major economic developments then taking place along the East African coast. These began with a marked growth in the hitherto-marginal slave trade, particularly at first in the Kilwa region, more especially from 1780 to 1810 as a result of French demand for slaves in Mauritius and Bourbon. This was succeeded by the discovery that cloves could be successfully grown on Zanzibar and by the development of flourishing plantations. British pressure on Sa'īd to end the export of slaves to "Christian" markets came to fruition in 1822, when he reluctantly signed what became known as the Moresby Treaty. In the event, however, it made very little difference, both on the coast and in the interior, since slaves were being increasingly required for the plantations on both Zanzibar and on Pemba and for export to the Persian Gulf and beyond.

Increasing commercial activity brought Sayyid Sa'īd sufficient wealth to buy ships and pay troops. It also attracted to the East African coast migrant Indians, who became heavily involved in the country's economic expansion; and, together with the Arabs who were beginning to make profits from their clove plantations, Indians helped to finance the new upcountry trading caravans.

The increased economic activity that centred upon the islands of Zanzibar and Pemba served to enhance the importance of the smaller towns that stood on the mainland opposite. It also attracted an influx of European traders, of which the most important were the Americans. They were the first Westerners to conclude a trade agreement with Sayyid Sa'īd (1833) and the first also to establish a consul at Zanzibar (1836). (Their prime achievement was to capture the cloth trade to East Africa —so that cheap cotton cloth henceforth came to be known there as Americani.) The British followed with a trade agreement in 1839 and a consul in 1841. The French made similar provisions in 1844, and some Germans from the Hanseatic towns moved in at about the same time. British trade, however, never flourished, and in fact died away; but by 1856 America and France were both making purchases in East Africa of over $500,000 a year, while exports to India, particularly British India, were higher still. Some of the main items of trade, such as ivory, were traditional, but copal, sesame, cloves, cowries, hides, and coconut oil were also important. Because of this increased activity, Sa'īd's economy in due course became less dependent upon the export of slaves, and he therefore showed himself more ready than he might otherwise have been to accept the so-called Hamerton Treaty of 1845, by which the export of slaves to his Arabian dominions was forbidden.

Diminution of the slave economy

Since by this time the revenues from Sa'īd's East African territories had overtaken those he received from Oman, it is understandable that in 1840 he should have transferred his own capital from Muscat to Zanzibar. At his death in 1856, Zanzibar was firmly established as the East African coast's main centre, from which major new incursions into the interior had begun to radiate extensively.



Principal peoples of East Africa, 15th–19th centuries, and (inset) major East African trade routes in the Indian Ocean.

II. The interior before the colonial era

THE STONE AGE

The coast was never more than East Africa's fringe. Beyond the harsh *nyika*, which lay immediately inland and was nowhere pierced by a long, navigable river, thornbush country extended to the south, sometimes interspersed with pleasanter plains toward the centre, while to the north cooler forested highlands ran into harsher country. Westward lay the Rift Valley and, beyond, the regions of the great lakes whence the Nile ran northward through its usually impassable as-Sudd (marshes).

Sources of information

Since there are no written records antedating the last century or so for this region, its history has to be deduced from often uncertain linguistic, cultural, and anthropological evidence; from oral traditions—where they are available, which at best is only for recent centuries—and from archaeological findings. Since investigations and analyses are still at a very early stage and since the first hypotheses have proved vulnerable to criticism, the statements that follow must be only tentative. Moreover, bolder categories such as the Bantu are strictly only linguistic and must be treated with caution. Furthermore, all accounts of tribal migration must allow for innumerable short-run moves and may refer only to small—if important—minorities. They must also take account of probable interactions with other peoples en route and often, indeed, of extensive absorption. Above all, care must be exercised over anachronistic concepts of "tribe."

Two features of the pre-19th-century period may be stressed: first, although it seems to have been in this part of Africa that man first developed, in the three or four most recent millennia the key innovations in man's evolution seem to have occurred elsewhere; second, the extensive agricultural revolution in East Africa, which took place during this time, had the vital consequence that sizable populations grew up in areas of adequate rainfall, which could not be easily brushed aside by subsequent alien invaders.

During the earlier stages of the Stone Age down to *c.* 50,000 BC, hand ax industries were established in the Rift Valley areas of Kenya and of Tanzania (especially at Olduvai Gorge) and along the Kagera River in Uganda. During the Middle Stone Age (thence to *c.* 10,000 BC), new stone-making techniques evolved, and the use of fire was mastered. Spreading to other parts of East Africa, in the Late Stone Age man clustered into specialized hunting and gathering communities from which may have developed some still-existing ways of life. The largest number of relevant sites is close to the homeland of the Hadzapi—the last contemporary hunters and gatherers—and to that of the Sandawe, who are physically and linguistically akin to the Bushmen of southern Africa. Remnants of other hunting and gathering communities—such as the Pygmies of western Uganda—or at least the memory of them, are found in many places. Latterly, they often lived in precisely those highland regions where agriculture and animal domestication in East Africa first occurred.

Food production and the keeping of cattle seem to have begun in the highland and Rift Valley regions of Kenya and of northern Tanzania in the 1st millennium BC and to have derived from Caucasoid peoples who were probably southern Cushites from Ethiopia. Some traces of these interlopers remain among, for example, the Iraqw of Tanzania; and it may be that the age-old systems of irrigation found throughout this region owe their origins to this period as well. Agriculture preceded the smelting of iron in these areas, and hunting and gathering continued to be important for the domestic economy. It looks as if in due course southern Cushites spread deep into what is now southern Tanzania, but, so far as has been ascertained, food production did not develop in the period BC elsewhere in Tanzania, nor in what is now Uganda.

THE SPREAD OF IRONWORKING AND THE BANTU MIGRATIONS

It is still far from clear when and whence iron smelting spread to the East African interior. Certainly there was no swift or complete transfer from stone to iron. At Engaruka, for example, in that same region of the Rift Valley in northern Tanzania, a major Iron Age site, which was both an important and concentrated agricultural settlement using irrigation, seems to have been occupied for over a thousand years. Significantly, its styles of pottery do not seem to have been related to those that became widespread in the 1st millennium AD. It is a reasonable assumption that its inhabitants were Cushitic speakers, but it would seem that its major period belongs to the middle of the 2nd millennium AD.

The spread of cultivation techniques

The major occurrences of the 1st millennium AD involved the spread of agriculture—more particularly, the cultivation of the banana—to the remaining areas of East Africa. Simultaneously or perhaps previously went the spread of ironworking, with which "dimple based" and related pottery was often associated, and fairly certainly too the diffusion of Bantu languages—except in the core of the Cushitic wedge and to the north of an east–west line through Lake Kyoga. If, as seems probable, proto-Bantu languages had their origins in the eastern interior of West Africa, it does not seem inconceivable that over a lengthy period of time some of its speakers, probably carrying with them a knowledge of grain agriculture and conceivably a knowledge of ironworking, should have diffused along the tributaries of the Congo River to the savanna country south of the Congo forest into what is now the region of Shaba (Katanga). Nor does it seem inconceivable that the banana, originally an Indonesian plant particularly suitable in tropical conditions, should have spread to that same region up the Zambezi Valley (certainly the Malayo-Polynesian influences in Madagascar in the 1st millennium AD are well attested in other respects). At all events, the linguistic and archaeological arguments for a fairly rapid eastward and northward expansion during the 1st millennium AD from the Katangan area now have wide acceptance. Bantu languages came to dominate most of this region (many Cushitic speakers in what is now Tanzania seem to have switched over to them or to have been eliminated). More varieties of banana developed in East Africa than anywhere else in the world. Ironworking was soon prevalent. Remarkably similar pottery styles were being used; and, where rainfall, soil nutrients, and the absence of the tsetse fly allowed, population growth increased decisively.

Sometime before the middle of the 2nd millennium AD, some of the most interesting developments were occurring in the interlacustrine area—*i.e.*, the region bounded by Lakes Victoria, Kyoga, Albert, Edward, and Tanganyika. Vague accounts of rulerships in various parts of this area date from the first half of the 2nd millennium AD, and it is at least possible that they existed—though they may well have been judicial arbitrators or ritual leaders rather than more strictly political figures. Whether they had their origins in roving Cushitic or Nilotic cattle keepers from the north or northeast—as has been variously suggested—is impossible to say, though some such explanation would not be difficult to believe. What seems certain is that around the middle of the present millennium a sudden cultural and political climax was marked by a short-lived, though widely acknowledged, dynasty of Chwezi rulers.

The Bigo enclosure

Many are ready to assume that these Chwezi people are to be associated with the great earthwork sites at Bigo, Mubende, Munsa, Kibengo, and Bugoma, in western Uganda. That at Mubende seems to have been a religious centre. The largest is at Bigo, where a ditch system, over six and a half miles long, some of it cut out of rock, encloses a large grazing area on a riverbank. It looks as if it comprised both a royal capital and a well-defended cattle enclosure. Its construction must certainly have required a considerable mobilization of labour—which, apart from indicating that it must have been the work of a substantial political power, would support the view that the distinction between cultivators and a pastoral aristocracy, which later became typical of this area, is of very long standing. Radioactive carbon dating suggests Bigo was occupied from the mid-14th to the early 16th century. This correlates with the evidence of oral tradition

that around the turn of the 15th century the Chwezi were supplanted in the north by Luo (Lwoo) rulers of the Bito clan (who provided the dynasties that ruled in Bunyoro, Koki, Buganda, and parts of Busoga) and that they were superseded to the south by various Hima rulers of the Hinda clan (in Ankole, Buhaya, Busubi, and around to the southeast of Lake Victoria). Under these (and the corresponding Nyiginya dynasty in Rwanda) powerful traditional rulerships among the interlacustrine Bantu persisted after the middle of the 20th century.

Their relatively common experience was reinforced in the aftermath of the Chwezi dynasty by the prevalence among them of a variety of (often commemorative) Chwezi religious movements. In some areas these took the form of spirit-possession cults; in others, pantheons of deities were developed. In various guises—sometimes in support of the existing political order, sometimes against it—they spread into Bunyoro, Buganda, Busoga, Ankole, Buha, Rwanda, Burundi, and even to Nyamwezi country, in what is now Tanzania. So extensive a diffusion of a basically common religious tradition in any other part of the East African interior before the much later arrival of Islām and Christianity was rare indeed.

The *ntemi* kingdoms

In northwestern Tanzania, dynasties of a pre-Chwezi kind apparently spread from the interlacustrine area during the middle centuries of the present millennium. *Ntemi* (as the office was called) became prevalent among both the Sukuma and the Nyamwezi. They (the *ntemi*) were probably as much ritual leaders as political rulers. Certainly they do not seem to have exercised before the 19th century that "state" authority that was characteristic of the later interlacustrine rulers. By about the 16th century there may have been an extension of this style of chieftainship southward into southwestern Tanzania. At all events, the chiefly groups among the Nyamwanga, the Nyika, the Safwa, the Ngonde, the Kinga, the Bena, the Pangwa, the Hehe, and the Sangu have common traditions of origin, and it seems clear that they are to be distinguished from their significantly different, matrilineal neighbours in southern Tanzania, Zambia, and Zaire. There also seem to have been secondary movements of *ntemi*-like institutions in the 18th century to Ugogo, Safwa, Kaguru, Kilimanjaro, and Usambara. At the same time, the development of chieftainship in these other areas of Tanzania may originally have occurred independently of influences from elsewhere.

Although it is believed that some of the northeastern Bantu were once ruled by chiefs, these did not survive. The spread of some Bantu to the northern coast of East Africa during the 1st millennium AD is supported by the memory of a settlement area named Shungwaya situated to the north of the Tana River. (Its name appears on European maps from 1596 onward.) The traditions and origins of several Bantu peoples who then moved southward are traced to this area. It would seem as if it had its heyday as a Bantu settlement area between perhaps the 12th and the 15th centuries and was subjected to a full-scale invasion of Cushitic-speaking Galla in the 16th century. Bantu are no longer found so far north. There is controversy as to whether the ancestors of the present Kamba and Kikuyu of Kenya were from Shungwaya, but it would seem that they probably broke away from there some time before the Galla onslaught. It has been suggested, indeed, that the Kikuyu spread through their present territories from 1400 to 1800.

The old Cushitic wedge checked them from spreading farther westward. This extended, as it would seem to have done for two or more millennia past, over both sides of the Kenyan and northern Tanzanian Rift Valley. In the middle of the present millennium it was subjected to one of the multiple waves of invading Nilotic peoples —who were partly agriculturists and partly pastoralists—that moved into much of the northern and northwestern parts of East Africa during the three or four centuries that ensued.

MIGRATIONS OF NILOTIC AND CUSHITIC PEOPLES

The supersession of the Chwezi by Luo dynasties in the northern interlacustrine region at about the end of the 15th century resulted from the migrations of Nilotic peoples southward—in this instance, it has seemed, from a cradleland in what is now the Sudan. For some 18 generations or so Bito rulers of Luo origin held sway over the kingdom of Bunyoro-Kitara, to the east of Lake Albert. Though at first their dominion seems to have been widely extended, they began to be rivalled in the 16th and 17th centuries by the rise of Buganda, under its ruler, or *kabaka*. Working on interior lines and based upon a particularly fertile region, Buganda developed a strength and cohesion that from the 18th century onward was to make it—with Rwanda—one of the two most formidable kingdoms of the region.

The Luo dynasties

Their Luo rulers and such followers as had accompanied them were soon fully absorbed into the Bantu population of these kingdoms. Immediately to the north (where the Bantu did not extend) there occurred the greatest independent expansion of the Luo peoples, who formed the Acholi; provided ruling groups for peoples to the west who came to be called Alur; and bred the Jopaluo and Jopadhola to the east and also the sizable Luo populations who, between the mid-16th and the mid-18th centuries, came to settle on the northern side of the Kavirondo Gulf to the northeast of Lake Victoria and spread thereafter to its southern shore as well.

Over to the east, into the former Cushitic domain that centred upon the Rift Valley, there appears to have been, in about the middle of the present millennium, a great expansion of Kalenjin peoples, whom it has been usual to call Nilo-Hamites but whom it would now seem proper to call Nilotes (Highland Nilotes to distinguish them from, among others, River–Lake Nilotes such as the Luo). They seem to have absorbed most of the previous southern Cushites who remained there and also to have successfully held the core of this ancient wedge, if not its earlier dimensions, against further Bantu incursions. By 1700, however, a second expansion into this old protrusion was beginning. During the 18th century the Masai (Plains Nilotes, as they are now being called) spread over most of the area, until they came to be found as far south as Gogo country in central Tanzania. Already divided into pure-pastoralist and mixed-agriculturist subtribes, they were soon to be found to the east near Mt. Kilimanjaro. The earlier Kalenjin thus found themselves confined to the hillier country between the Rift Valley and Lake Victoria, where—constituting the Keyu (Elgeyo), the Suk (Pokot), the Nandi, the Kipsigis, and the Tatoga of more recent times—they entered into a variety of interactions with their various Luo and Bantu neighbours. Farther to the north in the areas beyond Mt. Elgon a shorter run series of migrations by other Plains Nilotes was simultaneously taking place. First, in the 17th century, the Lango began moving southwestward (and became much affected by their River–Lake Nilotic neighbours, the Acholi). Then, in the 18th century, the Teso, Karamojong, and others began also to move in various southward directions.

It has been suggested that all these Highland and Plains Nilotic migrations were set off, both before and after the middle of the present millennium, by successive pressures from the eastern Cushitic Galla to the north, who were themselves being pressed upon by their own Cushitic neighbours, the Somalis.

If the basic Somali populations first spread into the horn of Africa from farther south, the first important Somali leaders were those who crossed over from Arabia in the 10th and 12th centuries. From that time onward the Dir, Darod, and Issa Somalis began to press steadily upon the Galla to the south. By the middle of the 16th century they had taken over much of northern Somaliland and the Ogaden region of Ethiopia. Since they were Muslims, the Somalis became involved in the Muslim conflict at this time with Christian Ethiopia. Already, moreover, the Galla in the 15th century had begun their own most wide-ranging expansion, which by the end of the 16th century was affecting much of the Ethiopian highlands. What was more, their southern thrusts were pushing on past Shungwaya and the Tana River and by the 17th century were even penetrating areas inland from

Mombasa. Behind them, however, the Somalis were also once more on the move and by the end of the 17th century were occupying all the territory on which the Galla had previously settled north of the Juba (Giuba) River.

The Somali "sultans"

The Somali "sultans" were scarcely more than figureheads and lacked any firmly institutionalized political power. They were often indeed less important than the elders of their clans. This, with variations, was the case with most of the leaders of the Nilotes as well. Indeed, Nilotes established "states" only where, as in the interlacustrine area over to the west, they came to rule other peoples who may very well have had traditions of rulership before they arrived. Such distinctions were to be of immense importance for the future. Among the Galla, the Masai, the Kalenjin, and—by adoption—some of their immediate Bantu neighbours, there were age systems of one kind or another. Farther south there were Bantu who had *ntemi* and other vague versions of chieftainship; there were somewhat similar figures among the River–Lake Nilotes. Only in the interlacustrine area were there powerful rulers and well-established kingdoms.

DEVELOPMENTS IN THE 19TH CENTURY

Political activities. In the 19th century there were a number of new movements and new political developments among the enlarged populations of East Africa before the advent of Western imperialist forces. In the south some parties of Ngoni—themselves originally refugees from the Zulu wars in southeastern Africa—moved northward following the breakup of their main body after 1845, at the north end of Lake Nyasa. Some Ngoni groups made their way to Songea; another struck north to Lake Victoria. They carried, to the area west of Lake Tanganyika, a new style of raiding and appear to have precipitated among such peoples as the Holoholo and the Ndendehule, the Sangu, and the Bena—in part, at least, as a safeguard against Ngoni raids—the creation of new political institutions, including more powerful rulerships. The most notable of such developments were, in the first place, among the Hehe, where, under Munyigumba and then on his death, in 1879, under his son Mkwawa, a powerful state was built up; then among the Nyamwezi, where between 1870 and 1884 the warrior chief Mirambo established a powerful personal rulership; and also among the Kimbu, where, between 1870 and 1884, Nyungu and his *ruga-ruga*s (or bands of warriors) created a dominion that survived his death.

Nothing quite so striking occurred to the northeast. Conflict persisted between the smaller Chagga rulers on Mt. Kilimanjaro. Around 1800, Orombo and, after 1860, Sina of Kibosho and Rindi of Moshi built up powerful combinations; but no one established a single political authority. In the middle decades of the 19th century, Kimweri enjoyed a considerable dominion in the region of the Usambara Mountains, but on his death, sometime in the 1860s, his rulership disintegrated. Further such disintegrations occurred in the 19th century among the rulerships of Buha and Buzinza, at the south end of Lake Victoria, and among the Buhaya kingdoms on its eastern shore. The kingdom of Mpororo, to the westward of Buhaya, had already broken up. By the 1890s its neighbour Nkore seems to have been in danger of disintegrating as well. Earlier in the century, Toro, to its north, had broken away from Bunyoro, previously the most extensive of the kingdoms in this area—while fragmentation was almost endemic among the Busoga kingdoms to the east.

The kingdom of Rwanda

But there were also political growing points in the interlacustrine area, where one of the largest kingdoms, Rwanda, first consolidated its rear by annexing Lake Kivu, then, in the aftermath of a succession war, swallowed the small kingdom of Gissaka, to its east. It failed to defeat Burundi, to the south, but under its *mwami*, or ruler, Kigeri IV (who reorganized its military forces) it extended its control by raiding to the north.

Its power was equalled in this region only by the kingdom of Buganda. Having annexed the large area of Buddu, to its southwest, in the late 18th century, Buganda thereafter generally refrained from any further territorial extensions. Its rulers steadily increased their authority at home by enhancing the power of appointed chiefs at the expense of the clan leaders, while abroad they preferred to make satellites rather than subjects of their neighbours. They had a good deal of success eastward in Busoga and southward along the western shore of Lake Victoria and around its southern rim. In the 1870s and 1880s Buganda's protégés were on several occasions installed in petty rulerships in Busoga. In 1869 Bunyoro successfully survived Buganda interference in one of its succession conflicts (as Nkore did in 1878) and indeed in the 1870s and 1880s was renewing its strength. As in Rwanda and farther south, military reorganization was of vital importance. Bunyoro's improved position turned much on its new military formations, the *abarasura*, while Buganda's successful predation owed a good deal to its new military efforts under the *mujasi*, or military commander, as well as to the building of a formidable fleet of canoes.

To the north and northeast the previous migrations of the Luo from west to east were followed in the 19th century by a new wave of migrations from east to west. The Lango, for example, further expanded in two southward and westward waves toward Lake Kyoga and toward the Victoria Nile, where they ran up against the Acholi. To their south the Teso and the Kumam were also moving west and south. A flourishing trading network developed around Lake Kyoga, and during the course of the 19th century the Teso in particular were converted from pastoralism to agriculture.

Activity was rife also among the pastoral peoples to the east. In about 1850 the Turkana began to migrate from a base west of Lake Rudolf. South of them the Suk were the terror of their neighbours, but by the 1880s they in their turn were being threatened by the Turkana. To the east the Somalis were once more on the move. After 1850 one of the Darod Somali groups, the Marehan, crossed the Juba in force. In 1865 they went on to break the Tana Galla and by 1880 had turned on the Boran. Pagan peoples in this region were now being dominated by Muslims, and peasants by nomads from the north. Moreover, after 1880 especially, the Ethiopians, fired by their new imperial ambitions, were advancing too.

Southward stood the Masai, the warrior people of the plains and open plateaus north and south of the string of Rift Valley lakes west of Mt. Kenya. From 1830 onward their various subtribes were engaged, under the auspices of their rival *laibon*s, or ritual leaders—among whom Mbatian, who succeeded his father, Subet, in 1866, was the most famous—in a succession of internecine conflicts largely over cattle and grazing grounds. Their wars denuded the Laikipia and Uasin Gishu plateaus of their former Masai occupants and increased the diaspora of Masai, the so-called Wakwavi, who, being deprived of their cattle, switched to agriculture. They also helped the Nandi, who, with the Uasin Gishu Masai now troubling them no more, took to raiding on their own account from a base between the Rift Valley and Lake Victoria. Under the leadership of their *laibon*-like *orkoiyot*s the Nandi and their kinsfolk, the Kipsigis, were soon the new powers in the land. Some of their neighbours who lived in open country put up defense works against them—the Baluyia, to the west, for example, built mud walls around their villages—while others, such as the Teita (Taita), the Kamba, and the Kikuyu, who lived on higher ground and in forest country, were rather better placed and from their carefully guarded fastnesses could defy the Masai. On the edges of their country they even entered into some permanent trade and marriage relations with the Masai. Where the soil was fertile, moreover, such people considerably increased their populations. By the 1880s, for example, Kikuyu country was very populous. Though they had no chiefs, "prominent men" were accorded a recognized status among them, and by the close of the century some of these were fighting each other for local supremacy.

Arabs and East African trade. On the coast, following the death, in 1856, of Sayyid Sa'īd, his erstwhile domin-

ions in East Africa were split off from the imamate of Muscat. With the death of the last Myenyi Mkuu in 1873, the authority of the Albusaidi sultans on Zanzibar itself became complete, although there were still revolts against them on the coast—particularly at Pate and Mombasa (where the Mazrui retained their pre-eminence despite successive defeats)—and at Kilwa, to the south.

Closing of the slave markets

This arose chiefly from the Sultan's acceptance of the further measures against the East African slave trade pressed upon him by the British consul at his court. By the 1860s some 7,000 or so slaves were being sold annually in the Zanzibar slave market. In 1873 a treaty with the British closed the market at Zanzibar, and Sultan Seyyid Barghash, by two proclamations in 1876, reduced the export from the mainland to a trickle. As it happened, however, there was then a final period of unprecedented slaving on the mainland, where the trade in slaves had generally been closely connected with the trade in ivory and the demand for porters was still considerable.

Trade in the East African interior began in African hands. In the southern regions Bisa, Yao, Fipa, and Nyamwezi traders were long active over a wide area. By the early 19th century Kamba traders had begun regularly to move northwestward between the Rift Valley and the sea. Indeed, it was Africans who usually arrived first to trade at the coast, rather than the Zanzibaris, who first moved inland. Zanzibari caravans had, however, begun to thrust inland before the end of the 18th century. Their main route thereafter struck immediately to the west and soon made Tabora their chief upcountry base. From there some traders went due west to Ujiji and across Lake Tanganyika to found, in the latter part of the 19th century, slave-based Arab states upon the Luapula and the upper reaches of the Congo. In these areas some of those who crossed the Nyasa–Tanganyika watershed (which was often approached from farther down the East African coast) were involved as well, while others went northwestward and captured the trade on the south and west sides of Lake Victoria. Here they were mostly kept out of Rwanda, but they were welcomed in both Buganda and Bunyoro and largely forestalled other traders who, after 1841, were thrusting up the Nile from Khartoum. They forestalled, too, the coastal traders moving inland from Mombasa, who seemed unable to establish themselves beyond Mt. Kilimanjaro on the south side of Lake Victoria. These Mombasa-Pangani traders only captured the Kamba trade by first moving out beyond it to the west. By the 1880s, however, they were operating both in the Mt. Kenya region and around the Kavirondo Gulf (where some of them became particularly associated with Mumia of the Wanga, or Hanga) and were even reaching north toward Lake Rudolf.

Suggestions that the Sultan of Zanzibar might at this time establish his dominion over the East African interior prompted Sultan Barghash to send a Baluchi force to Tabora. But the idea was never pursued. A comparable notion, however, led Khedive Ismāʿīl Pasha of Egypt to appoint in 1869 the Englishman Samuel White Baker as governor of the Equatorial Province of the Sudan in order that he might carry the Egyptian flag to the East African lakes. Though Baker reached as far south as Bunyoro in 1872, he was soon obliged to leave. His successor, Charles George Gordon, proposed to circumvent both Bunyoro and Buganda by going straight up the Nile's banks. Gordon also encouraged the ultimately frustrated expedition of McKillop Pasha, which the Khedive sent around the horn of Africa and down the East African coast to the mouth of the Juba River. Mutesa I, *kabaka* of Buganda, frustrated Gordon's efforts on the Nile; and by the early 1880s, with bankruptcy in Egypt and the Mahdist revolt in the Sudan, only remnants of the Egyptian enterprise remained.

III. The colonial era

THE ARRIVAL OF THE EUROPEANS

The Egyptian incursion had been the climax to the search by many European explorers for the headwaters of the Nile. This quest had obsessed the later years of the Scots missionary David Livingstone and had prompted the discovery in 1858 of Lake Tanganyika by the English expedition of Richard Burton and John Hanning Speke. Speke returned first to discover Lake Victoria in 1858 and then with James Grant in 1862 became the first white man to set eyes on the source of the Nile, which Speke named Ripon Falls. By circumnavigating Lake Victoria 12 years later, H.M. Stanley stilled the controversy that had ensued in Europe over Speke's claim.

The origins of missionary incursions

The revelations of these explorers, the example of David Livingstone, concern in western Europe over the East African slave trade, and the Catholic and evangelical fervour that existed there inspired the invasion of the East African interior by a motley collection of Christian missionary enterprises. Johann Ludwig Krapf and Johannes Rebmann of the Church Missionary Society, who had worked inland from Mombasa and had, in the 1840s and 1850s, journeyed to the foothills of Mounts Kenya and Kilimanjaro, were followed by a British Methodist mission. Catholic missionaries reached Zanzibar in 1860 and settled at Bagamoyo in 1868. An Anglo-Catholic mission first tried to establish itself in the Shire highlands, then in 1864 transferred to Zanzibar. Anglican missionaries arriving in Buganda in the mid-1870s at the request of Kabaka Mutesa were soon followed by Catholic White Fathers—there and elsewhere on Zanzibar's Tabora route—while the London Missionary Society sent men both to Unyamwezi and to Lake Tanganyika.

There were, of course, a number of localized religious movements among the peoples of East Africa during the 19th century. These included the Mbari cult among the Nyakyusa, the Nyabingi in Rwanda, and the Yakany movement north of Mt. Ruwenzori. None of them, however, spread in quite the way that the Chwezi movement had earlier. Islām, on the other hand—spread widely at the instance of the Zanzibari traders and long established on the coast—had secured a scattering of converts in the interior as in the key kingdom of Buganda.

This was the scene onto which Christian missionaries first entered. Although by 1885 there were nearly 300 of them in East Africa, they did not, initially, win many converts, and those they at first obtained came only from among freed slaves and refugees from local wars. After 1880, however, they made important conversions in Buganda, and by the end of the century Christianity was spreading in the Lake Victoria area over most of the region in which the Chwezi movement had previously percolated—and before very long over a much larger area as well. There were a good many African peoples who, secure in their own beliefs and rituals, were impervious to such alien religion, others who were actively hostile, to them, but others, too, who found that one or other of them met some of their deepest needs. In Buganda in 1876 there were even Muslim martyrs, when their ruler, the *kabaka*, turned against them, and, a decade later in similar circumstances, Christian ones too.

Philanthropic, commercial, and eventually imperialist ventures followed these evangelical endeavours. Nothing, however, of great moment occurred until in 1885, when a German, Karl Peters, riding a tide of diplomatic hostility between Germany and Britain in Europe, secured the grant of an imperial charter for his German East Africa Company. With this, the European scramble for east central Africa began. The genesis of this movement is subject to much controversy. Certainly all the major—and several minor—European states wanted new imperial possessions from which they could draw raw materials to feed their industrial revolutions and in which they could establish new markets. They also had some large-scale strategic interests and some strong feelings about what they saw as a call to civilize Africa.

There had already been rivalry over the Somali areas among the Egyptians, the Italians, the Ethiopians, the British, the French, and the Sultan of Zanzibar. When the Egyptians in 1884 were forced to withdraw from Harer, Zeila, and Berbera to consolidate their strength against the Mahdists in the Sudan, the Italians, somewhat encouraged by the British, who were fearful of French designs, proclaimed a protectorate in 1885 over the Eritrean coast. In the same period the French estab-

Italy's imperial pretensions

lished a protectorate in the Jibuti area, while the British declared a protectorate over the Somalis (who were becoming fearful of Ethiopian ambitions) in the area about Berbera. The Treaty of Uccialli (Wuchale), by which in 1889 the Italians believed that they had made Ethiopia into their protectorate, was disputed by the emperor Menelik II, who in 1896 inflicted a resounding defeat upon them at the Battle of Adowa. In the meantime, various Italian enterprises had managed to establish their claims to the area that was soon to become Italian Somaliland, and first in 1891 and later in 1894 various Anglo-Italian protocols settled the boundaries between the Italian and British spheres of influence. Moreover, at Harer in 1896, the British and the Ethiopians agreed to a treaty to define their boundaries.

The Anglo-German agreement of 1886

In east central Africa the key occurrence was the Anglo-German Agreement of 1886, by which the two parties agreed that their spheres of influence in East Africa should be divided by a line running from south of Mombasa, then north of Kilimanjaro to a point on the eastern shore of Lake Victoria. This ended any further question of Zanzibari dominion in the East African interior and began the extraordinary process by which the territories and subsequently the nations of East Africa were blocked out first upon the maps far away in Europe and only later upon the ground in East Africa itself, invariably, moreover, without regard for any social or cultural boundaries in the areas concerned. A British East Africa Association was created in 1887 to compete with the German company and it was transformed in 1888 into the Imperial British East Africa Company. The ensuing rivalry between the two European companies provoked such rivalry along the coastline and in the interior that the British and German governments were eventually obliged to conclude a further Anglo-German Agreement (1890), whereby German activity was confined to the area south of the 1886 line; Zanzibar was left to become a British protectorate; the boundary on Lake Victoria was bent due west to the border of the Congo; and the Germans acknowledged British claims to the headwaters of the Nile—while maintaining a crucial stretch of the "Cape to Cairo" route upon which much British romanticism had been lavished. Further international agreements—such as the Anglo-Congolese Agreement of 1894—and a good many detailed adjustments followed.

COLONIAL RULE

Both the British and the German companies soon surrendered such territorial controls as they had to direct imperial administrations—as did the various Italian enterprises. The German government took over from the German East African Company in 1890, while the British company handed over its claims upcountry to a British protectorate administration in 1893. In 1895 it also surrendered its territory nearer to the coast. Only gradually, however, was even a limited colonial administration extended upon the ground—in the western British territories by 1914, in parts of the northern territories as late as the 1920s.

The extension of administrative control was marked by violent conflict, in which the Europeans used mercenaries, sometimes from elsewhere in Africa, sometimes even from India. Innumerable military clashes in the German sphere included a major conflict on Mt. Kilimanjaro in 1892, a protracted war with the Hehe between 1888 and 1898, and stern resistance until 1899 from the Yao ruler of the Makonde plateau. The British, except for one occasion, successfully avoided clashes with the Masai, but they launched five distinct campaigns against the Nandi between 1895 and 1908. There was war in Buganda between 1891 and 1893 and again between 1897 and 1899, protracted conflicts in Bunyoro, a number of encounters in Busoga and Ankole, and, in the next two decades, countless clashes with the peoples throughout the northern British spheres.

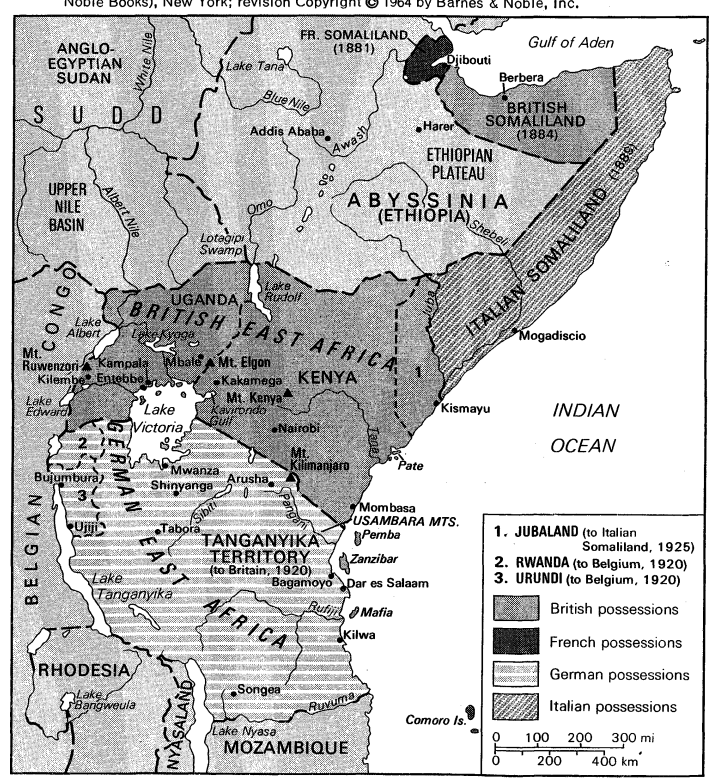
Nothing in this region equalled the resistance of the Somalis under Sayyid Muḥammad Abdille Hassan, a leading member of the militant Ṣalīḥiyah branch of the Aḥmadīyah order. Deeply perturbed by the threat to his people's Muslim faith by Christian colonial activity, he raised a great revolt against the British in 1900 and withstood them four times before (at Illig in 1905) agreeing with the Italians to settle in their sphere. He also fought the Ethiopians. He took up arms again between 1908 and 1920 until his stronghold at Taleh fell, and he died shortly after.

Somali resistance

Resistance was also sharp elsewhere. In 1888–90 a major assault under Bushiri was mounted by the coastal peoples against the Germans. Elsewhere the resistance of the Nandi and the Banyoro against the British, and of the Hehe against the Germans, was both skillful and courageous. As in Somaliland, some of the most forceful resistance was mounted in association with religious movements. Of these, some, such as the Yakany cult, which originally developed against the Nile traders, inspired a considerable upheaval against the British in the West Nile district of Uganda in 1919. To the south the Nyabingi cult, which had first developed in resistance to Rwanda depredations, became a scourge to the British in Kigezi District through much of the 1910s and again in 1928. Other movements, such as the Mumbo cult among the Gusii or the Kamba exorcism cults, seem to have been precipitated by the European pressures. Of these, the most notable was the fiercely repressed Maji Maji rising of 1905–07. Originally a protest against an ill-advised cotton scheme, it rapidly developed—in association with the Kolelo cult—into a widespread revolt against the Germans throughout the southern parts of their territory.

Some African peoples had welcomed the imperial rulers from the outset, seeing in them likely allies who could defend or enlarge their own local interests. Alliances thus developed—between, for example, Sangu sultans, Masai *laibons*, Buganda chiefs, and the European rulers—which often became the basis for a stabilized colonial order.

From W. Shepherd, *Historical Atlas:* Harper & Row, Publishers (Barnes & Noble Books), New York; revision Copyright © 1964 by Barnes & Noble, Inc.



European colonies in East Africa to 1914.

From their capital in Dar es Salaam, the Germans divided their colony into 14 "stations," and, over and above their links with certain local chiefs, they widely extended their Zanzibari inheritance of appointed *liwalis*, *akidas*, and *jumbes*.

The British sphere, as seen above, became divided between two British administrations. From the very beginning it had two focuses deep in the interior: the remnants of the Egyptian Equatoria Province and the promising Christian missions in Buganda. The Imperial British East Africa Company sought, originally, to take over the first —and thus early on directed its attention to the far interior. But during 1890–92 its caravan leaders were diverted to Buganda, because of a favourable opportunity that presented itself to them there as a result of an oligarchal revolt against its *kabaka*, Mwanga, in 1888. Capt. Frederick (later Lord) Lugard and his successors, by interposing themselves between the Muslim and Christian parties in the internecine conflict that ensued, were able by 1894 to establish British control over the most powerful African regime in the interior. And when, as a result of its impending bankruptcy, the company announced its withdrawal from Buganda, a British imperial protectorate was eventually established over "Uganda" in 1894. Since it was obvious that its commissioner would have difficulty in exercising authority nearer the coast, in 1895 a separate British East Africa Protectorate was established over the areas nearer the coast. The boundary between the two originally ran on the east side of the Rift Valley, but in 1902 it was shifted westward to the vicinity of Mt. Elgon. It became in due course the international boundary between Kenya and Uganda.

Buganda and the Uganda Protectorate

Thereafter, the two British protectorates developed very differently. Despite the Sudanese mutiny and the Kabaka's revolt in 1897–99, the new regime in Buganda—secured as it came to be by the so-called Uganda Agreement of 1900—was locked with a victorious oligarchy of Christian chiefs, and Buganda's inherited political institutions were readjusted so that under a new and less powerful *kabaka*, Daudi Chwa (1897–1939), three senior chiefs formed a three-man Buganda government, while the others came to be stratified into a three-tiered hierarchy of county, subcounty, and parish chiefs with both administrative and judicial powers. In the years that followed, this Buganda model was exported in one way or another to the rest of the Uganda Protectorate. In some areas former war-band leaders or petty rulers found themselves uncomfortably pressed into being administrative chiefs. Especially in the "stateless societies" in the north, where specialized political offices had not previously existed, the system was largely novel. Between 1895 and 1905 a Buganda general named Kakunguru had established his own Buganda-like "empire" there, but it was taken over by the British, who then used "Buganda Agents" in extending their rule there and elsewhere. After 1910 these were steadily replaced by local men, but everywhere a hierarchical structure, as first elaborated in Buganda, became established, and British officials in Uganda took pleasure in declaring, "The best method of developing the country is through the native governments."

From an imperial point of view, the advantage of working through "native governments" led British officials next door to proclaim paramount chiefs among the Kikuyu, Masai, and Baluyia. The belief that such appointments secured a better position for Africans led many of them there (and later in Tanganyika) to press for them. But in the East Africa Protectorate there were major problems. Scarcely anywhere within its boundaries was there any previous, distinctly political office on which to build. There was some possibility that, as in the south, *liwalis* and *jumbes* would be employed, but, when the hub of the country was shifted upcountry from Mombasa to Nairobi in 1902, the hope that local chiefs could be enlisted into British service was soon frustrated. Though an attempt was made to enlist local courts of elders after 1908, they were no use for ordinary administrative purposes. The result was that the Kenya administration came to base itself thereafter not upon any pre-existing indigenous authorities but upon its own imposed structure of directly appointed chiefs and headmen. There can be little doubt that they were more appropriate to many local societies than were those ritual leaders or petty rulers who were pressed into administrative service. But the appointed chiefs in Kenya (as the proponents of paramount chieftaincies very well understood) could neither withstand the colonial administration nor provide the more sympathetic administrators with a "native" instrument of development. Further attempts to enlist the cooperation of local leaders in the creation of local native councils in the 1920s never matched the success of the development of councils in Uganda, and local administration in Kenya not only remained very largely under European direction but was mainly controlled from the centre as well.

In several respects World War I (1914–18) was very important for East Africa. There was protracted conflict between the British (and the South Africans) and the Germans in Tanganyika—both sides used large numbers of Africans—and it was only in 1918 that the latter finally surrendered. Britain was then granted a League of Nations mandate over Tanganyika. Important new developments in the administration of the country then occurred when Sir Donald Cameron (governor 1925–31) and his able secretary for native affairs, Philip Mitchell, introduced "indirect rule," the system whereby local authority was formally vested in traditional African authorities. By contrast with Kenya there had in fact been a variety of African chiefs in this region, as seen above, in the past (though none of them with the power of the Uganda rulers), so that the policy of indirect rule could be pursued there with some success.

The new order. Important patterns common to the three territories now began to emerge. A money economy was beginning to be adopted, and Christian missions were becoming widespread in all three. Western education was primarily in the hands of Christian missionaries, and much that was dramatically new in East Africa derived from this connection. Out of the mission-school system came the beginnings of a Western-educated African elite. With no local university education and little opportunity to travel abroad, they combined pride in their heritage with the belief that since they were powerless to overthrow the colonial system their best course was to seize the opportunities it provided. Some of them among the Kikuyu of Kenya, with Johnstone (later Jomo) Kenyatta in their midst, vehemently defended during the late 1920s their people's practice of cliterodectomy (female circumcision) against the denunciations of the missionaries. Samwil Chiponde of Tanganyika, on the other hand, was heard to declare in 1925, "People can say what they want, but to the African mind, to imitate Europeans is civilisation." Probably, however, Kabaka Daudi Chwa spoke for the majority when in his pamphlet "Education, Civilisation and Foreignisation" (1935) he declared himself broadly in favour of the first two but opposed to the last.

The effects of missionary education

It was usually these men, who as chiefs, clergymen, schoolmasters, prosperous farmers, and entrepreneurs, led the way in the immensely important processes of aggregation—and particularly of political aggregation—that in a variety of ways took place at a local level. Previously, many of the bonds uniting those who spoke the same language and possessed closely similar cultural practices had been tenuous. But as, in the 20th century, they found themselves caught in a larger framework and a fast-changing environment, their sense of a common tribalism was heightened. This development was often encouraged by the colonial authorities, whose colonial administrative structures often provided the chief institutions within which it advanced. Local administrative boundaries were often adjusted to meet the resulting strains.

"White man's country." It was precisely these processes of aggregation that became caught up in the major uncertainty that hung over East Africa in the first half of the 20th century: whether it was or was not to become a "white man's country." The doubts here stemmed from a basic dispute over whether the economic development of East Africa should be based on African peasant farming or on European estates. Though there were always champions of African production, most European governing circles at the beginning of the 20th century

believed that the right way to proceed, wherever conditions allowed, was to build up the production of European estates. As the railway, which had been built through the British East Africa Protectorate between 1895 and 1901, showed little sign of paying its way, strenuous efforts were made to attract white settlers to the Kenya highlands in order to make both the railway and the protectorate pay—the assumption being that all necessary labour could be procured from the neighbouring well-populated African areas. German settlers had meanwhile been encouraged to settle in the German sphere.

Opposition to the settler doctrine

The settler doctrine did not go unchallenged even among British officials in the Kenyan highlands. In Uganda, more particularly under its then governor, Sir Hesketh Bell (1905–10), the production by African peasants of cotton was actively fostered—especially in the south central and eastern parts of the country; while moved in part by developments in Uganda, the German governor Albrecht von Rechenberg (1906–12) made a notable effort to encourage African peasant production in his territory and even secured German government finance to build Tanganyika's central railway so as substantially to forward its prospects. His plans were, however, resisted by the German settlers, who, after his departure, gained a political pre-eminence that exceeded that achieved at this time by the British settlers to the north. There, those British administrators who were concerned to withstand settler pressures upon Africans were by 1910 in full retreat (especially after the dubiously manipulated transfer of the Kenyan Masai into a compact southern reserve), while even in Uganda a handful of European planters were now assuming an influence that threatened the predominance of the African peasant sector.

The effects of World War I

World War I had a paradoxical effect upon this ambiguity. On the one hand, it destroyed the position of the German settlers—who were only slowly replaced by a new influx of British and other settlers into Tanganyika. On the other hand, the British settlers to the north used the leverage that their government's need of them during the war provided to establish their claims to elected representation on the legislative council and to seats on the governor's executive. Their advance to power was symbolized by the change in name of the territory in 1920 from the East Africa Protectorate to Kenya Colony.

Two subsequent occurrences were of major importance. First, in 1921 the collapse of the world primary produce market ruined the Uganda planters. (African peasant production, however, survived and with the addition of coffee to the existing cotton crop thenceforth provided the unrivalled base for the Uganda economy.) Secondly, an attempt by the Kenya settlers in the early 1920s to secure complete control of their country's internal affairs was sharply rebuffed. Essentially because they became embroiled in a dispute about the political rights of the Asian population in Kenya—the result of a larger migration paralleling theirs—they faced the opposition both of the British government of India and of various philanthropic bodies in Great Britain. Thanks to missionary intervention, the Devonshire White Paper of 1923, which sought to resolve the dispute, declared that African interests would thenceforth be paramount in Kenya. Although this turned out to be little more than a verbal smoke screen, the White Paper also stated, "His Majesty's Government cannot but regard the grant of responsible self-government [for the British settlers, that is] as out of the question within any period of time which need now be taken into consideration," and it said as well, "It will be necessary to see how matters develop, especially in regard to African representation." These statements were in fact adhered to for 40 years. But they did not stop the settlers from having a preponderant influence over the government of Kenya, an influence they regularly employed to secure for themselves important economic advantages, for example, in the construction of railway sidelines. Even in Uganda, moreover, European and Asian monopolies managed to control the processing and marketing, if not the actual production, of the country's cash crops (very largely thanks to the Uganda government's passion for tidiness), while in Tanganyika the small body of British settlers enjoyed a position after 1921 that those in Uganda had previously possessed.

Through a series of commissions—Ormsby-Gore (1924), Hilton Young (1927), Wilson (1929), and finally a Joint Select Committee of both houses of the British Parliament in 1930–31—the prevailing uncertainty was aggravated by a concerted attempt of the British government to create an East African federation. Although the Kenya settlers were doubtful of its effects, Africans and Europeans in both Tanganyika and Uganda alike were certain that it would lead to settler rule from Nairobi and steadily opposed it—none more so than the Kabaka of Buganda and the governor of Tanganyika, Donald Cameron. Thanks to the help of a liberal lobby in Great Britain, the proposal to federate was eventually shelved. But in 1935 the ambiguity in eastern Africa was then sharpened once again when the Fascist armies of Italy attacked the African kingdom of Ethiopia, and Italian activity in Somaliland was simultaneously increased.

THE COLONIAL ECONOMIES

20th-century economic needs

At the beginning of the 20th century the economies of East Africa were at once simple and poor, needing as a first step to improvement the creation of a cash economy. The next necessary step was to link the cash economy with the world market, since only in this way could any economic return provoke economic production beyond the existing subsistence, handicraft, and hunting stage. Real disadvantages were inherent in dependence on an outside market. In 1929, for example, sisal dropped from £40 a ton to £12, while in 1930 coffee dropped from £120 a ton to £70. Enormous advantages, nevertheless, flowed from the higher prices that were paid during the 1920s and during the great boom of the 1940s and early 1950s. The aliens, into whose hands the levers of economic power had fallen, brought to East Africa, moreover, new and essential economic skills. The Asians, particularly, took risks in expanding retail trade into the remotest parts of East Africa and were often vital to its growing economic networks.

Viable exports were the key to any kind of development. Moreover, their growth was central to the expansion of government revenues, without which improved provision for health, education, and transport would have been impossible. In Zanzibar cloves were the single most important source of weath and revenue; in Uganda cotton made up from 55 to 85 percent of the country's exports between the 1910s and the 1950s, by which time coffee was beginning to play a large part. Total export earnings in Uganda rose from £1,500,000 in 1921 to £47,400,000 in 1951, while government revenues over the same period climbed from £800,000 to £22,300,000. Though Tanganyika did not share in the prosperity of the 1920s and was badly hit by the 1930s slump, its total exports rose from £1,100,000 in 1921 to £40,500,000 by 1951. During that time, however, government revenues increased from £900,000 to only £11,900,000. Kenya's position, meanwhile, was different yet again. Though exports of £1,800,000 in 1921 grew to only £24,100,000 in 1951, government revenue climbed from £1,300,000 to £23,400,000 over the same period. Kenya's sizable deficits on external trade were partly offset by invisible earnings accruing from Nairobi's position as East Africa's main commercial and industrial centre and by Kenya's surplus on interterritorial trade. In Italian Somaliland the budget rose from a mere 2,500,000 lire in 1908 until it averaged 74,000,000 lire in the early 1930s. Local revenue, however, never approached expenditure and, indeed, produced considerably less than half the colony's budget. Exports, mainly of cotton, bananas, hides, and skins, gradually increased, but the adverse balance of trade was always marked; it reached a peak of 131,000,000 lire in 1929 and was still 29,000,000 lire in 1934. The position was just as bad in British Somaliland, to the north. There in the late 1930s imports were twice the value of exports, while as late as 1937 the colony's budget was still only £213,139.

The mineral finds in East Africa were in no way comparable to those of West, Central, and South Africa. There was a short-lived gold rush to Kakamega, in Kenya, in the 1930s and a spectacular find of diamonds near Shinyanga in the early 1940s, which in due course contributed significantly to the Tanganyikan budget; and in the 1950s copper began to be mined at Kilembe, in western Uganda.

Improvements in transportation

Railways were extended between the wars. The Uganda railway, hitherto running only through Kenya, reached Kampala, the main commercial centre of Uganda, in 1931 (it was not taken on into western Uganda until the 1950s). The central Tanganyikan line was meanwhile extended to Mwanza, at the south end of Lake Victoria, following a controversial British government decision of 1924. The advent of motor trucks, more especially after 1920, was no less revolutionary in its effect in the rural areas, though the building of a network of macadam roads in Kenya, Uganda, and Tanganyika had really to wait until the 1950s. Few novelties were more important than the bicycle. This was usually the first item an African would buy with any surplus cash he had.

Subsistence was, of course, the major concern of most Africans throughout this period, as it was of their colonial administrators. Concerned, as they were, by the threat of famine, they often insisted that particular "famine" crops should be grown. Large government food stores were built; and, later, soil erosion caused much concern. Throughout, the departments of agriculture of the colonial administrations were of prime importance in guiding the production of existing crops and developing new ones. Their research stations wrestled with crop diseases and produced new crop strains; and, especially after 1930, the scientific capability (if not yet the economic knowledge) of the colonial agricultural services reached a very high standard.

In Kenya's Nyanza Province cotton production was tried, but without success. Nor did it ever make quite the headway in Tanganyika's West Lake Province that it did in Uganda, to the north. African coffee cooperatives flourished on Mounts Kilimanjaro and Elgon; but African coffee production in the Mt. Kenya region was held back by settler opposition. Sugar production began in Uganda in 1924 and under Indian management successfully supplied the local market. The discovery in 1935 that the pyrethrum daisy yielded the most potent insecticide yet known stimulated its growth in the Kenyan highlands. Rubber was tried but never brought much return, nor did flax. A groundnut (peanut) scheme in Tanganyika that the British government sponsored in the years after World War II was a spectacular failure.

Because coffee played such a large part in European production, the settlers in East Africa had the particular distinction of being at one and the same time substantial capitalist employers of labour and resident farmers. Yet because Africans almost always possessed their own means of production, they never quite became a proletariat similar to the working classes of Europe. But in the decades immediately preceding World War II, most African societies, once they had made the initial adaptation to a money economy—to which they were impelled by the imposition of taxation and simultaneously attracted by a variety of new needs—were making very little progress. Although there were some signs—for example, in Kikuyu country—of an increase in economic activity, it was not until the 1940s and 1950s that a new economic momentum gathered force.

OPENING A BREACH: SOMALIA

World War II served to mark the beginning of the end of European rule and European predominance in East Africa in two respects. First, relations between the colonial rulers and a new African elitist generation worsened. In Kenya, for example, leaders of several local proto-nationalist political associations were imprisoned in 1940, while in Uganda the government's stringent handling of the wartime economy alienated many educated Africans. Second, the defeat by the British and their allies of the Italians in the horn of Africa and the restoration of Emperor Haile Selassie to Ethiopia in 1941 raised the question of the future of Somaliland. In the event, Somalia became an Italian trust territory in 1950—on the important condition that it should become independent in ten years.

Preparations for independence

This unusual requirement was faithfully carried out. Under its pressure the Italian administration set about preplanning the transfer of power. It opened a School of Politics and Administration at Mogadishu in 1950; in 1952 it instituted an imaginative five-year educational development plan; and by 1956 all districts and provinces were under the direct charge of Somali administrative officers. Municipal elections had already been held, and district councils were set up. A Somali Youth Club, formed in 1943, had by 1947 become the 25,000-strong Somali Youth League. Standing for election to the Legislative Assembly of (Italian) Somalia in 1956, at which time the assembly was granted full powers over the internal affairs of its country, it won 43 of the 60 open seats, and its candidate, Abdillahi Ise, became Somalia's first prime minister. Despite much political dispute in the years that followed, Abdillahi's government survived to take Somalia into independence on July 1, 1960.

These developments exercised very little influence over events in British Somaliland in the north, where British control had been restored after the short-lived Italian conquest in 1940. But eventually elections to the territory's legislative council were held in February 1960, and, as the victorious Somali National League declared its intention of joining hands with Somalia forthwith, the British granted independence to their territory three days before the Italians did to theirs.

THE CRISES OF THE 1950S

So shortsighted a policy was, however, only part of a general British reluctance to prepare for independence in East Africa. After World War II, in lieu of an East African federation, the British created the East African High Commission to oversee various intraterritorial services as well as the common market that had grown up between the three main territories. The still-persisting uncertainty in the evolving political order was then embodied in the doctrine of "partnership" between Africans, Europeans, and Asians in East Africa, which at the instance of the British government found numerical formulas for their representation in the legislative and executive bodies of its territories. Objections to these formulas from the European minorities were soon overtaken by the objections of Africans, who saw in the contradictory relationship of representation to population ratios a device whereby the preponderant influence of the European community over government could be continued.

By the late 1940s many Africans in Kenya, and more particularly the Kikuyu, had become profoundly alarmed about their future political prospects. Long aggrieved by the loss of some of their lands to settlers in the first decades of the century, they remembered the armed repression of a political gathering in Nairobi and the exile in 1922 of its leader, Harry Thuku, and recalled the subsequent suppression of various other such associations. When in 1944 the Kenya African Union (of which Jomo Kenyatta became president in 1946) was formed, it was soon made clear too that the possibility of their having any share in the government of their country was for the time being almost endlessly remote.

The Kikuyu revolt

By 1950 the Kikuyu were heading for revolt, and oaths were being pledged to the movement known as Mau Mau. Outbreaks of violence in 1952 caused the new governor, Evelyn Baring, to declare an emergency and to request that British troops be flown in. In some respects the violence that occurred took the form of a civil war between the "haves" and the "have nots" among the Kikuyu themselves. Most of the fighting took place in Kenya's forests, and it was not until 1959 that the emergency was lifted. Although some Africans had by then been allowed to vote for their own legislative representatives, the old regime held on against increasing strains. Some Europeans were beginning to think of compromising with the Afri-

can politicians, while others merely hardened in their attitudes amid them. New conflicts were at the same time racking the African elitist groups, and the older chiefs no longer seemed to be effective political intermediaries. From 1954 onward the Swynnerton Plan committed Kenya's Department of Agriculture to a more determined development of African agriculture and began to give new shape to the Kenyan economy.

In neighbouring Uganda the 1950s opened with government declarations in support of the dubiously formulated partnership doctrine, and in 1953 a sudden fear that an East African federation was once again being contemplated touched off a major crisis in the central kingdom of Buganda that led to the deportation by the British of its *kabaka*, Mutesa II. In an effort to settle the ensuing furore, the British reverted to their earlier declarations that Uganda was "primarily an African country"—and in a crucial step for the whole of East Africa broke through, in the Namirembe Agreement of 1955, the existing formula for a "balanced" racial representation in its legislative and executive councils and gave to Africans a handsome majority.

The focus then shifted to Tanganyika, which had become the chief testing ground for the partnership doctrine. There, under the leadership of Julius Nyerere, the Tanganyika African Association (originally set up in 1929) was transformed in 1954 into the Tanganyika African National Union (TANU). Carried along on a wave of discontent at increased European control of agricultural development (much of which may have been well intentioned) and helped by the absence of sharp tribal rivalries, it contested and won the multiracial elections of 1958, thus opening the way to African majority rule.

THE APPROACH OF INDEPENDENCE

Although Tanganyika was the least well endowed economically and educationally of the four British East African territories, its nationalist movement was organizationally united to a degree that was unique in East Africa, and it was the first to win independence, in 1961.

Economically, Uganda had prospered during the mid-1950s largely because of a coffee boom. As a result, educational and economic progress was marked—the Owen Falls hydroelectric dam was opened, for example, in 1954. But Uganda's nationalist movement was never united, largely because, as early as 1954, it had become clear that independence was approaching before too long, so that the main issue was not how to attain independence but how power should be distributed upon its attainment. Accordingly, during the 1950s neither the Uganda National Congress (founded in 1952) nor the (largely Protestant) Progressive Party (1955) nor the (Roman Catholic) Democratic Party (1956) nor the (breakaway) United Congress party made much headway. The main split was between the Baganda and the other peoples of Uganda (who in combination formed a majority). By 1960 the Baganda were represented by Kabaka Yekka (KY; The King Only) party, while the latter was represented by the Uganda People's Congress (UPC), led by Milton Obote. A surprising victory by the Democratic Party on a minority vote in 1961, and its formation of the first African government of Uganda, only precipitated a political deal between KY and the UPC, which, following another election in April 1962, led to their coalition in a new government under Milton Obote. Uganda achieved independence under this government on October 9, 1962.

Proliferation of parties in Uganda

Zanzibar meanwhile was somewhat similarly confused, first by rivalry between the (essentially African) Afro-Shirazi Party and the (largely Arab-led) Zanzibar National Party (ZNP) and then by the emergence of the Zanzibar and Pemba People's Party (ZPPP) and of the Ummah Party. Elections took place in July 1957 and in January and again in June 1961. After these last elections a ZPPP–ZNP coalition government was formed. Though the two parties had won only a minority of the votes, they remained in office despite another equally unsatisfactory election in July 1963 and took the country to independence on December 10, 1963.

Kenya, meanwhile, though benefitting from the highly successful handling of the economy during Baring's governorship, was first stopped by its colonial government in the aftermath of Mau Mau from having country-wide political parties and then saw two established in 1960 —the Kenya African National Union (KANU) and the Kenya African Democratic Union (KADU). The former represented mainly the Kikuyu and the Luo, the latter most of the smaller tribes. KADU formed a government with non-African support following the general elections of 1961, then formed a KADU–KANU coalition in 1962. KADU was decisively defeated by KANU (with Kenyatta at its head) in the elections of May 1963. Independence for Kenya followed on December 12, 1963.

IV. Independence and after

THE BEGINNINGS

Demography and language

As independence approached, there were just over 25,000,000 people in the four former British territories of East Africa and 2,500,000 in the newly constituted Somali Republic. Tanganyika, with almost 364,000 square miles (942,000 square kilometres), had 8,700,000 inhabitants (most of them clustered in the south and southwest); Uganda, with 91,000 square miles (236,000 square kilometres), had 7,000,000; Kenya, with 225,000 square miles (583,000 square kilometres), had 9,000,000 (clustered in the south and southwest); and the islands of Zanzibar and Pemba combined had a mere 300,000. Only 2 percent of the total population was non-African; there were about 100,000 Europeans and about 330,000 Asians. English was the official language in East Africa, Italian of the Somali Republic. Tanganyika, however, steadily pressing for Swahili to be given pride of place, had 70 percent of its newspaper circulation in that language (as compared to about 40 percent in Kenya). Half of Uganda's daily newspaper circulation continued to be in English, though at the core of the country there was a large newspaper circulation in Luganda as well. The gross domestic product for the four territories was estimated to be £557,000,000 in 1961, of which £302,000,000—£155,000,000 of it of a non-monetary kind—came from a variety of primary industries. This provided a very low per capita income, which varied from £20 a year in Tanganyika to £31 in Kenya (1961 figures). Only just over 4 percent of the African population was in salaried employment, and, although 50 percent were under 16, of these only 2,000,000 were having any kind of education. Makerere College in Uganda, founded in 1921, had become a university college by 1949 and by 1963 had produced 20 of the 49 cabinet ministers in the three larger East African countries, but Nairobi's Royal Technical College did not become a university college until 1958, and the University College of Dar es Salaam was not founded until 1962. At independence, therefore, though the East African elite was markedly smaller than that of West Africa at a comparable time, it was significantly larger than that of either the Congo or the Central African countries.

THE SOMALI REPUBLIC

At independence in 1960, though the politics of the Somali Republic was still conditioned by tribal allegiances, its first major problems arose from the last-minute marriage between the former Italian trust territory and the former British protectorate. Urgent improvements in communications between the two areas were necessary, as were a number of readjustments in their legal and judicial systems. Administrative and bureaucratic structures had to be realigned and delicate political arrangements made. The first independent government was formed by a coalition of the Somali Youth League (SYL) and the northern-based Somali National League (SNL). Abdi Rashid Shermarke (from the south) became the prime minister; Mohammed Ibrahim Egal, who had briefly been prime minister of British Somaliland, became minister of education; while in July 1961 Aden Abdullah Osman (a long-standing SYL figure) was confirmed as first president of the republic. Even so, discontent fermented in the north, expressing itself in December 1961 in an abortive coup d'etat by British-trained

officers. Potentially more serious was the resignation in September 1962 of Egal. But he then participated in forming the Somali National Congress, and thereafter in all the more prominent alignments, effective cross-linkages between the politicians of the two former territories were formed, and it soon became clear that the merger had been successfully accomplished.

In the first postindependence elections, on March 30, 1964, the SYL captured 69 of the national assembly's 123 seats and increased its strength to 90 thereafter. When, at the instance of Osman, Shermarke was replaced as prime minister by his former close associate Abdirizak Haji Hussein, a six months' political crisis ensued, which was concluded only with the enlargement of the Cabinet that Abdirizak had originally put together.

Somali realignment

For the Somali Republic the crisis came at a peculiarly inopportune moment. Late in July 1964 at a meeting in Cairo the member states of the Organization of African Unity (OAU) passed a resolution, in the face of open Somali opposition, whereby they pledged to respect the frontiers existing on the achievement of independence. Ever since the two Somali territories had been linked together at independence, the new republic's prime concern had been to secure the accession to it of French Somaliland and of the adjacent areas of Ethiopia and Kenya in which compact Somali populations lived. In 1962, moreover, the British government had undertaken to decide the future of the Somali areas of Kenya before internal self-government was established there, and in October 1962 it had even appointed a commission to ascertain public opinion in the area. Though an overwhelming Somali majority favoured union with the Somali Republic, the British, anxious to retain the goodwill of the African government of Kenya, which was opposed to the transfer, refused to authorize the decision. Accordingly, the Somali Republic broke off relations with the British in March 1963, and by the end of the year it had turned to the People's Republic of China for £7,000,000 of aid and to the Soviet Union for £11,000,000 of military assistance. Guerrilla operations in Kenya's northern frontier district soon followed.

The conflict with Ethiopia, meanwhile, was even more serious. The mobile Somali population in northeastern Ethiopia had long been at odds with the Ethiopian government, and at the end of 1963 the Somali Ogaden tribe came out in open revolt against it. Fighting also broke out between Somali and Ethiopian regular soldiers and was ended only when OAU intervention and Sudanese mediation secured the Khartoum Agreement of March 30, 1964. Local conflict was nevertheless endemic, periodically giving rise to such episodes as that at Webi Shebeli in which three Frenchmen were killed in February 1968. The situation was aggravated by the support that the Russians gave the Somalis and the U.S. to the Ethiopians. Despite animosity over French Somaliland (now the French Territory of Afars and Issas) half of whose population was Somali, 60 percent voted for the maintenance of French rule on March 19, 1967.

In June 1967 Shermarke defeated Osman for the presidency and in July nominated Egal as prime minister. This change of government soon led to a change in the "Greater Somali" policy, which had so far proved to be ineffectual. Violence was now eschewed, and the right of Somalis everywhere to self-determination was stressed. At an OAU meeting in the former Belgian Congo (now Zaire) in September 1967 Egal laid the basis for a friendly meeting in Addis Ababa with the Ethiopians later in the month. On October 28, 1967, following some mediation by the president of Zambia, Kenneth Kaunda, Egal and President Kenyatta of Kenya signed the Arusha Understanding. Ambassadors were exchanged (and a British ambassador returned to Mogadishu). In July 1968 Shirmarke and Egal were received in state in Nairobi, and the state of emergency in northeastern Kenya was ended. In September 1968 Egal was welcomed in Paris by the French president Charles de Gaulle, and further meetings in 1967 and 1968 took place between Egal and Haile Selassie, the emperor of Ethiopia. Clashes on the ground continued, however, and despite Egal's success in the national assembly there were always those who openly questioned his policies.

But under his leadership, the SYL won 73 of the 121 seats at the general elections of March 26, 1969, and in the aftermath Egal secured almost unanimous support for his government. There were now, however, strong feelings that malpractices during the election had been excessive, and the triumvirate that ruled the country—Shermarke, Egal, and the minister of the interior, Yasin Nur—was obviously becoming very unpopular. Shermarke was assassinated in October 1969 (apparently as a result of a family quarrel), and the denouement came one week later, on October 21, when the army led by Maj. Gen. Mohammed Siyad staged a coup d'etat. A 25-strong Revolutionary Council of officers and Cabinet of young civilians, which were in consequence set up, reorganized district administration, expelled the U.S. Peace Corps, and increasingly sought Soviet support. Egal's initiatives toward a Greater Somalia were shelved.

Economic difficulties

One major reason for the military takeover was a strong feeling that there was all too little to show for the $200,000,000 of foreign aid that had been received since independence. There had been a five-year development plan in 1963–68 and a short-term plan for 1968–70. Though a deep-sea port had been constructed at Berbera with Russian money and deep-sea berths at Kismayu with American money, there had been failures in wheat, cotton, and oilseeds developments. Foreign disbursements in aid of revenue were still required, and the only considerable exports were sheep, goats, and bananas. Nor did the export of these commodities proceed according to plan. Though a Livestock Development Agency had been established in 1966 and a meat-processing plant had been built at Kismayu, the sheep and goat trade, which was largely in private hands and most of whose produce went through Berbera to Arabia and the Persian Gulf, was threatened by the consequences of overgrazing and erosion in the herding areas. Moreover, the Banana Marketing Board was experiencing increased difficulty with its sales of bananas to Italy, for, after the closure of the Suez Canal in 1967, stocks had to be conveyed via the Cape and tended to spoil. From an early stage the republic had been an associate member of the European Economic Community (EEC) and there were protracted negotiations over the possibility that it might join the East African Community, but by 1972 no such link had been effected.

ECONOMIC COOPERATION

Close links, especially economic ones, among the former British territories in East Africa were maintained after independence. In 1961 a commission, led by Sir Jeremy Raisman, had produced an acceptable formula for a more equitable fiscal redistribution among the three mainland territories: the East African High Commission was re-formed to become the East African Common Services Organisation (EACSO), and the pre-existing common currency, common market, and customs union that they shared (with some participation by Zanzibar) was maintained. It seemed for a time, indeed, as if the links among the three territories would be considerably reinforced by the growing association among their African leaders. This had been inaugurated in the Pan-African Freedom Movement for East and Central Africa, which first met at Mwanza in Tanganyika in 1958, and had before very long provoked Julius Nyerere (of Tanzania) and Tom Mboya (of Kenya) to talk of that formerly forbidden possibility—an East African federation. The climax came when, following a euphoric OAU meeting in Addis Ababa in 1963, Kenyatta, Obote, and Nyerere issued, on June 5, 1963, a declaration of intent to create an East African federation before the end of the year. Since no proper investigation of its implications had, however, been made, the project soon collapsed. Links among the countries were further strained as Tanzania (as it became in 1964) worried increasingly about their economic advantage to Kenya, and had it not been for the signing of the Kampala Agreement of 1964, which was modified at Mbale in 1965 and which complemented Raisman's fis-

cal equalization by an arrangement to equalize the distribution of industry, the common market could very well have collapsed.

The agreement, however, was never put into operation, since Kenya refused to ratify it. In 1965 each country decided to establish its own central bank and its own currency; and when, in 1966, Kenya's share of interterritorial exports climbed to an unprecedented two-thirds, Tanzania announced that it would place severe restrictions upon its imports from there. The three governments agreed, however, to establish a ministerial commission, under an independent Danish economist, Kjeld Philip, and after considerable negotiation signed in Kampala on June 6, 1967, a new Treaty of East African Cooperation, by which, on December 1, 1967, they replaced EACSO with the East African Community. Unlike its predecessors, its headquarters were to be in Arusha, Tanzania, rather than in Nairobi, while the headquarters of the various common services were to be decentralized. An East African Development Bank was set up, and, in an effort to check imbalances in interterritorial trade, arrangements for a transfer tax were instituted. The customs union and the common market accordingly survived. Considerable strains were nevertheless imposed upon it in 1971 both by the refusal of Nyerere to recognize the military regime that came to power in Uganda and by the differential responses to the world monetary crisis that followed upon the collapse of the Bretton Woods system in the West.

Though there was much talk of Somalia, Rwanda, and Burundi joining the community and of Zambia becoming an associate member, the community was still, despite protracted negotiations, confined to its original three partners. There had been an important event in 1963, when the Kenya, Uganda, and northern Tanganyikan railway system had eventually been linked to the central Tanganyikan line. Following Rhodesia's unilateral declaration of independence in 1965, a pipeline, moreover, was built between Tanzania and Zambia, and after lengthy negotiations major assistance was procured from the Chinese government for a "Tanzam" railway. Much could be said for establishing closer connections between the East African Community and its like-minded neighbours. The difficulty was to see how newcomers could be associated with a system that, uniquely for Africa, had been in operation for over half a century.

Tentative economic moves toward the West

During these years the East African governments had also to decide their relationship with the EEC. Originally in 1962 they had decided that they would not apply for associate membership of the kind possessed by the West African countries. Since association with the EEC did nevertheless promise obvious advantages, negotiations were set in motion; but they broke down, both in March 1965 and more particularly in November 1966, over the EEC's insistence upon reciprocal rights. Following East African concessions, however, an agreement of association was eventually signed at Arusha on July 26, 1968, whereby East African goods were to be admitted duty free into Europe in exchange for 59 types of European goods being admitted duty free into East Africa.

THE ZANZIBAR REVOLUTION AND THE TANGANYIKA MUTINY

Upon these and related matters the political leaders of East Africa regularly conferred. Interests such as these were always, however, secondary to their main preoccupations, which were primarily concerned with their countries' internal affairs.

Within two months of Tanganyika's independence, Nyerere had dramatically resigned its prime ministership (to which Rashidi Kawawa, the vice president of TANU, temporarily succeeded) so that he might weld his party more firmly to his support. In November 1962, following the declaration of a republic, he was elected to the presidency by 97 percent of the votes cast. But these were only the first steps of a program designed to consolidate the fragile authority of government at the national level. Subsequently, the first legislative acts of the new republican government turned leading TANU officials into area and regional commissioners. (As such, they succeeded to the positions formerly held by colonial district and provincial commissioners.) In 1963 the chiefs were abolished, with the most acceptable among them becoming executive officers to local councils. In view of Tanganyika's shortage of manpower, its considerable size, and the fluctuations that had previously beset its political structures, these were viewed as shrewd measures.

Despite such precautions, Tanganyika was rocked, before long, by two important occurrences. First, on January 12, 1964, the racial and political crisis in Zanzibar erupted into violent revolution, precipitated, it seems, by the attempt of the ZNP government to purge the Zanzibar police of those whom it believed untrustworthy. The outbreak led to countless killings—particularly of Arabs. Those of any prominence who survived were swiftly imprisoned for indefinite periods. Though the Sultan, who had only just succeeded his father—himself only very recently the successor to Sultan Khalifa—escaped to England, his palace on the waterfront at Zanzibar now became the Peoples' Palace. A Revolutionary Council was set up, and in a highly symbolic gesture the entrée into the previously highly exclusive Arab club was thrown open, and the land—so much of it formerly owned by the Arab ruling class—was nationalized. In this way the African and local Afro-Shirazi majority vigorously overthrew the supremacy of the long-established Arab regime.

Zanzibar revolutionaries

At the outset, the exotic figure of so-called Field Marshal John Okello (a wanderer from Uganda) dominated the scene, but it was not long before he was brushed aside by two more permanent figures. One was the dockworker leader of the Afro-Shirazi Party, Abeid Karume; the other was the colourful figure of Mohammed Babu, leader of the Ummah Party, an Arab by birth and a Communist and revolutionary by conviction. Between them they quickly made Westerners unwelcome in Zanzibar and secured support both from the East Germans and from the Chinese for the development of the islands. A people's republic was inaugurated, and Zanzibar appeared to have become a Communist outpost.

The initial eruption was portentous for Nyerere. He had, as it happened, long-standing connections with Karume, whom he had encouraged and supported since the mid-1950s; and very soon after the outbreak Tanganyikan police were welcomed by the revolutionary regime in the island. Their return a few days later coincided, however, with a second event that unhinged the government of Tanganyika in January 1964.

On January 19, 1964, aroused no doubt by the Zanzibari example, a mutiny occurred at the Colito barracks, outside Dar es Salaam. Led by Hingo Ilogi, the mutineers locked their British and African officers in the guardroom, seized a number of government buildings, and rampaged through the town. Sympathetic outbreaks occurred at two other military stations upcountry. Nyerere went into hiding, and the Tanganyikan government came to a halt. The minister of defense, Oscar Kambona, managed, however, to make contact with the mutineers, and two days later the President emerged from hiding but was then put to the humiliating experience of seeking British help. On January 25 a Royal Marine Commando force was landed, and the mutineers were speedily rounded up. The position of the existing government was therefore fairly readily restored.

Nyerere next turned his attention to the new regime in Zanzibar, boldly effecting an agreement with its Revolutionary Council in April 1964 whereby the new United Republic of Tanganyika and Zanzibar (later renamed Tanzania) was set up. He remained president of the new republic, Karume first vice president, and Kawawa second vice president. Several Zanzibari leaders were appointed to the union Cabinet, and their Revolutionary Council was empowered to appoint 36 out of a total of 189 members of the national assembly upon the mainland. Although control over Zanzibar itself remained in the Revolutionary Council's hands, Tanganyika felt that it nevertheless possessed some check upon the potential dangers that its offshore revolutionary island presented.

Internally, Zanzibar went very much its own way. By 1970 it had received $12,000,000 of Chinese aid, had

welcomed a 400-strong Chinese mission, and had had its 3,000-man army trained by them. But, though Karume called the Chinese his best friends, he did not allow them to dominate him. He permitted 120 East Germans on the island and accepted a United States donation of over $1,000,000 for a technical school. Despite staff shortages, 80 percent of the children were receiving free primary education. Karume himself took an interest in a housing program. There were important developments in the improvement of water supplies, and as a consequence of the land distribution each cultivator was entitled to a three-acre plot. Since the hiring of labour was prohibited, there were problems at the picking seasons; but in 1969 the value of clove production rose to TShs. 80,000,000 (from TShs. 40,000,000 in 1968).

The problems on Pemba

Even so, there was much discontent on Pemba, the northern of the two islands. People's courts were set up, and in 1967 an alleged plot against the regime led to some key Zanzibaris being handed over to the Revolutionary Council by the mainland authorities and to the summary execution of four of them. In 1970 there were also stories of the forced marriages of young Arab and Persian girls to the Zanzibari leaders. Such occurrences were an embarrassment to Nyerere. They gave grounds to his detractors' attacks and, for all its unusual completeness, placed the Zanzibari revolution low on the list of models to which the world's revolutionaries now looked.

Then in April 1972 Karume was assassinated. Abeid Jumbe, however, succeeded him, and there were fewer changes in the revolutionary regime than might have been expected.

Both the original revolution and the mutiny were rather more than embarrassing upheavals in Nyerere's orbit: they underlined the tenuousness of political authority in East Africa and the latent vehemence of those who were excluded from the privileges enjoyed by its ruling elites. Nyerere himself had long been an idealist—for example, in his support for interracial harmony at a time when this was far from popular; but he had long been a pragmatist too—as his association with the last British governor, his resignation as prime minister, and his union with Zanzibar had shown. In the heat of the events of 1964 his idealism and his pragmatism fused. Elitism now not only seemed inappropriate to Zanzibar but, on the impoverished mainland, unwarranted and dangerous as well.

UJAMAA IN TANZANIA

In the latter part of 1964, therefore, he arranged that TANU members in groups of ten houses should form cells and that their leaders should be required to act as agents for the superior units of state and party. (There was some conflict with the village development committees that were instituted in 1963, until it was decided that these should be composed of cell leaders.) The presidential commission that was set up in January 1964 recommended that no similar fusion of party and state should take place. Tanzania, however, became a one-party state and at the general elections in September 1965 instituted an ingenious arrangement by which two candidates were chosen by TANU for each of the contested parliamentary seats. Nyerere was once more elected president, but a number of sitting members lost their seats to other TANU candidates, who, it was hoped, would be more careful of their constituents' interests. In 1966 a civil form of compulsory public service was introduced. Protests by university students at being required to serve with others of their generation were firmly rebuffed, and checks upon the pretensions of the Tanzanian elite were initiated.

The Arusha Declaration

The Arusha Declaration of February 1967, which Nyerere drew up, declared that TANU would henceforth establish a democratic Socialist government of the people and would see that the government exercised effective control over the principal means of production and would pursue policies that would facilitate collective ownership of the country's resources. More specifically, it laid down that no civil servant was to own more than the house in which he lived or own shares of stock or have an income from more than one job. The banks were nationalized immediately after, and majority control over a number of foreign concerns was instituted. There were negotiations over compensation—some of them protracted—but banking operations were taken over by the national bank of commerce. Nyerere issued a further pronouncement in July 1967 in which he called for a new attitude within the educational system that would be more appropriate to the self-help and egalitarianism that were the order of the day. He confirmed his interest in Ujamaa (Familyhood) Socialism, as he called it, by increasing the development of *ujamaa* villages and by encouraging the national development corporation that had been set up in 1965 to initiate large-scale Socialist production. Perhaps the key statement of Nyerere's policy appeared in a pamphlet on Socialism and Rural Development in which he proposed that the people of Tanzania should move from being a nation of individual peasant producers to a nation of *ujamaa* villagers where the people would cooperate directly in small groups that would cooperate together for joint enterprises. By 1969 it was reckoned that 10 percent of the population were to be found in *ujamaa* villages. In a presidential circular of March 20, 1969, government officials were instructed to encourage the "growth of cooperative production, cooperative marketing and distribution and communal rather than private expenditure patterns," to emphasize the "advantage of living together and working for the good of all," and to discourage "the continuance of private individual farming"; and in presenting the second five-year development plan (which was to run on from 1969) Nyerere announced that not only was there to be still more emphasis upon rural development but also that *ujamaa* villages were to be established throughout the country.

The first five-year plan

The first five-year plan (1964–69) had shown some success. In 1964, 15 percent of the wealth produced in the country was being reinvested; by 1967 the figure was 23 percent. Of the total investment, moreover, 70 percent had been from within the country—and teacher enrollments had nearly quadrupled. But industry still accounted for only 7 percent of the gross domestic product —less than had been hoped. Sisal prices, on which a great deal had depended, dropped catastrophically after 1965; and the census showed that the population was growing at a much faster rate than had hitherto been thought. Nyerere himself estimated that there were proportionately fewer children going to primary school than in the past—fewer, indeed, than 50 percent. In these circumstances he called for hardheaded decisions to be made and, while facing the fact that primary education for all could not be attained before 1989, insisted that for the good of the nation it would be best during the period of the second plan to give one child seven years' education rather than two of them four.

Though Nyerere carried most of his countrymen with him, he lost the support of Oscar Kambona (the former minister of defense and secretary general of TANU), who fled the country in 1967. In 1968 he clashed with the national assembly, from which he expelled six members. A year later a conspiracy against him was uncovered, and in February 1970 those concerned were sentenced to long prison terms. Nevertheless, at the ensuing general elections on October 30, 1970, Nyerere was re-elected with 95 percent of the votes in a 72.3 percent pool, while more members of parliament held their seats under the new electoral arrangements than in 1965. In 1971 the new assembly provided for the compulsory acquisition of all rented buildings worth over TShs. 100,000.

Between 1965 and 1970 Nyerere had been his own foreign minister. He had retained Tanzania's membership in the Commonwealth but after 1964 had sought to adopt a middle, nonaligned position in international affairs and in particular had exchanged visits with leaders of the Chinese People's Republic. Though between December 1965 and July 1968 he broke off diplomatic relations with Britain because of his hostility toward Britain's Rhodesian policy, with equal impartiality he vehemently denounced the Russian invasion of Czechoslovakia in 1968. Within the OAU, Tanzania played a leading part in the campaign against the white-dominated countries to the south. Dar es Salaam became a base for various exiled

leaders of the South African National Congress, the Zimbabwe African National Union, the South West Africa People's Organisation, and the National Liberation Front for Mozambique (FRELIMO). At the same time, Tanzania provided a refuge for exiles from Malawi, and, despite some official meetings in 1969, differences with Hastings Banda, president of Malawi, remained considerable. There were differences, too, with Nigeria when, on April 13, 1968, Tanzania accorded recognition to the breakaway state of Biafra. There were also differences with Uganda when, in 1971, after a military revolution against Milton Obote, Nyerere at first refused to recognize the succession of Gen. Idi Amin.

POLITICAL UNREST IN UGANDA

Amin's coup brought to an end the nine-year primacy in Uganda of Milton Obote. The political problems Obote had faced were far more serious than any that had confronted Nyerere, and they precluded the effective elaboration of the Socialist doctrine he professed.

Although Uganda was still much underdeveloped, its economic position was better than Tanzania's; in 1966, for example, Uganda's population was estimated at 7,740,000 and Tanzania's at 10,717,000; but Uganda's growth rate was reckoned at 5.6 percent per annum, while Tanzania's was 4.2 percent. The consequences can be seen in the statistics on education: in 1966 Uganda had 633,000 children in primary and junior secondary schools; Tanzania had only 100,000 more.

In 1968, 77 percent of Uganda's exports came from its two long-standing primary products—coffee and cotton. Production of coffee was limited by the world quota system, which in 1967 allocated to Uganda 167,000 tons. Even so, it was now the country's highest foreign-exchange earner. Cotton production was also doing well. In 1966 a new record of 445,000 bales was produced. Tea production was increasing, too. In 1967, 24,000,000 pounds were produced; 30,000,000 pounds in 1968; and 39,000,000 in 1969. Uganda, in fact, now had a higher tea yield per acre than any other country except Japan. Tobacco and sugar production were also going up, and tourism was on the way to being the country's third-largest foreign-exchange earner. Industry was expanding —at twice the rate of the economy as a whole. In 1968 a contract was signed by which Japan undertook to take all Uganda's copper for the next five years.

Economic changes

The economic organization of the coffee and cotton crops had by now set a pattern for the Uganda economy as a whole. Production was in the hands of its peasant producers—some large, some small, many employing labour and, if not owning their own land, having relatively easy access to it. Processing was increasingly in the hands of farmers' cooperatives, while marketing was the responsibility of coffee and lint marketing boards. Sugar production was largely privately owned; while the government controlled the production of tea through the Uganda Tea Growers Corporation; the Produce Marketing Board controlled the sale of such crops as tobacco, maize, beans, groundnuts, sesame (simsim), and chillies; the National Trading Corporation had a monopoly of the distribution of shirts, cement, steel windows, and hoes; and the Import Licensing Board promoted the "Ugandization" of commerce and industry. Perhaps the most efficient institution was the Uganda Electricity Board, responsible for the Owen Falls Dam and for planning a second dam on the Nile.

Politically, it was Buganda and all that Buganda stood for that haunted Obote's era. In 1963 a number of his Kabaka Yekka (KY) associates had published a pamphlet arguing for closer association with his Uganda People's Congress (UPC). In June 1963 nine KY members joined the UPC, and in October Obote confirmed his alliance with the leaders of Buganda by securing the election of Kabaka Mutesa II as first president of Uganda. He was troubled, however, by the possibility that the Buganda members might switch their support to his Democratic Party (DP) opponents. In February 1964 the DP won four out of six seats on the Kampala municipality and all six seats in Jinja. Obote therefore did everything he could to encourage further shifts to his own party and already in January 1964 had begun to talk about a one-party state. Changes of allegiance, both from the KY and the DP, did in fact ensue, until with 60 UPC members as against 15 KY members and 16 DP members in the national assembly Obote took the fateful step on August 24, 1964, of ending his coalition. When on the last day of 1964 the DP leader changed sides, only nine of the original 24 DP members were left. By this time the UPC was in control of all the district councils outside Buganda; and in July 1965 six out of the remaining 14 KY members finally joined the UPC as well.

All this was of first importance because by 1964 conflict with Buganda was acute. There was a long-standing clash over the administration of Buganda's finances and over the size of Uganda's subventions to Buganda. The Kabaka, moreover, was actively sponsoring an attempt to frustrate a referendum in the western counties of Buganda as to whether they should now revert to Bunyoro. Obote held the referendum on November 4; the "lost counties" were returned to Bunyoro; a Buganda government fell from power; and Obote emerged triumphant.

Even so, there were during 1965 several motions of no confidence in the UPC in the district councils. At the end of April 1964 the UPC's secretary general had been replaced, and by 1965 party rivalries ran deep. During 1965 there was also much anxiety over Obote's support to the rebellious groups in the northeastern Congo.

The Kabaka versus Obote

The ensuing climax was reached on February 4, 1966, when, during Obote's absence from Kampala, Daudi Ocheng won Cabinet support for a motion accusing Colonel (later General) Amin—and by extension Obote himself—of peculation in the course of their Congo contacts. Obote ordered a rigorous enquiry, but, following the capture on February 15 of Buganda's UPC by a moderate faction, now emerging as opposed to him personally, at a Cabinet meeting on February 22 he arrested five of his leading colleagues. Immediately after (angered by Kabaka Mutesa's appeal at this time to the British high commissioner for assistance), he declared himself executive president, and on April 15 he extracted the support of the national assembly for a new constitution, which they had not even seen; by this the federal provisions of the constitution agreed at independence were abrogated. The Buganda *lukiko* ("council") thereupon demanded the expulsion of the central government's offices from Buganda by May 30, and Obote in turn arrested the three chiefs who had moved the motion. As soon as this news became public, violence broke out in Buganda, and in response, on May 24, 1966, Obote despatched Amin's forces against the Kabaka's palace. Mutesa escaped to Britain.

Three major consequences ensued. Obote was now heavily dependent for the maintenance of his regime upon the support of the army. He had handled it cautiously when it had joined in the series of mutinies set off in Tanzania in January 1964. From 1963 onward it had had to contend with a violent revolt by the Konjo (Bakonjo) and Amba (Bwamba) peoples of Mt. Ruwenzori against their continued incorporation in Toro. It had been concerned as well with events on the southern frontier, where, following the Rwanda revolution of 1961, more than 100,000 Tutsi refugees had crossed into Uganda. It was also preoccupied with the protracted revolt in the southern Sudan. By the end of 1968 over 150,000 refugees had crossed into Uganda from The Sudan, and various consequential episodes gravely impaired the relations between the two countries. After 1966 Obote had to be particularly solicitous of the army's desires. Before long, over 10 percent of Uganda's budget was devoted to its upkeep—nearly the total of the Kenyan and Tanzanian military budgets together. But he sought also to divert power from the commander in chief and to build up his own "praetorian" guard.

Following the attack on the Kabaka's palace, he not only abolished the centuries-old rulership and its government in Buganda but in 1967 abolished the rulerships in the neighbouring kingdoms also, thus bringing all the local governments that had hitherto fashioned themselves

upon the Buganda model much more directly under his own government's control. Because of the relative autonomy that each of these had during the colonial period, they had been able to act in considerable independence of the central government. In a series of amendments to Uganda's local-administration legislation, Obote's government had already sought to curb their freedom. Now with the destruction of the prototype in Buganda it was able altogether to reduce them to submission.

Having proceeded violently against the Buganda kingdom and then having reduced its replicas in the rest of the country as well, President Obote, as he had now become, had in effect destroyed the major political resource that Uganda had inherited upon its independence. Since he was then essentially a dictator dependent upon the army, he became greatly exercised by the need to recreate his political authority and his political resources. During 1967 he had a third constitution (since independence) extensively discussed, and he then embarked upon a series of "meet the people" tours, during which he repeatedly stated that there was now in Uganda one state, one government, one parliament, and one people. He also declared that he and his ministers had come to meet their masters. This, however, availed him very little—particularly in Buganda. Altering course, he refurbished his populist and Socialist creed, announced a move to the left and a new political culture for Uganda, and published his "Common Man's Charter." At a conference in December 1969 this was endorsed by the UPC, whom Obote urged to uproot ruthlessly such elements of feudalism that might still remain. In May 1970 he proclaimed that the government would assume majority control of a number of commercial enterprises, and in July 1970 he published his *Proposals for New Methods of Election of Representatives of the People to Parliament*, in which he proposed to adopt the Tanzanian scheme of multiple candidates from the governing party but also to add a complex provision that required candidates to obtain votes from four different parts of the country simultaneously. Its purpose, so Obote declared, was to break the hold of Uganda's tribal masters and to insist that the national assembly and its members were national.

During these years Obote was much assisted by the decision of the majority of senior civil servants, many of them Baganda, to stay at their posts. He was further assisted by the death in London on November 21, 1969, of Kabaka Mutesa. Not only, however, was his attempt to secure popular support ineffectual but he also lost three of his closest associates when A.A. Nekyon and Godfrey Binaisa left the Cabinet in May 1967 and Cuthbert Obwangor was dismissed in July. He further alienated support in 1968 when he proceeded against Mayanja and Neogy, a Muganda politician and a journalist. At the UPC conference in December 1969 he was himself the object of an assassination attempt. By the end of the 1960s Uganda was to all appearances a dictatorship; an elaborate spy network existed, and a great many people were afraid to speak freely. Most leading politicians were fearful, moreover, lest in the forthcoming elections they should lose their seats.

The denouement came in January 1971, when, during Obote's absence at a Commonwealth conference in Singapore, a forceful attempt to establish control over the army before the elections were held misfired, and Amin retaliated with a coup d'etat.

Despite a change in leadership, the old problems remained. Though Amin had Mutesa II's body returned to Buganda and there buried in state, he was careful to declare that the kingdoms could not be restored. Though he, too, went on meet-the-people tours and assiduously proclaimed concern for the opinions of the elders, he had the elections postponed; and the local governments found their position further attenuated. In August 1972 Amin ordered expulsion of most of Uganda's Asians, who were the dominant force in the nation's economy.

KENYATTA'S KENYA

Kenya was dominated throughout these years by Jomo Kenyatta. A month after independence, Kenya suffered a military mutiny similar to that of its two neighbours, but, as nothing untoward resulted, the authority of the government was considerably strengthened thereafter. In the first year of independence KADU still stood in opposition to the KANU government but following several changes of allegiance went into voluntary liquidation in November 1964. On August 14, 1964, Kenyatta announced that the regional authorities were to be abolished, and under the terms of a constitutional amendment in November he became president of the new republic rather than prime minister.

To a greater degree than in either Uganda or in Tanzania, parliament in Kenya throughout the ensuing decade provided an open forum for lively national debate, but successive amendments to the constitution continued to increase the authority of the president. Built very deliberately upon the existing edifice of colonial administration, this was chiefly exercised—again in contrast to Uganda and Tanzania—through the bureaucratic hierarchy of provincial and district commissioners (all of them African by 1965). In December 1964 control over the provincial administration was transferred from the Ministry of Home Affairs to the president's office, and as the 1960s wore on the provincial administration became very close not only to the president himself but also to his party.

Though by 1965 Kenya had become in effect a one-party state, considerable intraparty differences were apparent. There were differences in foreign policy over the degree of Kenya's associations with the "East" and with the "West": differences too between Luo and Kikuyu; between the major Luo politicians Oginga Odinga and Tom Mboya; and more particularly between those who argued for the interests of the majority and those who affirmed that they were not going to rob people of property and then claim that they were carrying out the tenets and the requirements of African Socialism. In May 1965 Kenyatta's government, in effect associating itself with the latter, published its Sessional Paper No. 10 on *African Socialism and Its Application to Planning in Kenya*, which explicitly rejected the expropriation of property and the full nationalization of the means of production and argued for a mixed economy tempered by state control so as to procure both a fairer distribution of wealth and increased equality of opportunity. Thereafter, the breach between the two factions became substantial. At a party conference in March 1966 Odinga was ousted as KANU's sole vice president, and in April he resigned from the government. He then assumed the leadership of the newly formed Kenya People's Union (KPU) with 30 other dissenting members. The government responded by introducing a constitutional amendment that required members who changed their party allegiance to resign their seats. There followed the "little general election" of June 1966, in which, though Odinga himself was returned and the KPU won a majority of the votes, only nine KPU members were elected to face the large KANU majority.

Kenyatta's conservative Socialism

The KANU government harassed the KPU remorselessly. It denied the latter permission to hold meeting, had Bildad Kaggia arrested early in 1968, prevented Odinga from going abroad in March 1968, and in August disqualified KPU candidates from the local government elections. Finally, in January 1969, when the Minister of Education banned Odinga from speaking at the University College in Nairobi, the government sent troops against the students who protested.

By this time, however, owing to Kenyatta's advancing years, there were new strains within the government itself, essentially over who should be his successor. Tom Mboya had shown himself to be the government's ablest parliamentarian, and as secretary general of KANU and minister of economic planning and development he was holding key positions. But he was a Luo, and many Kikuyu, flushed by the now-extensive Kikuyu dominance of government, bureaucracy, and African commerce in Kenya, were determined that power should continue to remain in Kikuyu hands. By 1969 there was a new wave of Kikuyu oath taking, and on July 5 Mboya was murdered. Kikuyu confronted Luo at his funeral, and when

Kenyatta visited Luo country in October 1969 open violence erupted. Thereupon he had Odinga and his fellow KPU parliamentarians detained and their party altogether banned.

He next proceeded to hold a general election on January 3, 1970. With only one party allowed to contest it, its conclusion was foregone. Many real contests in the primaries, however, preceded it; almost half the sitting KANU members—including five Cabinet ministers and 14 junior ministers—lost their seats.

He and his regime lived on into the early 1970s. There was a treason trial in 1971 involving some prominent Luo. But in March 1971 Odinga was released from detention, and a few months later he rejoined KANU.

During these latter years there had been a strong move toward the Africanization of commerce in Kenya; and both there and in the neighbouring countries this meant that large numbers of Asians were deprived of their livelihoods. By 1966 Kenya had a population estimated at nearly 10,000,000. Its gross domestic product was now £385,000,000, while in 1968 its gross farm revenue attained a new record of £70,000,000. By that time, 1,209,680 of Kenya's children were in primary school. All of these were levels of achievement well above Tanzania's. Kenya's industries were, moreover, dominant in East Africa. Its primary production included tea, maize, rice, coffee, cotton, sugar, wheat, sisal, pyrethrum, beef, mutton, and dairy produce. Although between 1964 and 1968 sisal production dropped, pyrethrum production went up from 4,300,000 tons to 12,000,000; tea from 20,000,000 tons to 30,000,000; wheat from 128,000,000 to 225,400,000; sugarcane from 591,400,000 to 919,300,000; and maize from 87,000,000 to 351,000,000 tons. Kenya's economic policy was at one, moreover, with its sociopolitical policy. Kenyatta, by contrast with Nyerere and later with Obote, preached the doctrine of personal endeavour, which was embodied in his slogan "*Harambee*" ("Self-help"). Private ownership and private industry were actively encouraged, and it was strongly argued that in view of the gross poverty that existed it was irresponsible not to go out for every possible kind of economic contribution. Such income disparities as resulted, it was argued, could be checked by policies of progressive taxation.

Kenya's "*Harambee*" policy

In East Africa there was as a result a keen debate between the merits of this Kenyan method of development and that of Socialist Tanzania. The outcome of the debate was still unresolved in the early 1970s.

BIBLIOGRAPHY. B.A. OGOT and J.A. KIERAN (eds.), *Zamani: A Survey of East African History* (1968), the best single-volume survey; R.A. OLIVER *et al.* (eds.), *History of East Africa*, 3 vol. (1963–73), the most ambitious account so far; P.L. SHINNIE (ed.), *The African Iron Age* (1971), contains relevant and authoritative articles by H.N. CHITTICK and J.G. SUTTON; *Azania* (annual), the journal of the British Institute of History and Archaeology in East Africa, with authoritative "pre-colonial" articles; G.S.P. FREEMAN-GRENVILLE, *The Medieval History of the Coast of Tanganyika* (1962), subject now to correction, but still valuable; C.S. NICHOLS, *The Swahili Coast: Politics, Diplomacy and Trade of the East African Littoral, 1798–1856* (1971), a very full study; B.A. OGOT, *A History of the Southern Luo*, vol. 1, *Migration and Settlement, 1500–1900* (1967), an up-to-date account based on oral tradition; ROBERT L. HESS, *Italian Colonialism in Somalia* (1966), an authoritative and informative study; I.M. LEWIS, *The Modern History of Somaliland, from Nation to State* (1965), the best book on the subject in English; G.H. MUNGEAM, *British Rule in Kenya, 1895–1912* (1966), well researched; GEORGE BENNETT, *Kenya, a Political History: The Colonial Period* (1963), brief but authoritative; C.G. ROSBERG and JOHN NOTTINGHAM, *The Myth of "Mau Mau": Nationalism in Kenya* (1966), somewhat controversial, but essential reading; M.P.K. SORRENSON, *Land Reform in the Kikuyu Country* (1967), a detailed account of the land consolidation program and its setting; CHERRY J. GERTZEL, *The Politics of Independent Kenya: 1963–68* (1970), a full discussion of central political issues; D.A. LOW and R.C. PRATT, *Buganda and British Overrule, 1900–1955* (1960), a record of the colonial relationship; D.E. APTER, *The Political Kingdom in Uganda* (1961), a stimulating if sometimes uncertain study; C.C. WRIGLEY, *Crops and Wealth in Uganda* (1959), an outstanding essay; F.G. BURKE, *Local Government and Politics in Uganda* (1964), deals mainly with three districts in northern, western, and eastern Uganda; D.A. LOW, *Buganda in Modern History* (1971), discussion of various topics from *c.* 1860 to 1970; M.F. LOFCHIE, *Zanzibar: Background to Revolution* (1965), a standard account; ANDREW ROBERTS (ed.), *Tanzania Before 1900* (1968), summaries of important recent research; I.N. KIMAMBO and A.J. TEMU (eds.), *A History of Tanzania* (1969), patchy but useful; JOHN ILLIFFE, *Tanganyika Under German Rule, 1905–1912* (1969), a stimulating and authoritative study; R.A. AUSTEN, *Northwest Tanzania Under German and British Rule: Colonial Policy and Tribal Politics, 1889–1939* (1968), a case study of colonial rule; G. ANDREW MAGUIRE, *Towards "Uhuru" in Tanzania: The Politics of Participation* (1969), a valuable case study of Tanganyikan nationalism.

(D.A.Lo.)

East African Cultures

The cultures described in this article are those of the region of Africa stretching from the upper Nile Basin, through Uganda, Kenya, Tanzania, and parts of central Zambia to southern Mozambique, Rhodesia, Botswana, and South Africa. The cultures of Ethiopia and Somalia are excluded, as are those of the predominantly matrilineal peoples of northern Mozambique and Zambia (see ETHIOPIAN AND SOMALIAN CULTURES).

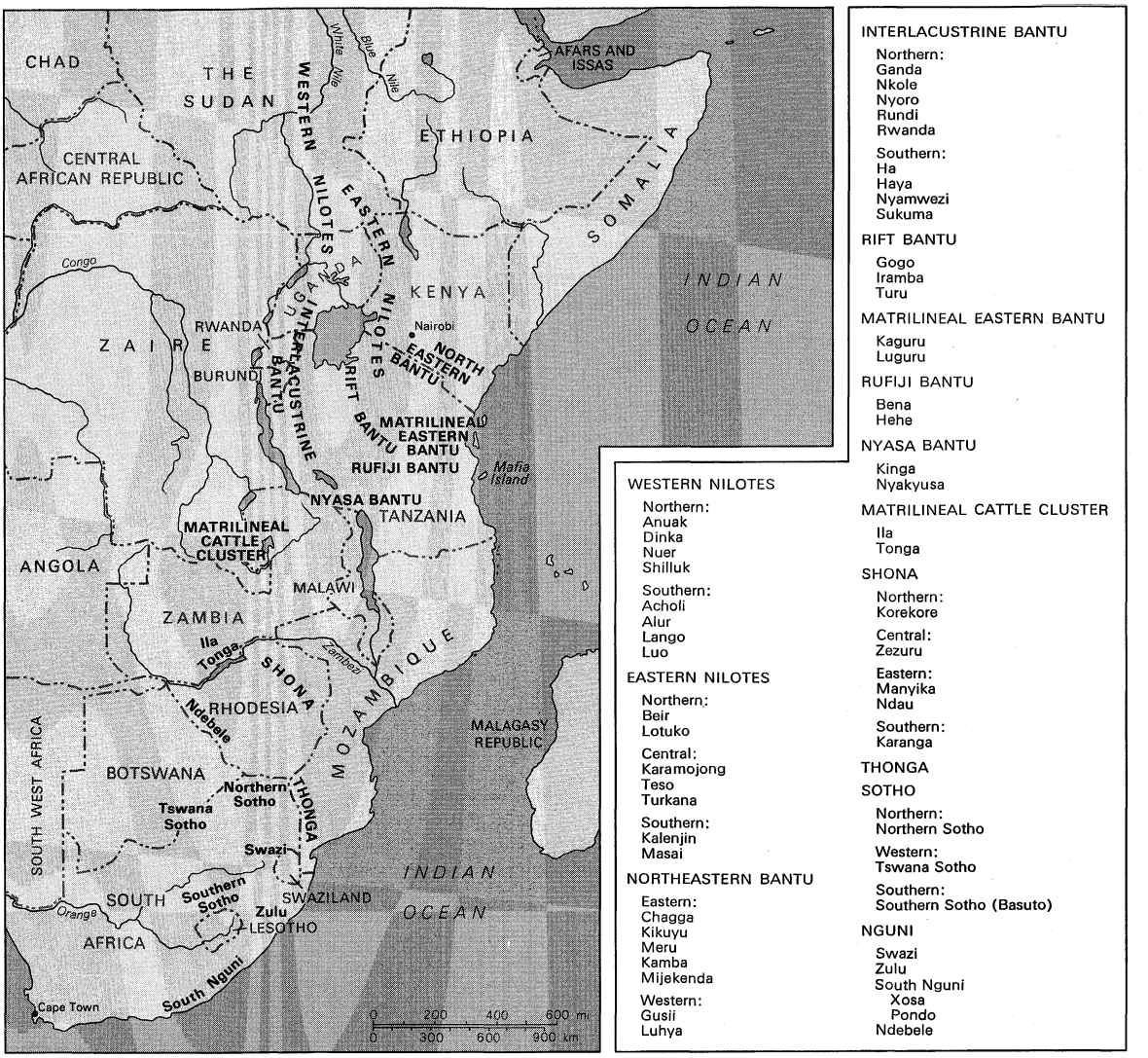
Peoples and their cultures are nowhere susceptible to precise classification, and this is especially so in this region. A good deal of the information necessary for a full account of the peoples of the area is still unavailable or uncertain. Reliable ethnographic and historical studies are limited, both in geographical and cultural respects, and only a handful of cultures have been adequately described in detail. The conventional classification, which is not entirely agreed on by all anthropologists, is somewhat arbitrarily based on a combination of linguistic and ethnohistorical characteristics; but many of these cut across similarities and differences of an economic, political, or religious kind. In the following brief and simplified account only the more important peoples are mentioned. The contemporary population figures are only approximate indications, since demographic data tend to be unreliable in most of the countries concerned and since in many countries enumeration by ethnic or cultural group is avoided.

THE PEOPLES OF THE AREA

In the north of this region, in the upper Nile Basin, the western Nilotes are separated geographically and linguistically into northerners (Nuer, 300,000; Dinka, 500,000) in The Sudan, and southerners (Luo, 900,000, in western Kenya; Lango, 350,000, Acholi, 275,000, and the Alur, 120,000, in northern Uganda). The eastern Nilotes, often referred to by the hybrid misnomer "Nilo-Hamites," are separable into three clusters: the Beir and Lotuko in the southernmost Sudan (data unavailable); the Teso (375,000), Karamojong (130,000), and Turkana (120,000) in northwestern Kenya and northeastern Uganda; and the Kalenjin (500,000) and Masai (230,000) in west central Kenya and northern Tanzania. The Nilotic languages have been ascribed to the Eastern Sudanic family of African languages.

Languages and population

The languages spoken in the rest of East Africa belong to the Bantu, or the southern branch, of the Niger-Congo family. The northeastern Bantu live in the highlands (Kikuyu, 1,500,000; Meru, 450,000; Chagga, 300,000) and in the lowlands eastward to the Indian Ocean (Kamba, 1,000,000; Mijekenda, 400,000); and people of similar cultures are found in western Kenya (Luhya, 750,000; Gusii, 300,000). Farther west, the interlacustrine (living between lakes) Bantu comprise the peoples of the large kingdoms, whose language, history, and state systems are all closely related (Ganda, 1,000,000; Nyoro, 110,000; Nkole, 425,000; Rwanda, 2,600,000; Rundi, 4,000,000). In northwestern Tanzania are a number of peoples, culturally related in some ways, divided among smaller chiefdoms (Haya, 300,000; Ha, 300,000; Sukuma, 1,000,000; Nyamwezi, 400,000). The other peoples of mainland Tanzania can be roughly categorized as follows: Rift Cluster (Gogo, 300,000; Turu, 200,000; Iramba, 140,000); Matrilineal Eastern Cluster (Luguru,



INTERLACUSTRINE BANTU
- Northern:
  - Ganda
  - Nkole
  - Nyoro
  - Rundi
  - Rwanda
- Southern:
  - Ha
  - Haya
  - Nyamwezi
  - Sukuma

RIFT BANTU
- Gogo
- Iramba
- Turu

MATRILINEAL EASTERN BANTU
- Kaguru
- Luguru

RUFIJI BANTU
- Bena
- Hehe

NYASA BANTU
- Kinga
- Nyakyusa

MATRILINEAL CATTLE CLUSTER
- Ila
- Tonga

SHONA
- Northern:
  - Korekore
- Central:
  - Zezuru
- Eastern:
  - Manyika
  - Ndau
- Southern:
  - Karanga

THONGA

SOTHO
- Northern:
  - Northern Sotho
- Western:
  - Tswana Sotho
- Southern:
  - Southern Sotho (Basuto)

NGUNI
- Swazi
- Zulu
- South Nguni
  - Xosa
  - Pondo
- Ndebele

WESTERN NILOTES
- Northern:
  - Anuak
  - Dinka
  - Nuer
  - Shilluk
- Southern:
  - Acholi
  - Alur
  - Lango
  - Luo

EASTERN NILOTES
- Northern:
  - Beir
  - Lotuko
- Central:
  - Karamojong
  - Teso
  - Turkana
- Southern:
  - Kalenjin
  - Masai

NORTHEASTERN BANTU
- Eastern:
  - Chagga
  - Kikuyu
  - Meru
  - Kamba
  - Mijekenda
- Western:
  - Gusii
  - Luhya

Approximate distribution of the major cultural groups of East Africa.

200,000; Kaguru, 90,000); Rufiji Cluster (Hehe, 200,000; Bena, 200,000); and Nyasa Cluster (Nyakyusa, 250,000; and the Kinga, 60,000). In Zambia the people of the centre of the country (Tonga, 250,000; and the Ila, 60,000) are identified by their emphasized association with cattle. In Rhodesia, all but the southwestern area was occupied by the Shona peoples, though now they are interspersed with extensive European settlements. The Shona (about 1,000,000) divide into distinctive cultural subgroups: Korekore in the north, Zezuru in the centre, Manyika and Ndau in the east, and Karanga in the south. The Thonga peoples to the east, in southern Mozambique, are linguistically and probably ethnohistorically linked with the Shona. To the south are the numerous Sotho peoples: the Northern Sotho of the Transvaal, the Western or Tswana Sotho of Botswana and northwestern Transvaal (900,000), and the Southern Sotho (Basuto) mainly of Lesotho (1,000,000). The Nguni peoples are divided among the Swazi of Swaziland (410,000), the Zulu of Natal (2,000,000), and the South Nguni (Xosa [Xhosa] and Pondo, 2,400,000), with a 19th-century offshoot, the Ndebele (400,000), in the southern part of Rhodesia.

Cultural significance of cattle

The majority of East African peoples were subsistence farmers. Some combined animal husbandry with agriculture, and a few were primarily pastoralists economically dependent on their herds. This region has, however, been characterized as the East African Cattle Area because of the common and widespread interests—political, legal, mystical, emotional, and aesthetic—culturally associated with cattle. There is some validity in this view. Even those farmers who were dependent on cultivation for their livelihood gave great attention and high value to their cattle, in contrast to African cattle keepers in other parts of the continent. From the Nguni in the south to the Nilotes in the north, cattle had a special cultural emphasis and importance in the people's lives. But the particular kinds of significance and the ideas and values varied a good deal. Possession and use of cattle thus could be political prizes and the means of domination or they could express cooperative egalitarianism; cattle could embody deep mystical qualities or be treated as entirely profane; they could be vehicles of aesthetic expression but not always so. Much depended on the culture and other interests of the cattle keepers. There was also inevitable difference of emphasis between societies in which there were many cattle and those in which cattle were less numerous than human beings. In most cases, too, goats and sheep, though less valuable individually, could be used for almost all of the same purposes as cattle. Finally, cattle were not kept by peoples living in forest areas or in the extensive central areas infested with insect-borne disease.

As regards other cultural features, this area contained a wide variety of institutions, values, means of organization, and beliefs, and this was consistent with the very considerable historical movements and diverse origins of the peoples, as well as with the wide range of environmental conditions from rain forest to desert and coastal plain to high mountain. External influences from the Indian Ocean shores (Arab, Persian, Indian, Indonesian, and latterly European) spread slowly and irregularly and generally indirectly so that for the most part East Africa

remained isolated from the outside world until well into the 19th century. Technology was quite limited, and, despite a few notable exceptions in the major kingdoms, political and economic organization was comparatively simple and embraced small groups of people. The only world religion to affect the area before the mid-19th century was Islām, and it seldom spread beyond the coastal areas, despite many centuries of existence there.

FAMILY AND KINSHIP

Extended families. The isolated nuclear family of a married couple and their unmarried children (typical of Western industrialized cultures) was rare in traditional East African cultures. It occurred fairly regularly only in a few societies with highly individualized ownership of property and among some of the shifting cultivators of the central part of the region. Some kind of compound family homestead or hamlet was more typical. A patriarchal extended family, sometimes within a single fenced enclosure (as among the Swazi) was one form: this comprised an elderly man, with his wife or wives, his married sons and their wives and children, and sometimes younger brothers of the family head. Other forms occurred when brothers, and sometimes paternal cousins, lived together with their wives and children or when an influential man attracted close kinsmen and their wives. Even when something like the nuclear family was more emphasized, the practical and moral requirements of kinship obligations produced households containing a variety of dependents: aged kin, widows and divorcées and their children, and younger siblings and cousins.

Invariably, however, the unit of a woman and her children was the nucleus of family composition and domestic activity. Usually, each wife had her own house, sometimes partly or wholly fenced off, as among the Karamojong, or in some cases, as among the Gogo and Nyakyusa, her own room in a large house. A wife had her own kitchen hearth, where she cooked for her children, and often her own food store or granary. Older children sometimes had separate sleeping houses in the larger domestic groups. Even when a man had but one wife he often had his own separate house or separate room, where he kept his possessions and entertained visitors.

Polygyny

Virtually everywhere polygyny was the accepted ideal, even though only realized by a minority of men in most societies. Apart from male prestige and the benefits of more labour and more children, polygyny was encouraged by the widespread numerical preponderance of women and by the common custom whereby young women were first married to older men and then were inherited or remarried at the death of the husband. Pre-colonial statistics are not available but in Basutoland (Lesotho), for example, almost one man in five had two wives and one in 27 had more than two (1912). Among the nomadic Turkana of northern Kenya, largely untouched by modernization in 1949, 80 percent of men about 40 years of age or over had more than one wife. These were probably among the higher polygyny rates in this region, and it is likely that no more than 5 percent of men could afford second marriages in most of these societies. Such men would have been the more wealthy, and the higher officeholders in the chiefdoms.

With or without polygyny, in some cultures the basic unit of mother and children was especially emphasized as being a permanent "house," using the same word as for the physical dwelling. Among Nguni and Sotho peoples of the south and some of the eastern Nilotes in the north—to some extent in all cultures in which herds were large—each wife was allocated animals and sometimes land, for which she assumed managerial responsibility and from which she obtained her food supply. A wife's allocation became in due course the heritable portion of her adult sons. This group of brothers, also referred to as the "house," jointly inherited, and for a time or, even permanently, controlled and used the property.

Descent systems. In kinship relations beyond the domestic domain of the compound family, the predominant organizing principle was that prime rights and obligations, mutual assistance and coresponsibility, were intimately connected with common descent in the male line (*i.e.*, patrilineally). Wider kin groups were founded on this principle. There were only two areas in which matrilineal descent was the dominant mode: in east central Tanzania among such peoples as the Luguru and Kaguru, and in south central Zambia among the Tonga and Ila. In the narrow coastal strip from southern Somalia to northern Mozambique, the Swahili and Shirazi peoples, despite centuries of Arab-Muslim influence, gave more or less equal recognition to kin traced through both parents. In some areas (Mafia Island, for example) people could choose to associate with one or more groups in which descent was traced through both males and females from a known or hypothetical ancestor. Even wider, multilateral recognition of kinship, with little or no acknowledgment of descent lines, occurred among some peoples of southern Tanzania.

Lineage and clan organization

In those cultures in which descent and the lineal principle were followed, there was commonly an important lineage based on a founding ancestor of three to five generations preceding that of the older living men. Often, though not invariably, this lineage had a territorial basis, even though some members might live outside lineage land. The genealogically senior member was generally recognized as head and through him the lineage exercised overriding control over land and other property that members severally held. Even when lineage members were scattered among people of other lineages (as among the Tonga) and especially when there was considerable residential unity (as among the Tswana), there was a great deal of continuous cooperation and interdependence. Among the Nguni and the western Nilotes especially, this lineage formed one segment of a more extensive hierarchy of lineages, each more inclusive lineage being focussed on a more remote ancestor. Each level of lineage tended to be associated with a territorial area in which it held or at least was the focus of political influence or control. Beyond the lineage of largest span was the clan, a pseudodescent group in which genealogical connections were often unknown but which was focussed on a semimythical ancestor, a culture hero, who was one of the claimed founders of the culture group. Sometimes such clan founders were not known, but invariably clan membership was determined by descent and was associated with ritual prescriptions. Certain clans, or certain lineages within them, held special privilege over political office or in ritual leadership such as rainmaking or divination.

Marriage and bridewealth. Marriage was not permitted within some designated group—the lineage, one particular level of lineage in a hierarchy, or the clan. Usually, a man could not marry a girl known to be genealogically related within some degree. Among Sotho peoples, on the other hand, marriage was positively encouraged with close kin, including certain first cousins. Among the Lovedu of the Transvaal there was a strong injunction for a man to marry his mother's brother's daughter. Among the Southern Sotho a man's marriage to his father's brother's daughter was said to assist in keeping extended family property together. More generally among the Sotho, marriage was preferred with kin rather than with strangers. This was in marked contrast with many of the northern peoples, among whom there were sayings such as, "We marry those whom we fight": that is, men took their wives from other groups, not related by descent, with whom there was for that reason potential or active hostility.

Bridewealth composition

In most East African cultures, marriage was legitimated by the transfer of livestock (cattle preferentially) from the groom and his kin to the bride's father and his kin. Bridewealth amounts ranged from an average of 50 cattle and 100 or more goats among the nomadic Turkana, to two or three cattle among the Sotho. The size of bridewealth seems to have been directly correlated with (but not necessarily determined by) the number of kin on either side who participated in the transfer and with the general stability of marriage in a society. When livestock were few or absent, bridewealth was composed of other kinds of goods: barkcloth (among the Ganda),

foodstuffs, and ironware such as hoes, axes, knives, and spears. A legitimate marriage also required special ritual performance and the exchange of formal gifts. In most cases the transfer established legal paternity of all children born to the wife. In those societies in which bridewealth was high and divorce rare, a woman remained married to her husband even after his death, so that she could produce and rear his legitimate offspring by regularized consort with a pro-husband, preferably the dead husband's near kinsman. This was the case among both Nguni and the Nilotes; and among some of the Nilotes (for example, the Nuer) a woman could be married to a man already dead in order to raise children to his name (ghost marriage); and a wealthy woman could marry another woman, in both cases by the proper transfer of bridewealth.

An alternative to bridewealth was sometimes possible for a poor man (as among the Shona) and was the cultural norm in some societies of southern Tanzania. This was suitor-service, whereby the groom lived with his father-in-law and worked for him for an indefinite period. Where bridewealth was used, divorce required the repayment of the livestock to the husband, less a certain proportion if he wished to retain paternity and rights over any children of the marriage.

SOCIALIZATION AND EDUCATION

Early child training was the prime responsibility of the mother, but it took place within the milieu of the compound family domestic group in which a child found a range of partially substitute mothers among the other women and also among older sisters and female cousins. Universally, men showed much open affection for small children but were little concerned with their early education. If livestock were important, boys began herding work at least by the age of about eight years and often earlier among pure pastoralists such as the Turkana or Masai. They thus came under the tutelage of their fathers and other male kinsmen. Girls remained closely attached to their mothers, from whom they learned domestic and other skills and mores. Most typically, boys learned sexual knowledge and mores from their peers and those just slightly older, rather than from their parents. Throughout this region it was considered that parents and children, especially sons, should keep their sexual activities and interests quite separate. This was particularly characteristic of those societies in which adult males formed into age groups, as among the eastern Nilotes. There, informal, pre-initiation peer groups provided learning opportunities among equals; sometimes this was also the case for girls, as among the Rwanda. The recurrent dramatic ritual performances associated with marriage, birth, naming, puberty, and death provided some of the most important learning occasions as children and adolescents watched from the sidelines.

More formal education for males took place in connection with initiation and the formation of institutionalized age groups and their subsequent activities. Puberty or initiation "schools" were formally arranged among the Sotho and Nguni peoples and among some of the eastern Nilotes, although these were not universal in the region. Among the Masai, a young man's age group came under the authority and tutelage of a specified older age group in order to learn adult mores, the details of ritual performance, and the nature of political behaviour. In general, there was emphasis on learning among one's peers and from one's peers. Technical skills, however, were learned in the family context.

ECONOMIC SYSTEMS

Among the agricultural peoples, the typical residential pattern almost everywhere was one of scattered homesteads or of small hamlets, each surrounded by its arable land and perhaps some reserved pastureland. Compact villages existed in only a few societies. Round houses with wooden frames, mudded walls, and grass-thatched conical roofs were common throughout the region, but banana-fibre thatch was prevalent in areas in which bananas were staples. In central Tanzania, a number of peoples built low, mudded, thatchless houses, arranged in a hollow square to give an internal courtyard and cattle pen; and in a few areas somewhat similar houses were built in pits so that their earth roofs were only slightly above ground level. Conical houses without walls and thatched to the ground were built in northern Tanzania.

Among the mobile pastoral peoples—and in the semi-mobile cattle camps of more sedentary peoples, such as the Karamojong and Nuer—quickly constructed wind breaks and shade shelters took the place of houses. None of these people had developed a form of tentage.

Subsistence economies

Economies were all essentially subsistence, the household raising its own food supply and making its own clothes and houses. Specialization, other than ironworking, was not usual, although some individuals developed skills as part-time artisans in addition to their normal work: women potters, male woodcarvers, barkcloth makers, herbalists, and diviners provide examples. Even governmental officials (other than the higher officeholders in the larger states) were also farmers or herders. It was a moral as well as an economic responsibility for a man to ensure the production of sufficient food for his household. There was exceedingly little if any production for market exchange, although all households had to barter a little of their produce for minimal outside necessities, such as salt, iron goods, pottery, and medicines. East African farmers were mostly cereal producers: millets with some corn in the south, rather more corn in the central area, and sorghum with finger millet in the drier parts of the north. Quantitatively less important but invaluable as relishes and for variety were beans, sweet potatoes, groundnuts, squashes, and oil-bearing plants. In the wetter interlacustrine region of Uganda and the well-favoured mountain areas of Kenya and Tanzania and in other high-rainfall areas, bananas and plantains were prime staples. Beer was brewed from the staple cereal or from bananas.

Tools were limited to the soft-iron hoe, bushknife, and ax. A few peoples retained wooden digging sticks until the colonial era. In the prevalent semi-arid bush and woodland areas, as on much of the open savanna and veldt, shifting cultivation was normal. Field areas were cleared, cut vegetation then being fired on the site to provide ash fertilizer. Fields were fallowed or virtually abandoned as fertility fell after two to four successive harvests. Where soils were infertile, acid, and deficient in humus, available land was exhausted before natural regeneration restored fertility in old fields, and periodically households or even whole communities shifted to new areas of woodland. With the generally low populations fresh land was readily available, and there was little incentive to develop techniques for the maintenance or improvement of soil quality. In the areas more favoured in soils and rainfall and where populations were denser (typically the mountain areas), permanent farms were normal, with usually a more efficient technology, including crop rotation, soil retention measures on slopes, irrigation, digging in of green weeds, and, occasionally, the use of animal manure. Yields even in these areas were much below those of modern farmers. In many of the more arid areas with lengthy dry seasons, during which agriculture was impossible and soils were damaged, the people recognized more or less regular "hungry months" in the late dry season and early wet season, for harvests were barely sufficient to last out the year. Poor years of inadequate rainfall were common everywhere except in the mountain areas, as was a notable famine perhaps once every two decades. High production years, on the other hand, were uncommon, since the peoples lacked the techniques, labour, and seed to take advantage of better wet seasons. Food was therefore seldom more than minimally sufficient; diets were notably deficient in protein and, on the peoples' own admission, generally monotonous. Few East African farmers had enough livestock to provide more than very occasional meat, and that was provided mainly as a by-product of the ritual slaughter of animals. Hunting, except for the few remnant hunting groups such as the Dorobo and Hadza (Kindiga), was more a leisure pursuit than a food-provid-

Dietary deficiencies

ing activity; but women and girls everywhere collected wild produce, mainly for relishes but of much dietetic value.

Mixed farming, mainly in the north (southern Sudan, northern Uganda, and Kenya) and in the south (South Africa), allowed some augmentation of the prevalent cereal diet with dairy produce (milk, butter, and, still only infrequently, meat). Milk yields from cattle were low, though with a high fat content. Eastern Nilotic peoples, such as the Turkana, Karamojong, and Masai, also fairly commonly bled their animals to obtain a valuable food supplement. The Karamojong, as well as other peoples who might be classified as pastoralists, such as the Nuer of The Sudan, the Kipsigi, Gogo, and Sotho, supplemented arable produce from their herds; but households seldom had sufficient animals to give a regular supply of milk and meat.

Livestock for East Africans was a store of wealth and an investment. Domestic animals were the most valuable and valued assets to East Africans who possessed them, and even in outright famine years pastoralists were unwilling to exchange livestock, or even to slaughter them, for food. A man's animals were essential to his entire way of life. He and his male dependents could not marry without them, nor have legitimate children; he could not properly perform rituals nor acknowledge and entreat supernatural power except through sacrifice and by the sharing of meat with human participants; quite generally he could not make restitution for wrongs committed, nor give reward for favours and friendship, nor engage in gift-giving to reinforce useful social relationships without livestock. Political officeholders and men of rank and aspiration in the East African states sought to attract followers and to control supporters by strategic gifts or loans of cattle.

Land was seldom considered valuable property, if only because, with certain exceptions, it was plentiful, and because under shifting cultivation any particular piece of land tended to have only temporary usefulness. The common rubric in states and stateless cultures alike was that the household should have access to enough land to meet its residential and economic needs. There was little reason to have more than that since it could not be put to advantage. Only particularly favourable pieces of land, such as well-watered areas near streams or good arable land in the more densely populated areas had economic value through scarcity. A local group, such as an extended family, lineage, or village, however, often laid claim to an area, to the exclusion of others, in order to ensure future sufficiency at hand for its members. Group identity could also be expressed in this territorial fashion.

GOVERNMENT

Indigenous governmental systems in East Africa ranged from highly autocratic kingdoms to the "ordered anarchy" of completely decentralized tribes in which an egalitarian ethic and practice was culturally emphasized. The great interlacustrine kingdoms—Ganda in the east, and Nyoro, Toro, Nkole, Rwanda, and Rundi in the west —were efficiently organized predatory states in which everything centred on the hereditary king. In the western kingdoms he had certain superhuman qualities and abilities to reinforce his position, though he was not really a divine king, and these were particularly expressed in his intimate association with the national, magical drums. Only a few subordinate political offices were hereditary, so that appointment and dismissal of administrators and military leaders was tightly controlled by the king on the basis of loyalty and efficiency. In the western kingdoms there was a superior caste, Tutsi or Hima, comprising 10–15 percent of the population. This caste was pastoral and not agricultural, enjoyed certain privileges, and all the more important political and military appointments were made from it. In the Ganda kingdom, however, there was no more than an incipient aristocratic class. Each kingdom was integrated by a pervasive clientship hierarchy. The king held all property, much of which he allocated for use to major clients, such as senior officeholders, military leaders, and favourites. These, in turn, were patrons of their own similar clients by the suballocation of property and protection; and these, too, had their clients, down to the mass of the peasant-serfs. These peasants in the western kingdoms were an agricultural caste comprising about 85 percent of the population and were the autochthonous (indigenous) people of this area. There was a good deal of competition to amass clients at other patrons' expense in the struggle for power and prestige, for clients were effectively economic and political supporters. These kingdoms were conquest states that had gradually overpowered and incorporated their weaker neighbours.

Hereditary kingships

The southern kingdoms of the Sotho and Nguni peoples were rather smaller and less centralized, except for the Zulu empire created in the early 19th century. They were typically composed of homogeneous smaller chiefdoms. Subordinate chiefships and other offices were hereditary and tended to be supported by locally strong patrilineal groups, thereby retaining a degree of autonomy from the king. Officeholders and hereditary councillors met periodically in a national council at which public opinion could be represented to the king; chiefly lineages formed something of a permanent aristocracy.

A variety of smaller chiefdoms existed among the Xosa, Tswana, Venda, and others in the south akin to those larger kingdoms and among the Nyamwezi, Sukuma, Ha, Haya, and Soga in the northwest in the lacustrine area. Chiefs were autocrats, but the chiefdoms were rather unstable because of local competition with each other and the secessionary ambitions of junior branches of the chiefly lineages. Lesser chiefdoms, often rather loosely organized under comparatively weak chiefs, existed among the Shona and elsewhere.

Decentralized systems

In the central, eastern, and northern parts of the region noncentralized polities were general. Some were based on semi-autonomous local communities in which law and order were maintained and disputes and policy matters were settled on the basis of the kinship connections of the people, as among the Tonga. Farther north among eastern Nilotes and to a lesser extent among northeastern Bantu peoples, men were arranged into age groups cross-cutting kinship linkages. Each age group dealt with affairs between comembers and acted as supporting and coercive units in affairs between members of different groups. Where this system was most emphasized, as among the Masai and Kalenjin, each age group collectively passed through a regular sequence of grades as members grew older, with each grade involving public responsibilities for defense, police, political, judicial, and ritual activities for the community. Fraternal egalitarianism among members of the same age group was particularly stressed throughout their lives.

Among the western Nilotes and certain northern Bantu (the Gusii, for example), lineages acted collectively in defense of members' interests. Among some western Nilotes (Acholi, Alur, Lango) this pattern was associated with weak chiefships.

BELIEF AND AESTHETIC SYSTEMS

Ancestor and other spirit cults. Among the Bantu-speaking peoples, a belief in ancestral spirits was the characteristic focus around which religious ideas developed. These spirits of the dead required regular acknowledgment, propitiation, and supplication by congregations of their living descendants, whose welfare they affected or even controlled. Ancestral spirits of chiefs in all the southern kingdoms and chiefdoms were regarded as especially powerful and important because they gave legitimation and authority to their living successors and affected the welfare of the state and the fertility and well-being of the country and its people. In some cases, as among the Shona, the spirits of former chiefs could possess a living person, not necessarily a descendant, and issue information and instructions through their mouths under trance conditions. Nonchiefly ancestors were concerned only with the group of their own living descendants and with the sanctioning of proper behaviour and cooperation among them. Misconduct threatened major

canons of kinship morality and the perpetuation of group unity and was liable to punishment by the ancestors with misfortune, illness, and even death. Ritual performance in expiation and petition were the principal expressions of lineage unity, the interdependence of kinsfolk, and their common origins from and subjection to these ancestors. These spirits were not, however, concerned with the wider morality of proper behaviour to other people of different descent.

Ancestor cults were more developed among the southern peoples and were much less elaborated among interlacustrine and northern Bantu. There, and among the northern peoples without beliefs in the spirits of the dead, there was a concept of immanent supernatural power of impersonal character, sometimes residing in natural objects or invoked in cattle, to which ritual performance was directed. In these areas, too, and occasionally elsewhere, there was a belief in a remote, omnipotent, high god who was susceptible to human supplication and manipulation through sacrifice, prayer, and ritual. In the more arid parts of the region, rainmaking was important, and this was closely associated with the fertility of women, livestock, and fields. Rainmakers were ritual experts but seldom had much worldly power.

Medicine, magic, and sorcery

Ordained priests were rare, but specialist practitioners of supernatural power were widespread. These ranged from herbalists, whose main stock was natural substances magically treated (and often of some pharmaceutical value) to doctors who cured mainly by magical means and who could foretell the future and divine the causes of misfortunes. Among the southern peoples, particularly, doctors underwent lengthy apprenticeship; this was usually precipitated by spirit possession or other psychic experience. The most powerful doctors could detect and combat witches and sorcerers, though usually these specialists themselves could work magical substances so as to harm others on behalf of clients. Some belief in witchcraft was virtually universal in all these cultures but was more elaborated from Tanzania southward and less emphasized by the Nilotes. Witches were seldom claimed to be seen at work; essentially the belief was that certain individuals inherited or sometimes learned the power by which they could harm and kill their victims. Witches were associated with the night and with hyenas, snakes, and other disliked creatures; witches were the epitome of evil and represented the inversion of normality and morality.

Artistic expression. Compared with the rest of sub-Saharan Africa, the traditional cultures of this region were poor in aesthetic expression. Basketry and beadwork decoration of clothing and utensils were quite rich, but the plastic arts were almost totally neglected. In music, unaccompanied group singing was well developed for both social and ritual performance. Drums were principally associated with high rank and political office, reaching a peak significance in the sacred national drums of some of the interlacustrine kingdoms such as Rwanda and Nkole. Xylophones were prominent, especially among the Chopi of Mozambique and the interlacustrine Bantu. Ensembles of drums, xylophones, dancers, and singers were maintained at the courts of interlacustrine kings and involved large numbers of performers on state occasions. Except in the coastal areas long influenced by Arab culture, there was little sophistication in poetry; the words of songs were very simple, in marked contrast to the accompanying musical forms.

THE IMPACT OF THE MODERN WORLD

Cultural evolution and social change were at work in East Africa before the modern era, as climatic, demographic, economic, and political conditions changed and as innovation developed. A new empire, for example, was founded among the Zulu at the beginning of the 19th century, and older kingdoms expanded by conquest and dealt with problems of more extended administration. Tribes came into being as new areas were settled, as among the Chagga on the slopes of Kilimanjaro, while older cultural groups were augmented and often quite radically affected by migrants. The Bantu peoples were penetrating farther south, and the Masai and other pastoral groups were expanding their territories in the north during the 19th century. New or improved types of domestic plants percolated into the area from the outside world—plantains, rice, and corn, for example—creating new diets and calling for new agricultural techniques.

Political organization under colonial rule. Nevertheless, the potential for change increased tremendously with the advent of massive Western intervention. The whole area came under some form of white colonial rule, ranging from dominant interference to laissez-faire paternalism. The principal factors at work were the loss of political autonomy and the introduction of a cash economy. By the third quarter of the 20th century, however, there were areas and peoples, usually geographically remote and environmentally poor, that had still been only slightly affected: the semidesert areas and their nomadic peoples (such as the Turkana and Masai) and the arid bushlands of much of Tanzania and Mozambique, peopled by near-subsistence-level, shifting cultivators. And despite certain major changes, sometimes coercively imposed, nowhere had evolution gone so far as to obliterate the earlier major cultural identities and differentiation. Indeed, it was usually the case that colonial powers froze the ethnic-cultural pattern they found on their arrival, assuming it to be a static one. Under direct and indirect rule alike, the ethnic groups were taken as the units of new administration. In effect, though not always in intent, this gave a new emphasis to cultural identity and its symbols. Some ethnic groups gained unity for the first time under colonial rule, as among the Kikuyu or the Sukuma, and sometimes smaller autonomous but culturally related groups were enforcedly amalgamated, as among the Teso or Nyakyusa. With relatively few exceptions, the end of the colonial age showed a greater emphasis on ethnic (tribal) distinction than had previously existed. This was further emphasized as ethnic groups, especially the larger ones, became involved in modern politics as the power bases of the new African politicians and the units of competition for power and preference. Although quite radical changes had occurred in certain cultural features of most of these ethnic groups, traditional elements not only persisted but also were given prominence as symbols, as part of the ideology, of modern political conflict. On the whole, as a reaction to Western intervention and power, conservatism has been dominant, partly as a defensive mechanism, partly out of continued inability to utilize modern opportunities, and partly because traditional vested interests remained strong.

Employment of traditional rulers by colonial powers

If the indigenous political system was centralized, with recognized officials, the latter were used as local authorities with varying degrees of autonomy ranging from effective subservience in the south to comparative independence in the interlacustrine area. Yet chiefs, headmen, and others became in some degree dependent, for they were paid and could be dismissed by the colonial power. Commonly their position was ambivalent, as their subjects and their masters held different expectations of their roles. They were expected to undertake new tasks, such as the collection of taxes or the enforcement of agricultural regulations; and they had to accept new kinds of assistants and subordinates (responsible for new local government services) and new forms of popular representation, such as appointed or elected councils. They had to compete with a new class of Western-educated men who sought wealth and influence. Nevertheless, the chief, though in large part turned into an administrative official, remained important as a symbol and representative of his people until his office was abolished by independent African governments.

Formerly noncentralized societies had for the first time an organized, centralized administration imposed on them, and the indigenous leaders and notables lost their influence, since they were unable to assume radically novel roles. Political changes in such situations were sometimes considerable, as among the Kikuyu, although the new, wholly unfamiliar kind of government

was often chronically unable to cope with modern needs and responsibilities.

Economic and demographic changes. East Africans were typically first brought into direct contact with the outside economic system through labour migration to farms, plantations, railroads, and growing urban centres. In the south this had begun before the end of the 19th century and almost everywhere, except the southern Sudan, labour migration flourished in the period between 1910 and 1960. As a result, men learned to work for wages and to appreciate the significance of money and Western goods. The poorer areas continued to supply the distant, white-dominated labour markets; for even the rural subsistence economies now had at least minimal needs for cash to pay taxes, to buy some clothing and other store goods, and even to substitute money for traditional wealth in order to pay bridewealth, obtain cattle, and pay fines. Where white influence dominated, as in the south and in Kenya, Africans were positively discouraged from developing their rural economies. In Uganda and to a lesser extent in Tanzania, the development of cash-crop agriculture was fostered—cotton, coffee, tobacco, corn, cashew nuts, coconuts—linking local enterprise directly to the international market. In the kingdom of Buganda, in eastern Uganda, and among the Chagga and Nyakyusa of Tanzania, for example, considerable wealth was generated, leading to higher standards of living, a degree of individualism, and a spread of Western-influenced education, skills, and ambitions. Everywhere there was a tendency for older patterns of interdependence and cooperation to fade, whether among kinsmen or neighbours. Land tenure and economic production generally tended to become more individualized as group controls and communal enterprise slackened. In the environmentally favoured areas, as well as in those in which the Africans' lands were forcibly limited, population pressures became critical, thus inhibiting peasant, subsistence agriculture. Economic development —rapid or not—was threatened by this, and great strains were imposed on kinship and inheritance systems.

The introduction of a cash economy

Although the plow and the ox-wagon have been taken up by some African peoples, more particularly in the south, and even the tractor and truck in scattered instances latterly, the East African farmer still largely depends on hoe cultivation. In the poorer bush areas the traditional slash-and-burn, shifting cultivation persists. Under government pressure such techniques as contour ridging and crop rotation have been introduced; but traditional techniques continue to be only slightly modified by Western innovations.

20th-century urbanization

Apart from a few Arab-inspired coastal towns, there was no urban element in East African cultures until the 20th century. Modern cities, from Cape Town to Nairobi and many lesser towns, grew to serve primarily European and governmental interests; but they became the principal centres of multi-ethnic interaction for African workers and centres of cultural and political innovation. Until fairly recently the majority of African town dwellers were but temporary migrants who, after a period of employment, returned to their rural homes and preferred way of life. Even as the proportion of more or less permanent town dwellers has increased, men continue to visit their home areas and to take them as their cultural reference groups. So far, ethnic differentiation has been only slightly reduced and in some ways exacerbated by the urban experience. Particularly at the level of domestic life, kinship, and marriage, the influence of rural cultures continues. Outside of South Africa the people remained overwhelmingly rural in residence and orientation: in the early 1960s over 90 percent of Tanzanians lived in rural areas and about 85 percent of Kenyans; and even in South Africa the rural population amounted to about 53 percent. This tended to reinforce conservative cultural tendencies and the preoccupation with farming and small-scale social intercourse with kin and neighbours.

Christianity and witchcraft. In general, indigenous religious beliefs and activities have not persisted in the face of Western Christian and secular ideologies, the developing cash economy, and the shock of colonial domination. Intensive Christian mission activity and its close link with educational opportunity diverted converts from former beliefs in ancestral spirits or immanent psychic power. South African peoples in particular, especially Nguni, and others elsewhere in the area have developed many nativistic Christian churches in which Africans seek to become independent of European domination of worship and organization. Renewal of traditional religion and ritual, however, has been rare. Conversely, belief in witchcraft and sorcery has tended to strengthen, perhaps as an emotional and social reaction to insecurity and uncertainty in a period of change and the reduction of traditional sanctions. Short-lived but recurrent witch-finding movements and the widespread fame of certain doctor-diviners have been common throughout the area during the 20th century. This is especially noteworthy, since colonial governments and their modern successors sought to outlaw witchcraft accusations.

BIBLIOGRAPHY. There is no single general work devoted to the cultures of this area, but good summaries of ethnographic data are given in the various volumes of the east central and southern Africa sections of the *Ethnographic Survey of Africa*, ed. by DARYLL FORDE (1956–). Among symposia focussed on some parts of East Africa are JOHN MIDDLETON and EDWARD H. WINTER (eds.), *Witchcraft and Sorcery in East Africa* (1963); ROBERT F. GRAY and P.H. GULLIVER (eds.), *The Family Estate in Africa* (1964); and PAUL BOHANNAN (ed.), *African Homicide and Suicide* (1960). There are many studies of aspects of particular cultures. For the western Nilotes, E.E. EVANS-PRITCHARD's trilogy *The Nuer* (1940), *Kinship and Marriage Among the Nuer* (1951), and *Nuer Religion* (1956), are outstanding. G. LIENHARDT, *Divinity and Experience* (1961), is valuable on the Dinka. For the eastern Nilotes the most useful are P.H. GULLIVER, *The Family Herds* (1955), on the Jie and Turkana; PAUL SPENCER, *The Samburu* (1965); and P.H. GULLIVER, *Social Control in an African Society* (1963), on the Arusha. The northeastern Bantu are poorly described, but H.E. LAMBERT, *Kikuyu Social and Political Institutions* (1956), is reliable. The best studies of interlacustrine Bantu are L.A. FALLERS, *Bantu Bureaucracy* (1956), on the Soga, and *The King's Men* (1964), on the kingdom of Buganda; JOHN BEATTIE, *The Nyoro State* (1971); and J.J. MAQUET, *The Premise of Inequality* (1961), on the Rwanda kingdom, are also useful. Among the few good studies of Tanzanian cultures are PETER RIGBY, *Cattle and Kinship Among the Gogo* (1969); R.G. ABRAHAMS, *The Political Organization of Unyamwezi* (1967); P.H. GULLIVER, *Neighbours and Networks* (1971), on the Ndendehule; and MONICA WILSON's trilogy on the Nyakyusa: *Good Company* (1951), *The Rituals of Kinship Among the Nyakyusa* (1957), and *Communal Rituals of the Nyakyusa* (1959). ELIZABETH COLSON'S writings on the Tonga of Zambia are very valuable: *Marriage and the Family Among the Plateau Tonga of Northern Rhodesia* (1958), *The Plateau Tonga of Northern Rhodesia* (1962), and *Social Organization of the Gwembe Tonga* (1960). The best work on the Rhodesian Shona is J.F. HOLLEMAN, *Shona Customary Law, with Reference to Kinship, Marriage, the Family and the Estate* (1952). The southern cultures are more comprehensively described. A general and analytical account of traditional political systems is given in I. SCHAPERA, *Government and Politics in Tribal Societies* (1956). The most useful particular accounts are to be found in H.A. JUNOD, *The Life of a South African Tribe* (1912–13), on the Thonga; HILDA KUPER, *An African Aristocracy* (1947), on the Swazi; E. HUGH ASHTON, *The Basuto* (1952); the numerous works on the Tswana by I. SCHAPERA, of which the most general is *A Handbook of Tswana Law and Custom*, 2nd ed. (1955); A.T. BRYANT, *Olden Times in Zululand and Natal, Containing Earlier Political History of the Eastern-Ngumi Clans* (1929); and MONICA HUNTER, *Reaction to Conquest*, 2nd ed. (1961), on the Pondo.

Studies of cultural change in the 20th century are still meagre and often in scattered articles. The most useful are: P.H. GULLIVER (ed.), *Tradition and Transition in East Africa* (1969); A.I. RICHARDS (ed.), *Economic Development and Tribal Change* (1954), on the Ganda, and *East African Chiefs* (1959); P.H. GULLIVER, *Land Tenure and Social Change Among the Nyakyusa* (1958); I. SCHAPERA (ed.), *Western Civilization and the Natives of South Africa* (1934); HILDA KUPER, *The Uniform of Colour* (1947), on the Swazi; and B.G.M. SUNDKLER, *Bantu Prophets in South Africa*, 2nd ed. (1961). Two volumes in the Keiskammahoek Survey give a valuable picture of a South African reserve: M.E. ELTON MILLS and MONICA H. WILSON, *Land Tenure* (1953); and D.H. HOUGHTON and E.M. WALTON, *Economy of a Native Reserve*

(1953). There is an increasing number of studies of Africans in the new urban areas, such as AIDAN W. SOUTHALL and PETER C.W. GUTKIND, *Townsmen in the Making* (1957), on Kampala; J.A.K. LESLIE, *A Survey of Dar es Salaam* (1963); and PHILIP MAYER, *Townsmen or Tribesmen* (1961), on East London.

(P.H.Gul.)

East African Lakes

The majority of East African lakes lie within the great rift-valley system that forms a part of a series of gigantic fissures in the Earth's crust extending northward from the Zambezi Valley through eastern and northeastern Africa and the Red Sea to the Jordan Valley. In East Africa itself, southern, eastern, and western branches of the rift can be discerned. Occupying the Southern Rift is Lake Nyasa (Lake Malawi), which drains into the Zambezi River. Marking the course of the Western Rift are Lakes Tanganyika, Kivu, Edward, and Albert—the first two of which are within the drainage basin of the Congo River, while the other two constitute part of the Nile River drainage system. With the exception of Lake Rudolf, the lakes of the Eastern Rift are smaller than those of the Western Rift and comprise several independent inland drainage basins.

Located in a shallow downwarping between the Eastern and Western Rift highlands is Lake Victoria, which among the freshwater lakes of the world has a surface area second only to that of Lake Superior in North America. On a smaller scale, East Africa also includes some fine examples of crater lakes and, on Mt. Kenya and in the Ruwenzori Range, of glacial tarns, small lakes each of which occupies a basin, or cirque, scraped out by a mountain glacier. (For associated physical features, see CONGO RIVER; EAST AFRICAN MOUNTAINS; NILE RIVER; VIRUNGA MOUNTAINS; and ZAMBEZI RIVER.)

The eight largest lakes

Of the eight largest lakes—Victoria, 26,418 square miles (68,423 square kilometres) in area; Tanganyika, 12,700; Nyasa, 11,430; Rudolf, 2,473; Albert, 2,059; Rukwa, 1,100; Kivu, 1,042; and Edward, 850—only one, Rukwa, in Tanzania, lies wholly within a single political entity. The northern shore of Kenya's Lake Rudolf is in Ethiopia; Lake Victoria is divided among Uganda, Tanzania, and Kenya. In the west, the international boundary between Uganda and Zaire (former Democratic Republic of the Congo) runs through the centre of Lake Albert; the same boundary places two-thirds of Lake Edward in Zaire and the remainder in Uganda. Lake Kivu lies between Rwanda and Zaire; the waters of Lake Tanganyika are shared among Tanzania, Zaire, Burundi, and Zambia. Malawi and Mozambique have territorial waters on Lake Nyasa, and, since its independence, Tanzania has also advanced claims to its territorial waters because it also occupies a part of the lakeshore.

Although not all the shores of Lake Victoria are well settled, the lands marginal to the lake are among the most densely populated in Africa. The best known groups inhabiting these marginal lands are the Ganda, Soga, Luyia, Luo, Sukuma, and Haya, all these groups except for the Luo being Bantu-speaking. The lowlands at the northern and southern ends of Lake Nyasa are also well populated, but, in general, the shores of the Eastern and Western Rift lakes are sparsely settled.

On Lake Victoria the Luo and the Sese Islanders have shown a marked interest in the fisheries. Peasant fishermen operate in the northern part of Lake Tanganyika; the majority of the fishermen on Lake Nyasa also practice subsistence agriculture, except on Likoma Island, where there is little available agricultural support.

General description. *Eastern Rift lakes.* The surface levels of the lakes on the irregular floor of the Eastern Rift are of varying heights, rising from Lake Rudolf (1,230 feet) through Lake Baringo (3,200 feet) to Lake Naivasha (6,180 feet), after which there is a decrease in height to Lake Magadi (1,900 feet). The Omo River from the Ethiopian Highlands is the only perennial affluent of Lake Rudolf, which is the lowest of the major East African lakes. Although it has the typical elongated form of a rift lake, Rudolf is relatively shallow (240 feet), as are the other lakes of the Eastern Rift. Its eastern and southern shores are bounded by rocky margins of volcanic origin; the lower western and northern shores are mostly sandy sediments. South of Lake Rudolf the splaying continuation of the Eastern Rift into northern Tanzania is indicated by Lakes Natron, Manyara, and Eyasi.

Western Rift lakes. In the Western Rift, the northwestern and southwestern shores of Lake Albert are flanked by steep escarpments; wild ravines and fine cascades comprise a feature of these young tectonic (fault-formed) landscapes, the scale being greater on the Zaire than on the Uganda side. There is a considerable lowland area at the northern end of the lake, where, about 20 miles below Murchison Falls, the Victoria Nile enters Lake Albert, to leave almost immediately as the northward-flowing Albert Nile. The southern end of Lake Albert contains an alluvial flat and a papyrus-choked delta formed by the Semliki River, which carries the outflow from Lake Edward and drainage from the rain-soaked Ruwenzori Massif.

Lake Albert

Lake Edward, of which the deepest part (384 feet) is in the west under the Congo Escarpment, receives the Rutshuru River as its principal affluent. On the northeast it is connected with Lake George by the 3,000-foot-wide Kazinga Channel. At an altitude of approximately 3,000 feet above sea level, the surface of both lakes is nearly 1,000 feet higher than that of Lake Albert.

Separating the basins of Lake Edward and Lake Kivu are the Virunga Mountains, which thus divide the drainage system of the Nile River from that of the Congo River. With clear water, a broken shoreline, and a mountainous setting, including the relatively recent volcanoes of Nyamulagira and Nyiragongo, Lake Kivu possesses outstanding scenic attractions. Its outflow is southward by the turbulent Ruzizi River, which drops more than 2,000 feet on its way to Lake Tanganyika. This latter lake, long and narrow, is second only to the Soviet Union's Lake Baikal in depth, penetrating at its maximum to 4,710 feet, which is more than 2,000 feet below sea level. Typical, too, are the flanking escarpments, which often rise sheer from the lake; the only sizable lowland is the lower Ruzizi Valley. The drainage of the Malagarasi River system enters Lake Tanganyika about 25 miles south of Kigoma; at its southern end, on the frontier with Zambia, the 704-foot Kalambo Falls occur where the Kalambo River tumbles over the escarpment. The overflow to the Lualaba River, a tributary of the Congo, is via the shallow and sometimes obstructed Lukuga River outlet on the western side.

Lake Victoria. With its quadrilateral shape, relative shallowness (270 feet), and an area that is more than twice as great, Lake Victoria is quite different from Lake Tanganyika. Set in a region of erosion surfaces instead of tectonic escarpments, it is bounded by a shoreline of considerable variety: on the west a straight, cliffed coast gives way to papyrus swamp; headlands and deep indentations mark the intricate northern shores; a major inlet, the Kavirondo Gulf, is located on the east; and on the southern shores the Speke, Mwanza, and Emin Pasha gulfs lie amid rocky, granitic hills. Ukerewe, situated in the southeast, is the largest island in the lake; in the northwest the Sese Islands comprise a major archipelago. The Kagera River, largest of the affluents, may be considered as being the most remote headstream of the Nile. The outlet of the lake and the conventional source of the Nile is at Jinja, where after flowing over the now-submerged Ripon Falls, the Victoria Nile begins its journey toward the Mediterranean Sea through the sluices of the dam at Owen Falls.

Lakes Rukwa and Nyasa. Lake Rukwa is situated in a northwest- to southeast-trending side rift, parallel with the southern part of the Lake Tanganyika Rift and continuing the structural alignment of the northern end of Lake Nyasa. Lying on flat alluvium (soil, gravel, sand, and stone deposited by running water), Rukwa is extremely shallow (20 feet at the greatest depth), and any change of surface level causes great fluctuations in its area. Southeast of Lake Rukwa, beyond the volcanic mass of Rungwe Mountain, Lake Nyasa, third in size

among the East African lakes, has the same characteristics as Lake Tanganyika but in less extreme form. It is deepest in the north (2,310 feet), where, on the Tanzanian side, the Livingstone Mountains rise precipitously from the lake surface. In the northwest, however, there is a well-defined alluvial plain. From the east come the waters of the Ruhuhu River, and numerous streams flow across the Malawi plateaus to the western shore. In the shallower, southern part there are several lake plains and sandy beaches. The Shire River outlet at the southernmost end has an extremely small gradient in its upper section, but in its middle course the river is interrupted by cataracts.

Geological history. The East African rifts attained their present form mainly as a result of Pleistocene earth movements from 1,000,000 to 2,500,000 years ago, and the lakes must have been formed after the landscapes in which they are set. The shallowness of such lakes as Albert (maximum recorded depth 168 feet) and Edward (384 feet) is the result of the thick layers of sediment upon which they rest. In some lakes, too, volcanic activity has played a part in blocking drainage and shaping shorelines. Raised beaches indicate that the lake levels were higher and their surfaces were more extensive during rainy phases of the Pleistocene Epoch. In the Eastern Rift, for instance, Lakes Rudolf and Baringo were formerly part of one lake, from which there was a link via the Sobat River with the White Nile. Subsequent drier conditions caused the eastern lakes gradually to dwindle in size, with many fluctuations, to their present independent status.

Changes since the Pleistocene

The lakes of the Western Rift have experienced their own geological changes. Before the geologically recent blocking of the rift by the eruption of the Virunga Mountains, the drainage of Lake Kivu was probably northward into the Nile; the fossil record suggests that Lakes Edward and George were completely sterilized during a period of intense volcanic activity around their shores. In the middle Pleistocene (about 1,750,000 years ago), when the rise in the shoulders of the Western Rift resulted in the reversal of the westward-flowing drainage of such rivers as the Kagera, Katonga, and Kyoga-Kafu, Lakes Victoria and Kyoga were formed by water diverted from the northern section of the rift. Eventually, however, the drainage of most of Uganda returned to the Western Rift and to the Nile. The subsequent reductions in the level of Lake Victoria are indicated by a series of strandlines around its shores.

Climate. Those East African lakes that lie in inland troughs at, say, 2,000 feet or less above sea level, have a hot, dry climatic environment with a high potential evaporation. In the higher parts of the rift floors, however, climatic conditions approach those of the flanking highlands. In the Western Rift, moist Congo air is a source of the more humid conditions prevailing over Lakes Tanganyika and Kivu. The glacial tarns of Mt. Kenya and the Ruwenzori Range are in the frigid zone.

Influence of the lakes on climate

Lakes also tend to create or influence their own climates. This effect is most marked on the western and northern margins of the immense mass of Lake Victoria, where, in a zone 30 to 50 miles wide, temperatures rarely rise above 80° F (27° C) or fall below 60° F (16° C) and rainfall is well distributed throughout the year. Moreover, annual rainfall is high, being heaviest over the lake and decreasing inland from an average of 50 to 60 inches at the lake shores. At the northern end of Lake Nyasa, an annual rainfall total of about 120 inches results from similar influences reinforced by air convergence due to the funnel-shaped relief at the head of the lake. Sudden and dangerous storms are likely to arise over the waters of all the major lakes.

Hydrography. The levels of the East African lakes are perceptibly sensitive to climatic fluctuations. Seasonal ranges of level are small: no more than a foot on Lake Victoria, 1.3 feet on Lake Albert, and three to four feet on Lake Nyasa. Longer term fluctuations, with consequential effects on the shorelines, are greater: during the 20th century the extreme range recorded on Lake Victoria has been 10.3 feet, compared with 17.3 feet on Lake Albert and 18.8 feet on Lake Nyasa. In each case the recorded maximum occurred in the early 1960s. The effects of drought are enhanced in small, shallow lakes: Lake Nakuru, in Kenya, dried up completely in 1939–40; at the end of 1949 Lake Rukwa was estimated to be one-fifth its normal size; and Lake Chilwa, in Malawi, suffered a drastic reduction of area in 1968.

Lake levels

There is a significant correlation between rainfall and variations in the level of Lake Victoria, and Lakes Kyoga and Albert follow—with an appropriate time lag—the conditions of Lake Victoria. In addition to rainfall, other factors affect the longer term fluctuations of Lakes Tanganyika and Nyasa. The Lukuga River outlet of Lake Tanganyika tends to become blocked intermittently by silting, consolidated by swampy growth. Similarly, in the flat valley of the upper Shire River, periods of erosion alternate with the building of bars of silt and sand reinforced by reeds. As a result, between 1915 and 1935 there was virtually no outflow from Lake Nyasa.

Many of the lakes, especially those in the Eastern Rift, are brackish, but Baringo and Naivasha are exceptions in that they are freshwater lakes. At the other extreme of salinity, Magadi is a soda lake, in which the continuing source of sodium carbonate appears to be alkaline waters of deep-seated origin. Lakes Edward and George have the highest salinities among the lakes of the Western Rift, but the alkalinity is not excessive. The problem of the deeper lakes, such as Tanganyika, Nyasa, and Kivu, is that their deeper waters are permanently deoxygenated and thus comprise a biological desert: 75 percent of the volume of Lake Tanganyika and 99 percent of that of Kivu are within this category. Moreover, all three lakes contain lethal amounts of hydrogen sulfide in their deeper waters, and Lake Kivu also contains vast quantities of methane.

Geochemistry of the lakes

Animal life. *Fish.* Among the main genera of fish in the East African lakes are the mouthbreeders *Tilapia*, much the most important in number of species and in total quantity; *Haplochromis*, which, like the *Tilapia*, belong to the Cichlidae family, a group of small perch-like fish; *Clarias* (barbel) and *Bagrus* among the catfish; *Hydrocynus* (tiger fish); various carps, including *Labeo*, *Barilius*, and *Barbus; Protopterus*, a lungfish; *Mormyrus*, a member of the elephant-snout fish family; and *Stolothrissa tanganicae* (*dagaa*), a small sardine-like fish.

Fish population patterns

The more strongly saline lakes, such as Nakuru, Elmenteita, Manyara, and, above all, Magadi and Natron, have a severely limited fish life. Lake Kivu also has a fish population that is neither varied nor abundant. Although fish are present in enormous quantities in Lake Rukwa, the number of species is not large, and the stock is dominated by *Tilapia rukwaensis*, an endemic species of *Tilapia*. Successive droughts such as that of 1949 explain why there are so few species in Lake Rukwa; the years immediately following 1949, on the other hand, provide an excellent example of the amazing recovery powers of tropical fish populations.

The majority of the lakes, though, have a rich and varied fish life, of which a high proportion of species are endemic to the individual lake. The Cichlidae, for example, are especially prone to form new species, and there are about 90 species of the family in Lakes Victoria and Kyoga.

Lake Albert has a fish life that is related to that of the Nile; it includes Nile perch, tiger fish, and *Polypterus* (bichir). The physical barrier of Murchison Falls, situated near the northern end of Lake Albert, marks the frontier of a separate animal province formed by Lakes Victoria and Kyoga. Several of the Lake Albert genera are not found in these two lakes, which contain many unique species. Similarly, the rapids on the Semliki River have prevented the recolonization of Lakes Edward and George—otherwise particularly rich in fish—by fish from Lake Albert. On the other hand, the presence of Nile perch, tiger fish, and bichir in Lake Rudolf serves to indicate its former connection with the Nile. The transplantation of fish by man, however, has caused a man-made zoogeographical revolution, the full effects of which cannot yet be discerned.

Game animals. The hippopotamus is ubiquitous around the lake shores, except those of Lake Kivu; the crocodile is also widespread, although absent from Lakes Edward, George, and Kivu, each of which is sheltered from the spread of this reptile by falls in the outflow river, with cool mountain torrents and sunless forest as additional deterrents. There is an inverse relationship between the density of game and that of human settlement, but in recent times the establishment of national parks and game reserves has encouraged the increase of game population. Among the variety of game to be seen in the neighbourhood of the lakes are elephant, buffalo, and various antelopes.

Birds. Among the resident and migrant birds in evidence, waterfowl are especially noticeable. Several of the Eastern Rift lakes—such as Nakuru, Elmenteita, and Manyara—are famous for their vast congregations of flamingos. Forming the basis of a national park in which the emphasis is on aquatic birds, Lake Nakuru is an ornithologist's paradise; Lake Edward and the Kazinga Channel are also notably rich in bird life. The fish-eating birds—cormorants, darters, and kingfishers—are also part of the ecology of the lakes.

Plant life. The vegetation setting of the lakes varies from the semidesert, in which Lake Rudolf is situated, to the patches of closed evergreen forest on the western and northern shores of Lake Victoria. Between the two extremes and in accordance with the position of the individual lakes, bushland and thicket, grassland, savanna, or open woodland occur. The oil palm, which is characteristic of West Africa and the Congo region, is found on the shores of Lake Tanganyika. Heavily populated and intensively cultivated areas marginal to Lakes Victoria and Nyasa present vegetation types much modified by man.

The lake shores may comprise open landscapes of headland or beach or may contain plants associated with swamps, such as the giant sedge, *Cyperus papyrus*, which is the most prevalent.

Resources and development. Fishing is one of the major resources of the East African lakes, but the degree of exploitation varies considerably. During the early 1960s, for example, less than 1,000 tons of fish, mostly for local consumption, was taken annually from the well-stocked waters of Lake Rudolf, and even in the early 1970s this lake remained virtually unexploited. To the south, however, the much smaller Lake Baringo yields about 600 tons of fish annually. The introduction of black bass and several species of *Tilapia* forms the basis of a commercial fishery at Lake Naivasha, which is also a popular weekend resort for Nairobi residents and which, like Lake Rudolf, attracts many sport fishermen. At Lake Magadi, on the other hand, soda ash, the leading mineral export of Kenya, is extracted and refined.

The fisheries of Lake Victoria are important principally as a source of protein for the rural and urban population living in the vicinity of the lakes; the fisheries of Lake Kyoga are strongly developed. Although Lake Victoria is considered to be essentially a *Tilapia* fishery, there are also important catches of other species. The industry is controlled by independent fishermen, typically using planked canoes with outboard motors.

Hydroelectric power

Another important resource at Lake Victoria is the 150,000-kilowatt Owen Falls power station at Jinja. In addition to providing electricity for use in Uganda and Kenya, a dam at the falls enhances the potential of Lake Victoria as a storage reservoir for the Nile River.

The varied fish life of Lake Albert includes Nile perch and two genera of tiger fish, of which one, *Alestes*, provides two-thirds of the total weight landed on the Uganda shores. Calculations indicate, however, that the catch is below the maximum sustained yield of the lake. With the collapse of the Congo market in the early 1960s, a greater proportion of the output was absorbed by a developing home demand. A tourist attraction adjacent to Lake Albert is the Ruwenzori National Park, which extends down to the narrowing outlet at the northern end of the lake.

Lakes Edward and George support remarkably fertile fisheries. The Zaire shores of Lake Edward are enclosed within the Virunga National Park (Parc National du Virunga), while the Ugandan portion of the lake lies within the Kabarega National Park. Close to the north-eastern shore of Lake Edward, Lake Katwe, which lies in one of the numerous shallow craters within the Edward–George branch of the Western Rift, is a source of salt.

The fisheries potential of Lake Kivu is more limited than that of the other lakes of the Western Rift. Tourist attractions are being developed, and, at the outlet of the lake below Bukavu, a hydroelectric station on the Ruzizi River provides power to as far away as Bujumbura (Usumbura) on the northern shores of Lake Tanganyika.

As a whole, the borderlands of Lake Tanganyika are characterized by underdevelopment. Fisheries are dominated by the exploitation of *dagaa*. There is, however, a secondary commercial fishery based on other species; in the absence of shallow, marginal waters, however, *Tilapia* do not form a significant proportion of the catch. Despite the changes in water level to which it is liable and its remoteness from markets, Lake Rukwa is third to Lakes Victoria and Tanganyika among the freshwater fisheries of Tanzania.

In the southern part of Lake Nyasa, shallow, sheltered waters provide a good habitat for *Tilapia* and for *Labeo mesops* (a species of carp). Non-African commercial fishing, however, has intensified the demand upon the inshore species, resulting in a reduction of the fish population. In the northern waters the fish caught include the catfish and some species of carp. The total annual catch in Lake Nyasa and the upper Shire River area is about 20,000 tons.

European reconnaissance of the lakes

Exploration. Following in the track of earlier Arab trading expeditions, the British explorers Sir Richard Burton and John Hanning Speke reached the eastern shore of Lake Tanganyika in 1858. It was there, at Ujiji, that David Livingstone, the Scots explorer-missionary who had pioneered the Shire River route to Lake Nyasa in 1859, was met by the Welsh explorer Henry Morton Stanley in 1871, after which the two subsequently made a reconnaissance of the northern part of the lake. Then, in 1874, in the course of his east–west traverse of the continent, the English traveller Verney Lovett Cameron explored the Lukuga River outlet on the western shore of Tanganyika.

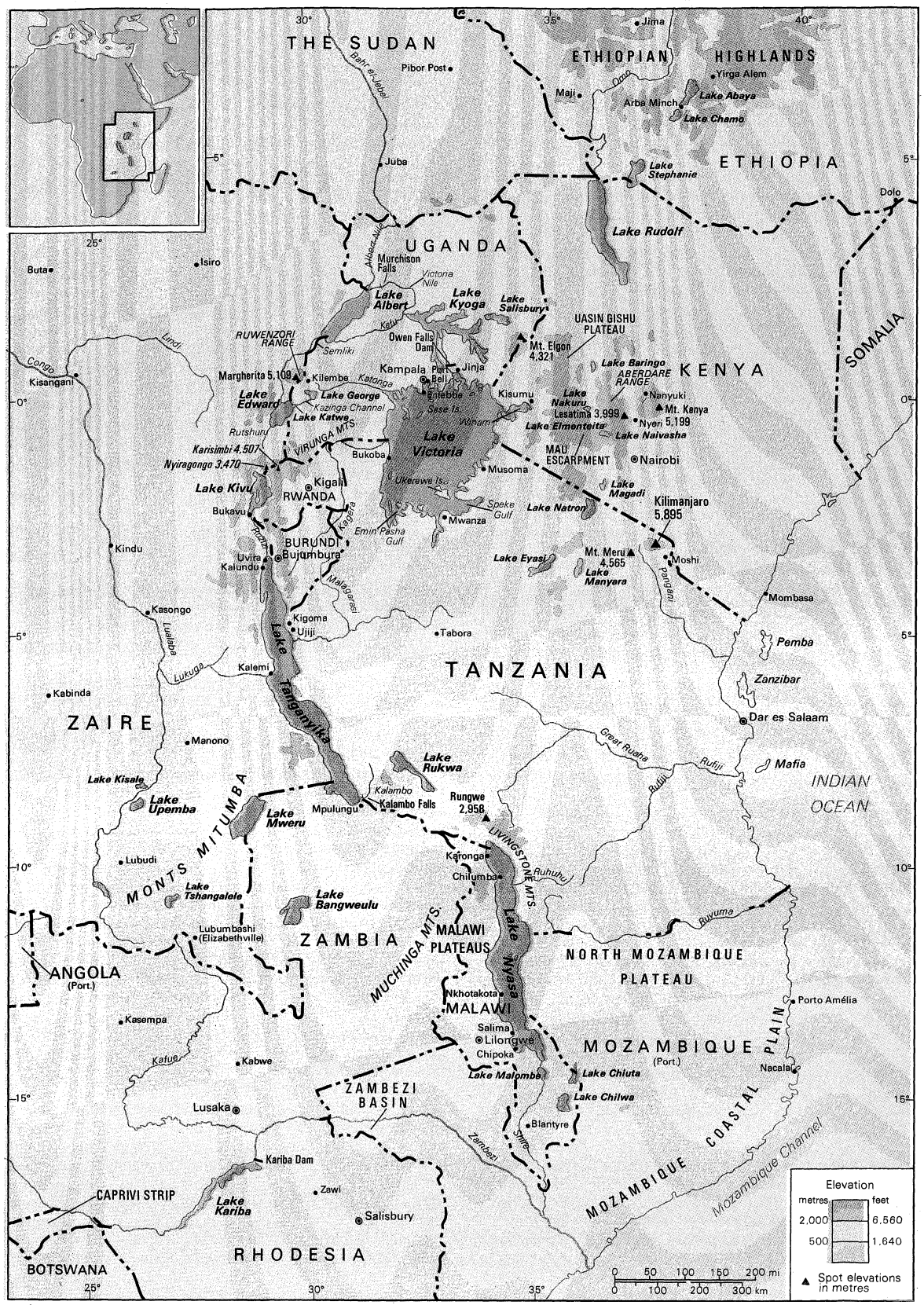
While Burton remained at Tabora in 1858, Speke journeyed northward and became the first European to view the waters of Lake Victoria. In 1862, after he returned to East Africa with a fellow countryman, James Augustus Grant, Speke stood above the Ripon Falls, which he proclaimed as the source of the Nile. The details of the Lake Victoria shoreline were filled in by Stanley when he circumnavigated the lake in 1875.

Meanwhile another Englishman, Samuel White Baker, together with his intrepid wife, had approached the East African lakes from the north, reaching Lake Albert in 1864. In 1888 to 1889 Stanley, in company with Emin Paşa, the German traveller, traced the Semliki River to its source in Lake Edward. Finally, in 1894, a German explorer, Adolf von Götzen, became the first European to visit Lake Kivu.

During the 1880s, Europeans explored the lakes of the Eastern Rift. Lakes Magadi and Naivasha were visited by a German traveller, Gustav Fischer, in 1883, and in the same year the Scots explorer Joseph Thomson reached the shores of Lake Baringo. Five years later two Hungarians, Count Samuel Teleki and Ludwig von Höhnel, reached Lake Rudolf.

Transport routes

Communications. Most of the East African lakes are reasonably accessible by road or air or both. Among the major lakes, however, Rudolf remains the least accessible by road. With the exception of Lake Albert, the main lakes of Uganda are well served by surface communications. There are good road communications to Lake Kivu, but access to Lake Tanganyika is more difficult, as is that to the eastern shores of Lake Nyasa. Scheduled inland-waterway services have been developed on Lakes Victoria, Tanganyika, and Nyasa; elsewhere,



Lakes and mountains of East Africa.

as on Lake Kivu, there are launch services. Otherwise, canoes and, to a smaller extent, dhows transport people and goods.

Fluctuations of the lake level and changing economic conditions have brought about the closure of a number of ports on Lake Victoria; the main ports in use are Kisumu, Musoma, Mwanza, Bukoba, Port Bell (serving the Kampala metropolitan area), and Jinja. Entebbe (in Uganda), which is no longer a lake port, is an international and regional airport. The East African Railways Corporation operates lake services on Victoria, including rail-ferry terminals at Jinja, Kisumu, Mwanza, and Musoma. Rail connections with the maritime ports are provided at the first three of these terminals. In the late

1960s the total lake trade—mostly in coffee, cotton, and cereals—remained well under 300,000 metric tons.

On Lake Tanganyika, the railways provide a service connecting Kigoma, terminus of the rail line from the Tanzanian capital of Dar es Salaam, with the Zambian port of Mpulungu. A Zaire company based at the railway port of Kalemi (Albertville) maintains links with Kigoma, Bujumbura (capital and lake port of Burundi), and Kalundu (for the Zairian town of Uvira). Malawi Railways, which provides services on Lake Nyasa, including the Tanzanian and Mozambique shores, has rail connections with Beira and Nacala on the coast of Mozambique. The lake rail terminal is at Chipoka, and the ports at which calls are most frequently made are Nkhotakota, Bhilumba (formerly Deep Bay), and Karonga, all on the western side of the lake.

BIBLIOGRAPHY

General: C.F. HICKLING, *Tropical Inland Fisheries* (1961), an ecological approach, with emphasis on the East African lakes; H.E. HURST, *The Nile: A General Account of the River and the Utilization of Its Waters*, rev. ed. (1957), on the hydrology of the Nile Basin, including its equatorial lakes; M.J. MANN, "Freshwater Fisheries," in W.T.W. MORGAN (ed.), *East Africa: Its Peoples and Resources* (1969), a brief account of the fisheries of the individual lakes of Kenya, Uganda, and Tanzania; ALAN MOOREHEAD, *The White Nile*, rev. ed. (1971), an account of the exploration of the White Nile between 1856 and 1900, containing much material about the East African lakes; S. and E.B. WORTHINGTON, *The Inland Waters of Africa* (1933), subjects of general interest related to the biology of the lakes of Kenya and Uganda.

Specific countries: Atlas of Kenya, 3rd ed. (1970); J.G. PIKE and G.T. RIMMINGTON, *Malawi: A Geographical Study* (1965), a general text that devotes much attention to Lake Malawi; *Traveller's Guide to the Belgian Congo and Ruanda-Urundi*, 2nd ed. (1956), contains useful topographical descriptions of the lakes that lie within Zaire, Burundi, and Rwanda; TANZANIA, SURVEYS AND MAPPING DIVISION, *Atlas of Tanzania* (1967); UGANDA, LAND AND SURVEYS DEPARTMENT, *Atlas of Uganda*, 2nd ed. (1967); R.H. TEMPLE, "The Lakes of Uganda," in S.H. OMINDE (ed.), *Studies in East African Geography and Development* (1971).

(S.J.K.B.)

East African Mountains

The East African mountains of Kenya, Tanzania, Uganda, Zaire, Rwanda, and Burundi are intimately related to a rift system, the fractures of which extend discontinuously between the Zambezi Valley and the Red Sea, and are flanked in many areas by highlands. Of the major mountains, all but one are of volcanic origin. Rising magnificently from the surrounding plateaus to altitudes of over 16,000 feet (5,000 metres), the highest peaks, despite their proximity to the Equator, are ice-capped.

Mt. Kenya, the Aberdare Range, and the Mau Escarpment are located wholly within Kenya to the north of Nairobi; Mt. Elgon lies astride the Uganda–Kenya border; Kilimanjaro extends along Kenya's southern boundary with Tanzania; and Mt. Meru is in northern Tanzania. The Ruwenzori Range stretches between Lakes Edward and Albert on the Uganda–Zaire border and, farther south, the Virunga Mountains extend along the contiguous borders of Uganda, Rwanda, and Zaire. (For a detailed discussion, see VIRUNGA MOUNTAINS. For associated physical features see EAST AFRICAN LAKES.)

The natural environment. *Topography.* The Aberdare Range, of which the highest peak is Ol Doiny Lesatima (Satima) at 13,120 feet (3,999 metres), and the Mau Escarpment rise steeply from the Eastern Rift. To the west, beyond the Uasin Gishu Plateau, Mt. Elgon emerges gently from a level of about 6,200 feet; but the spectacular cliffs of its western face dominate the lower plains of eastern Uganda, which lie at about 3,600 feet. The rim of Elgon's caldera (a vast explosion crater) is about five miles in diameter and contains several peaks of which Wagagai, at 14,178 feet (4,321 metres), is the highest.

Principal mountain features

A wide col, or saddle-shaped pass, separates the Aberdare Range from Mt. Kenya and forms the Nyeri–Nanyuki corridor. The second highest mountain in Africa, Mt. Kenya has a girth of about 95 miles at 8,000 feet, from which it rises boldly to its restricted summit zone. The craggy twin peaks of Batian (17,058 feet [5,199 metres]) and Nelion (17,022 feet [5,188 metres]) are closely followed in height by Point Lenana (16,355 feet [4,985 metres]).

Set amid low plateaus, Kilimanjaro is the highest mountain in Africa, rising to 19,340 feet (5,895 metres) at Uhuru peak on the Kibo cone. The generally smooth outlines of the cratered dome of Kibo are in marked contrast with the jagged form of Mawensi, or Mawenzi (16,890 feet [5,148 metres]); the two summits are connected by a seven-mile-long saddle, which lies at about 14,500 feet. Mt. Meru, about 40 miles southwest of Kilimanjaro, attains an altitude of 14,979 feet (4,566 metres).

The Ruwenzori Range runs parallel to the Western Rift, to which it drops steeply. The fall to the uplands of western Uganda, however, is more gradual. At its base the range is 75 miles long, and its greatest width is about 30 miles. The summit zone contains six distinct mountain massifs, which are separated by well-defined passes and deep river valleys. Mts. Baker and Gessi lie entirely within Uganda, while Mts. Stanley, Speke, Emin, and Luigi di Savoia form part of the Uganda–Zaire frontier. Of the ten peaks topping 16,000 feet, all but one are on Mt. Stanley, which includes the highest peak, Margherita, at 16,763 feet (5,109 metres).

The Virunga Mountains and their associated lava flows extend across the Western Rift. In the west, Nyamulagira, Nyiragongo, and Mikeno are in Zaire; Karisimbi—at 14,783 feet (4,507 metres) the highest of the Virunga volcanoes—and Visoke are centrally placed on the Zaire–Rwanda frontier; and farther east Sabinio (Sabinyo), Mgahinga (Gahinga), and Muhavura, also known as the Mufumbiro Mountains, are on the Rwanda–Uganda frontier. Not all of the cones culminate in craters, but several have crater lakes.

Glaciers

The relict glaciers that occur in the summit zones of Kilimanjaro, Mt. Kenya, and the Ruwenzori have little erosive force. Their more powerful predecessors, however, extended down to altitudes of between 12,000 and 10,000 feet and even lower and produced arêtes (sharp-edged ridges), cirques (glacial amphitheatres), rock tarns (rock basin lakes), U-shaped valleys, and moraines (boulders and other debris deposited by glacial action). Early glaciation also affected Mt. Elgon and the Aberdare Range. More than 30 small glaciers on the Ruwenzori together cover a surface of about two square miles, most of which is on Mts. Stanley and Speke; the lowest valley glacier descends slightly below 14,000 feet. Of the ten glaciers remaining on Mt. Kenya and covering less than 0.5 square miles, the largest are Lewis and Tyndall; the lowest tongue of ice reaches down below 15,000 feet. On Kilimanjaro, Kibo crater is strewn with giant blocks of ice, and the outer rims are covered with ice reaching down to 15,000 feet at the snout of the Penck Glacier on the wet southwestern moorlands. The 20th century has been marked in East Africa by a process of glacial retreat that has been rapid but neither constant nor conspicuous.

Geology. The peneplain of eastern Africa, dating from the Miocene Epoch (between 26,000,000 and 7,000,000 years ago), has been subject to a general elevational movement. The shoulders of the rift valleys have risen intermittently to produce highlands on which lavas ejected from fissures in the earth's surface have in some instances added considerable height. The most astonishing uplift is that of the Ruwenzori, the only East African mountains that are not volcanic. The ancient plateau surface of gneisses (coarse-grained, foliated crystalline rocks) and schists (foliated metamorphic rocks that can be split into thin flakes) was upfaulted on the west and upwarped on the east. Movements along the faults continue, and the Ruwenzori comprise an important earthquake epicentre.

Kilimanjaro is a volcano of complex structure and alkaline lavas situated at an intersection of fault lines. Shira was the first volcano of the group to become inactive, followed in turn by Mawensi and Kibo. Kibo retains its caldera—one and a half miles in diameter and 600

feet deep—within which there are successive inner cones and craters as well as fumaroles (holes or vents that emit gases).

The long-extinct volcano of Mt. Kenya has been much denuded, and the highest peaks consist of the crystalline nepheline–syenite (a granular rock of alkalic feldspar, nepheline, and other minerals), which plugged the former vent. Around this core are gently dipping lavas, agglomerates, and tuffs (rocks formed of compacted volcanic fragments).

Mt. Elgon belongs to the Eastern Volcanics of Uganda, which comprise soda-rich lavas and associated fragmental tuffs and agglomerates. The Western Volcanics are represented by the Virunga Mountains, of which Nyamulagira and Nyiragongo have been active during the 20th century. Major eruptions occurred in 1938 and 1948, and on several occasions a lava stream has reached the shores of Lake Kivu 12 miles south.

Drainage and soils. The Virunga Mountains separate the basins of the Nile and the Congo rivers and are the only East African mountains to form a drainage divide of continental stature. The entire drainage system of the Ruwenzori Range finds its way into the Semliki River, a tributary of the Nile. Because they are young, the mountain systems present good examples of consequent drainage (drainage that is determined by the initial slope of the land) such as the radial system of Mt. Elgon, in which streams radiate from a central area, and the parallel streams of the Aberdare dip slopes. The porous nature of volcanic materials often results in areas devoid of surface drainage.

The succession of soils is from the raw mineral type of the summit area, through the dark peaty loams of the Alpine zone and the strong brown loams of high organic content in the forest belt, to the ferruginous (iron-bearing) soils of the lower slopes. Volcanic material presents a range from the unaltered rock of the most recent eruptions to the well-developed fertile soils on longer exposed surfaces.

Climate. In a region of predominantly dry climate, the mountains are conspicuous as high-rainfall areas. Affected by the convergence of Indian Ocean and Atlantic airstreams, the Ruwenzori as a whole comprise the wettest and the cloudiest of the mountains, where moist conditions penetrate up to the peaks and about 80 inches (2,000 millimetres) of rain annually fall at 15,000 feet (5,000 metres). Kilimanjaro, Mt. Kenya, and Mt. Elgon are affected by their position in relation to southeasterly and northeasterly airstreams. Below 10,000 feet they are wettest on their southeastern and southern sides, with annual totals rising to 100 inches on Mt. Kenya; they are driest on their northern flanks, with less than 40 inches yearly. Rainfall decreases above a cloud ceiling at about 10,000 feet, especially on Kilimanjaro, where the Alpine zone is a veritable desert. On the western and southwestern slopes the diverted southeast trade winds are sucked up each of the mountains as westerlies, creating an increase of cloud and precipitation. At the summit of Mt. Kenya, temperatures seldom rise above the freezing point, and above 14,500 feet precipitation mainly occurs as snow.

The plant life of Mt. Kenya

Vegetation. The altitudinal succession of vegetation is well developed on Mt. Kenya. The succession emerges from the surrounding savanna (grassland with scattered bushes or trees) and begins on the lower slopes with a crescent of cultivated land. The montane forest extends upward from a lower limit of 5,000 to 7,000 feet to about 10,500 feet and includes giant trees, such as camphor and various figs, cedar, yellowwood, and the East African olive. From about 8,000 feet, the forest comprises montane bamboo, and at its upper limit parkland and low thicket fringe the succeeding zone of giant heather. At 11,000 to 12,000 feet, the heather zone gives way to the Alpine zone in which tree groundsel and the giant lobelia rise out of a ground vegetation of tussocky grassland and everlastings (composite plants, the flowers of which can be dried without loss of colour or form). Mosses and lichens survive up to about 15,000 feet, but bare rock and ice are exposed above that height.

The montane forest of Kilimanjaro is drier than that of Mt. Kenya. Bamboo is virtually absent, although it is abundant on neighbouring Mt. Meru, and there is no parkland zone. The heather zone is strongly represented, whereas the Alpine semidesert is poor in flowering plants. Mt. Elgon reaches into the Alpine zone, as do the summits of the Aberdare Range. On the northwest of the Ruwenzori, the lower slopes touch upon the equatorial forest, and the vegetation is moister and more luxuriant than that of the eastern mountains. Above the bamboo forest and the wooded parkland, the Virunga Mountains extend into the heather zone and, in the three highest volcanoes, into the Alpine zone.

The Afro-Alpine vegetation of the eastern mountains is unique. With the increase of temperatures in post-Pleistocene times (about 10,000 years ago) the cold-loving plants retreated to the mountains where they have been preserved and somewhat transformed. Despite the enormous distances that separate the mountains, plants in the respective Alpine zones are closely comparable. There are lobelia and *Alchemilla* (lady's-mantle) species common to all the mountains, although the tree groundsel species are limited within neighbouring mountains. The phenomenon of giantism is common, while dwarfism occurs at the highest altitudes.

Animal life. Elephants, rhinoceroses, buffalo, antelope, hyrax (one of several small, hooved mammals), bush pigs, and monkeys, including the black-and-white colobus, are among the main inhabitants of the montane forest. The bongo (forest antelope) and the giant forest hog have not been recorded from Kilimanjaro perhaps because of its lack of bamboo forest and its isolation from the mountains to the north. The golden monkey occurs in the Virunga Range and the chimpanzee in the Ruwenzori. Trout have been introduced into the streams of the more accessible mountains.

Mammals of the upper forests, including the leopard and antelope such as the duiker and the eland, penetrate into the moorland and Alpine zones, where the hyrax and the groove-toothed rat are the most obvious inhabitants. Birds include the lammergeier (one of the largest birds of prey, resembling the eagle and vulture), the mountain chat (a songbird), and the scarlet-tufted malachite sunbird (a small, brilliantly coloured songbird). Animal life, like the vegetation, shows resemblances that suggest a retreat of its distribution from the surrounding plateaus to the montane islands of refuge.

The human imprint. *Population.* Population is confined to the lower slopes, with the upper limit of settlement at about 7,000 feet. Bananas and millet are common subsistence crops, and coffee is an important cash crop. On the southern and eastern slopes of Kilimanjaro the Chagga have long used an effective system of irrigation based on the ridge and valley relief of the mountain.

Population densities

High population densities among the Chagga are matched by those among the Kikuyu and related groups on the slopes of the Aberdare Range and around the southern and eastern margins of Mt. Kenya. The Gisu have densely settled the western slopes of Mt. Elgon below the forest zone.

Population density is high on either side of the northern nose of the Ruwenzori; the Konjo hillmen live mainly on the eastern flank of the range. The pygmy Twa occupy the forests of the Virunga Mountains, the lower slopes of which are cultivated by the far more numerous Hutu.

Exploration. In 1848 the snow-covered summit of Kilimanjaro was observed by the German explorer Johannes Rebmann and in the following year Johann L. Krapf obtained a view of the snows of Mt. Kenya. In 1888 the Welsh explorer Henry M. Stanley glimpsed the Ruwenzori through a break in their cloud cover and equated them with the Mountains of the Moon of Ptolemy.

A number of expeditions on Kilimanjaro preceded that of 1889 in which the German geographer Hans Meyer conquered the summit of Kibo. Partial ascents of Mt. Kenya were made before H.J. Mackinder reached the summit of Batian in 1899. The Ruwenzori awaited the large-scale expedition of the Duke of Abruzzi in 1906

for the conquest of their major peaks, and in 1907–08 the Duke of Mecklenburg made a comparably comprehensive survey of the Virunga Mountains. The Elgon caldera had been visited in 1890, and in 1911 Wagagai was climbed. In 1932 a Belgian scientific mission explored the western slopes of the Ruwenzori and climbed several peaks.

Since 1931, air photography has assisted in the production of excellent maps of the major mountains. Under the stimulus of the International Geophysical Year of 1957–58, glaciological expeditions were mounted on the Ruwenzori, Mt. Kenya, and Kilimanjaro.

Kilimanjaro, Kenya, and Ruwenzori are names of African origin; the first two are of somewhat uncertain meaning, and Ruwenzori may come from Lunyoro words meaning "place of rain." Elgonyi was the Masai name for Elgon, but the Gisu name Masaba remains current in the Bantu languages. The names Kibo and Mawensi are of African derivation; and the highest peaks of Mt. Kenya were named after Masai chiefs. With one exception, Ensonga, the major mountains and peaks of the Ruwenzori have European names, while Virunga nomenclature is African.

Transport. The Ruwenzori form a barrier to east–west transport, and road connections between Uganda and Zaire pass south of the range. The Mau Escarpment and the Aberdare Range deflect the main roads to their northern and southern flanks, but each is crossed by a secondary road. The Nyeri col between the Aberdares and Mt. Kenya is used as a gateway to the north. There are circular routes around Mts. Kenya and Elgon, and the bases of all the mountains are reasonably accessible.

Resources. Copper ore is mined at Kilembe, Uganda, on the southeast flank of the Ruwenzori, where the narrow margin of ore reserves is a precarious factor in relation to future output. A hydroelectric installation in the nearby Mubuku Valley provides power for the mines.

Forestry and agriculture

The closed forests are mainly under forest reserve and are classed as presently productive of timber. The most important output of sawn wood has come from the more accessible mountains of Kenya and Tanzania, where cedar, podo (or yellowwood), and camphorwood are among the principal timbers. Land cleared for cultivation in the lower part of the forests forms rich agricultural zones in which a considerable variety of crops, including coffee, tea, wheat, pyrethrum (a chrysanthemum used in the production of insecticides), bananas, millet, root crops, and vegetables may be grown. Cattle are reared on the northern slopes of Mts. Kenya and Elgon and on Kilimanjaro. The narrow agricultural belt could be widened, and there is a tendency for the extension of cultivation into lower altitudes by the use of irrigation.

The mountains are an important tourist attraction. The whole of Mt. Meru is within Meru National Park, and other parks include Mt. Kenya above 10,500 feet, the moorland zone of the Aberdares, and a sector of the Kenya portion of Mt. Elgon. In the Uganda section of the Virunga Mountains, the Gorilla Game Reserve is situated on the northern slopes of Mts. Muhavura and Mgahinga (Gahinga). In the Rwanda and Zaire portions of the Virunga Mountains, gorillas are protected, respectively, in the Volcanoes National Park and the Virunga National Park (Parc National du Virunga).

BIBLIOGRAPHY. RENE JEANNEL, *Hautes Montagnes d'Afrique* (1950), extends its coverage to Ethiopia, but is chiefly important for the mountains of Tanzania, Kenya, Uganda, and Zaire. I.C. REID (ed.), *Guide Book to Mount Kenya and Kilimanjaro*, 2nd ed. (1963), is primarily intended as a rock-climbing guide to the central peaks; also contains information on the plant and animal life; PETER ROBSON, *Mountains of Kenya* (1969), is a guidebook designed primarily for the hill climber. *Tanganyika Notes and Records*, no. 64 (March 1965), is a special issue devoted to Kilimanjaro that contains an extensive bibliography. The *Traveller's Guide to the Belgian Congo and Ruanda-Urundi*, 2nd ed. (1956), provides a useful topographical description of the mountains within Rwanda and Zaire. X. DE GRUNNE *et al.*, *Le Ruwenzori: vers les glaciers de l'Equateur* (1937), contains contributions on the geology, physiography, plants, and animals on the East African mountains as a group.

(S.J.K.B.)

East Asian Cultures

East Asia as a cultural entity is defined in this article as comprising the Chinese and groups that have been derived from the Chinese ethnically or culturally and, secondly, those cultures that were greatly influenced by China at the time they became states. Among the first category, of course, are the Chinese themselves, the Taiwanese, and, with some equivocation, the Miao and Yao. The second category comprises such peoples as the Japanese, Koreans, Manchus, and Minchia. Other peoples have also been influenced by the Chinese, but, because they have not formed states of their own, this influence is difficult to determine and is often not emphasized. This is the case among, for example, the Nakhi of southwest China and the Ainu of Hokkaido.

Recognition of ties with China is the most characteristic feature of East Asian cultures. It unites peoples as disparate in many aspects of life as the Hui-hui of China's northwest, the Japanese, and the Manchus. The area thus has a special significance for many of the peoples themselves and is not simply a construct of social scientists.

It is to the south that it is most difficult to determine the boundary of East Asia. Here groups escaping the domination of Chinese power fled through China and finally into what are now designated as the states of Southeast Asia. Some of these peoples are closely related to others still within China. In addition, some peoples such as the Tai and Tibetans have been greatly influenced by the Chinese; but then again they have both been greatly influenced by the Indians. The Vietnamese are not only the most Sinified of the Southeast Asians but probably give the Chinese undue credit for the beginnings of civilization in the Red River Delta of North Vietnam. Vietnam is generally discussed in Western accounts in the context of a Southeast Asian culture area and is not discussed in this article (see SOUTHEAST ASIAN CULTURES).

To the north, the border of the East Asian cultural area is marked by the Great Wall of China, which passes within 35 miles (55 kilometres) of Peking. On the other side of the wall live those who, from about the 7th century BC on, have rejected the East Asian way of life.

GENERAL CONSIDERATIONS

People and languages. Some peoples in Asia more or less form a buffer zone around East as well as South Asia. These ethnic groups are of diverse origins and cultures. The Ainu, situated on the island of Hokkaido, appear to represent in ethnic terms a conservative autochthonous refugee group that cannot easily be connected with any other group. They are linguistically distinct from other groups but culturally akin to the peoples of the Soviet Maritime Province and Eastern Siberia. The Manchus, Xojen, Daghur, and Mongours are small ethnic groups situated along China's northern and northwestern frontier. The first two speak a language of the Tungusic branch of the Altaic linguistic family, while the Daghur and Mongours speak Mongolian languages, also of the Altaic family.

The Islāmic Hui-hui

The Hui-hui (also called Tungan in Chinese Turkistan and Panthay in southwest China), or Chinese Muslims, take their ethnic identity from their religion. Those aspects of life in which they differ from Chinese are related to their practice of Islām, although they also exhibit a greater esprit de corps and competitiveness. There is a natural tendency for members of an ethnic unit who can communicate readily with members of the dominant group to merge with that group, and many Hui-hui have seen themselves as predominantly Chinese. But it is also true that Chinese Muslims have regarded themselves as an alien and superior people to the Chinese.

Extending eastward from the Tibetan ranges and plateau, the Tsinling Mountains form a dividing line between border groups of Ural-Altaic linguistic affiliation to the north and Sino-Tibetan, Tai-Kadai, Austronesian (Malayo-Polynesian), Austro-Asiatic, and perhaps independent linguistic affiliations to the south. Most of the non-Chinese ethnic groups found along the borders of Tibet and China speak a language affiliated to the Tibeto-Burman branch of the Sino-Tibetan linguistic family.

The Lolo and Minchia have occupied the highland area of southwest China for over 1,000 years. A number of Lolo traits—among them a language that indicates an affinity to Tibeto-Burman, a tall stature with a high-bridged nose, the use of felt, a mixed pastoral and agricultural society, and the use of poisoned arrows—appear to indicate a culture with diverse and distant origins. The Lolo seem to be most alike in culture, language, and physical type to the peoples of southeastern Tibet. They appear to have originated in the same borderland that they now occupy, although perhaps more to the west.

The Minchia have been a part of a centralized political system since the 7th century, when their principal city, Tali, was the capital of the Nan Chao state. The Yüan dynasty emperor Kublai Khan conquered Nan Chao in 1253, but the area around Tali is to this day primarily inhabited by the approximately 500,000 Minchia. The Minchia have been rice cultivators for over 1,000 years and now are highly acculturated to Chinese ways of living. Many speak Chinese as well as Minchia.

The linguistic affiliation of the Miao and Yao to major linguistic groups is still in dispute, but there is no doubt but that the two are linguistically related to each other and have borrowed heavily from Chinese. Miao names for themselves are Hmong, Hmung, or Hmu, followed by a further word indicating subethnic affiliation. The Yao refer to themselves as Kim Mien, Kim Mun, and Yu Mien. The Miao number about 2,500,000 in China and 200,000 in Southeast Asia. The Yao number more than 600,000 in China and more than 100,000 to the south. Both derive from the Yangtze Valley region and, in the case of the Yao, probably from the mountains along the east coast of China. They appear to have been recognized by the Chinese as ethnic units for over 2,000 years. Both are distributed for the most part in small enclaves sometimes no larger than a village, although there are also larger settlements. The People's Republic of China has recognized a number of Yao autonomous counties (*hsien*), and an autonomous special administrative district (*chou*) has been set aside for the Miao in Hunan Province, as well as two autonomous regions in Kweichow Province. The sociopolitical fragmentation of the Miao and the Yao has apparently been caused by their countless attempts to escape domination and assimilation by the Chinese. The last independent Miao group disappeared at the end of the last Miao rebellion in 1870. Both Miao and Yao have drifted southward, and now both groups are found as far south as Thailand.

Ethnic and language groups in Taiwan

The original inhabitants of Taiwan (Formosa) were Austronesian (Malayo-Polynesian) speakers, but, within the last 300 years, after the establishment of trading posts on the island by the Dutch, the Austronesian speakers have been gradually pushed back into the mountains, in part by Hakka Chinese from Kwangtung Province but mostly by Minnan (Amoy) Chinese from Fukien Province, directly across the Formosa Strait. There are now about 12,000,000 Taiwanese. The majority speak one of several dialects of the Fukienese languages, a Sinitic tongue. A minority speak Hakka, another Sinitic language. Many mainlanders in Taiwan speak Mandarin Chinese. Within the last 25 years many Chinese have come to identify with the family, associational, and professional ties that they have established on Taiwan. As a result they and especially their children have come to associate more with Taiwan than with China.

Japanese is a composite speech in which Austronesian was the earliest ingredient. A later Altaic influence came to overlay Austronesian forms, and many Chinese words have subsequently been borrowed.

Korean linguistic affiliations seem to parallel those of Japan but with different emphasis. Altaic, specifically Tungusic, linguistic influences are very strong, and cultural influences from the north and east have been more prominent. The Chinese, for example, conquered Korea and established military colonies in the 1st century BC, one such outpost lasting until the 4th century AD.

The dominance of China. By definition the peoples of East Asia turn toward Chinese culture but with varying degrees of ambivalence. Perhaps the Koreans, Manchus, and Minchia have been the more assiduous students. One can contrast the attitudes of these nationalities with the open defiance and antagonistic acculturation maintained by the independent Lolo of the Ta-liang Shan and the belief of the Japanese from the 10th century on that they should be treated as equals rather than clients by the Chinese. On the other hand, such groups as the Miao, Yao, and some Lolo speakers lost their ability to control their affairs to any large extent beyond the village level, thus tending to accept the Chinese great tradition. In early China may be seen an approximation of the feudal system of Europe. Men who held political power and an authoritative position were overlords and not simply bureaucrats. The political structure was that of a lineage state; that is, it was considered that everyone, or at least everyone of importance, within the state was related. Ideally, therefore, a vassal was not only a subordinate—a client—but a relative as well. The lord should be a more direct descendant, a member of a senior generation, or in some way more closely related to the founder of the line than the vassal. If the kinship relationship between two men who wished to establish a vassal and lord relationship was not close enough, it appears that genealogical charters could be altered to fit social and political needs. The regularized system of interstate relationship, called tribute in the literature, between China and other countries was modelled upon a system of exchange between patron and client that existed in feudal China. According to the world view by which the tribute system operated, there was only one true universal culture—that of the Chinese. The tribute system was the means of emphasizing this fact. It consisted of a number of ceremonial acts, the most famous of which was the full kowtow, in which the ambassador from the foreign court prostrated himself full length in front of the emperor or the latter's representative. He thus expressed not only submission and subservience in military and civil affairs but also obsequious deference in cultural, ceremonial, and ethical matters. The tribute system was a means of carrying on diplomacy between the dominant power—that is, the "Middle Kingdom"—and other less equal powers and civilizations. Tribute was mostly an exchange of gifts, the Chinese often giving more in value than they received. The important element was the bringing of the various peoples together under the common umbrella of customary law and ritual, or attitudes toward virtue. The adoption of the Chinese ideographic system, calendar, Classical literature and art, court ceremonies, methods of government and law, and even dress and music were further aspects of the Chinese cultural hegemony.

The tribute system

The crucial period for the development of China and also East Asia was after the fall of the Han dynasty, about AD 220. Han society was comparable to Rome in its general complexity of organization and in its decline into equally fragmented pieces. Yet a central power was to reappear in the 7th century. The reasons for this revival were several, but the most important was that although the barbarian areas, especially those that threatened China, could fight well and at times conquer China, they never emulated it to the same extent that the Western barbarians emulated Rome.

There is some agreement that the pre-modernization periods in East Asia can be called "traditional" ("classical" has a different connotation and should refer to a period preceding the traditional). The traditional period has lasted a long time, not without change but with essentially the same types of social structures. In China the traditional period began during the T'ang dynasty; that is, during the 7th century AD. In Korea it began with the Yi dynasty in 1392 and in Japan with the Tokugawa shogunate in 1603.

TRADITIONAL SOCIAL STRUCTURES

It would seem that the peoples of East Asia originally had much the same type of social organization but that each group has developed a somewhat different world view.

Kinship patterns. The common bases upon which the luxuriant growths of East Asian social structure grew

are few in number and relatively simple. One is the tendency to maintain extended patrilocal families; that is, for sons to bring their wives to live within the same household or, if not, at least close enough for all descendants to form a single social, economic, and ceremonial unit during the lifetime of the father and very often the mother. Due to family or sometimes circumstances of wider importance, it may not be practical or even considered desirable for sons to live and work together, but almost always one of the children will be on hand to care for the aged parents. Sometimes, if there are no sons, a man will marry into the family, the husband of a daughter living with his wife's family.

Whether it is a woman or man who marries in, the effect seems to be the same. The spouse, or his or her children, tends to establish an identity for inheritance and descent through residence in the house into which they marry or into which they are born.

Main characteristics of the kinship system

Another basic element of East Asian social organization is the ego-centred kinship system. In such a system an individual does not place himself within a kinship group by reference to a group in which he has membership but rather calculates the distance of each relative from himself. He is thus the centre of the kin group, and those too far on the periphery cease to be relatives.

This ego-centred system has two forms. One is the kindred form, the type of kinship system found in the West, in which relatives on both the mother's and father's side are equally considered as kin. The second form may be regarded as an asymmetrical extension of the kindred form. This is the patrilineal segmentary lineage, which emphasizes the relationship between the parent and child on the father's side of the family. The line of descent from grandson to great-grandfather is thus given prominence.

Characteristic of the segmentary lineage is a tendency for residence to equal kinship affiliation. Thus, most of those who live within the same village or encampment are believed to be related. Members may move out to form colonies, however, and in this way lose their original identity. Other individuals and even families move into the residential unit, and in time they or their descendants are accepted as relatives. Within any residential area, however, there is a core group of old families, so to speak, toward whom the others gravitate and claim relationship of some sort.

The last basic element of East Asian social organization is ancestor worship, especially of one's patrilineal ancestors. Shamanistic religious practices often related to ancestor worship are prominent (see below *Traditional belief and aesthetic patterns*).

Apparently most of Europe and northern Asia at one time maintained a social organization of the type depicted above. It is also found in northern and eastern Africa. East Asia is exceptional because not only have many of the features of this model been maintained until recently, even down to the present, but because they were also accompanied by the rise of a centralized state.

Sociopolitical organizations. As stated above, Chinese civilization began as a lineage state (one in which political subordinates were ideally relatives of inferior rank in the lineage) and one in which the patrilineal segment of a lineage was of particular importance. Japanese political structure was also a political system tending toward a state based upon the segmentary lineage. It is almost impossible to maintain a strong centralized system of government along with a segmentary lineage social organization because forces for centralization are countered by those for segmentation. In China this "familial feudalism" came to an end with the Ch'in dynasty (221 BC), when a bureaucratic, authoritative system of government was introduced for all China. Although China subsequently became fragmented, indeed for a time became a feudal society, it never reverted to the lineage state as a form of government.

Japan entered into its bureaucratic period under the impact of Chinese influence, which was at its height during the Taika reform era beginning in 645. By the 11th century, however, during the Fujiwara period, local family-centred warrior bands held Japan in de facto control. With the rise to power of Minamoto Yoritomo (died 1199), the welding into a national state of local centres of power based on residence and descent groups reached its climax.

Korean sociopolitical organization managed to combine a number of seemingly conflicting factors as late as the 12th century: a bureaucratic structure, in part oriented around a Chinese-like civil service examination system, combined with a very class-conscious society in which local power rested largely in the hands of aristocratic estate holders. The residents on these estates—slaves, serfs, and nobles—were all united by common genealogical charters. Unlike Shang dynasty China, however, these lineage estates did not integrate to form a lineage state. Linguistic evidence for medieval Korean social organization can be found in the use of place-names along with patronymics. There are, for example, 84 place-names that qualify the Kim surname.

The traditional social structure in all three nations is similar in one respect: although segmentary lineages continued to exist, they were unable to dominate and determine the organization of the state. Nor can these be called feudal societies.

Significance of the tax system

Taxation and tax reforms and their relation to land tenure were important factors in traditional Chinese society. Reforms were initiated by the Northern Wei dynasty, which controlled North China from AD 386–534. At first, the reforms consisted of limitations to the amount of land any individual could control and of qualifications on inheritance rights. During the T'ang dynasty (618–907) this system was changed to a direct and, as compared to the past, heavy tax on land. The emphasis shifted from taxing each person equally, as in a poll tax, to a tax on resources. As a result, large estates apparently were broken up into small, independently owned farms and relatively small landlord-owned estates.

A second factor of great importance in shaping Chinese traditional society was the commercialization and urban orientation that developed in China. The percentage of persons living in urban centres in China even today is small, but by at least the Ming dynasty (1368–1644) and probably already by the Sung (906–1279) the influence of city life upon the nation was very strong.

Gentry society. This period in Chinese history is usually referred to as that of a "gentry society," a term of considerable ambiguity. On the one hand, one can speak of the scholar gentry or, more accurately, the degree-holding gentry—those called *shen-shih* in Chinese. Such individuals were granted a special legal as well as social status because of their success in the examination system. Some few members of this group, by continuing up the degree ladder, received bureaucratic appointments. These may be called the bureaucratic scholar gentry. Then there were others who, because of their position as landlords, may be termed landlord gentry.

To understand traditional Chinese society one should translate gentry simply as important and influential persons and then attempt to understand in what manner their pre-eminence was obtained and by what means their influence was exerted. Government service, the examination system, education, landed wealth, wealth in general, tenantry, and social mobility can best be examined as interrelating within a system of government, as well as of society. The summit of the system was the Imperial court supported by a minute elitist bureaucracy. The court and civil service system were organized as though the government was an autocratic, although benevolent, despotism. In a formal sense this was true. There was little an individual could do to counter the will of the central government if he met with its full displeasure. It is not fully understood how the theoretical or formal structure of the government, which can be deduced from examining old statutes, actually operated. It is, however, evident that the government of the various dynasties depended upon a client–patron system of structured relationships and on informal government to maintain stability in the countryside.

By late Ch'ing (1644–1911) and Republican times this relationship was called *kan-ch'ing*. To have good *kan-ch'ing* is to have a sense of well-being, a feeling of being at ease, especially in relationships between parties of unequal status who are not kinsmen. Those who have good *kan-ch'ing* are in a position to help each other rather than to compete with each other. In other words, it is a personalization of otherwise contractual relationships.

The personalization of contractual relationships

Landlord–tenant relations. Despite the superficially despotic central government, the landlord was primarily dependent upon the goodwill of the tenant if he was to get his rent. It is true that, should the tenant refuse to pay the rent, the landlord could ask for help from the magistrate, but such aid, in the short run at least, might prove to be very costly, as the landlord would be saddled with all court costs besides having to satisfy those greedy lower officials upon whose goodwill he had made himself dependent. Thus it was better for both parties, tenant and landlord, to keep the central government at a distance—especially military officials. The tenant, of course, normally paid the rent. For him not to do so and present no excuse would be tantamount to rejecting the landlord's right to the land. But the actual amount to be paid was open to negotiation. If drought, flood, too much rain at the wrong time of year, insect plagues, or other factors adversely affected the harvest, it was the landlord's obligation to remit as much of the rent as necessary for the tenant to live. The tenant was in a better position to know the actual growing conditions and the real condition of the harvest. If good *kan-ch'ing* existed, the landlord would not press overly for the rent, and the tenant would be fairly honest about paying it in good years.

The lack of security in rural areas meant that good relations between landlord and villager were necessary to protect the landlord against bandits. In an area in which the gentry were well respected, they were able to maintain an effective militia (mostly composed of villagers) for the defense of the village and their own property. But if the villagers hated the landlord, residents of the rural area where he lived might offer him no protection against bandits coming from a neighbouring area or might simply attack him themselves.

More symbolic contributions made by tenants to the landlord consisted of gifts of food and produce presented at festivals. During rites of passage (at birth, marriage, death, etc.) in the landlord's house, tenants as well as freeholders would contribute labour. The landlord would make monetary contributions to his neighbours when they needed help. Tenant children often worked as domestics in the "big house," and sometimes a landlord would take a tenant girl to be his "little wife" (second, subordinate spouse). The tenant was courteous to the landlord, calling him "master" or "elder."

In addition to forgiving rents at times of distress to the tenant who fulfilled his obligations, the landlord acted as an intermediary between the rural area and the functionaries of the central government. He had the poise, education, and background that made it possible for him to argue the case for an accused villager or attempt to have the tax rate adjusted or obtain a *pai-lou* ("memorial") in honour of a faithful widow or earnest local historian. It was also expected of landlords to obtain financing for bridges, schools, temples, and other semi-public buildings. With the aid of the village headman, the landlord organized the maintenance of public works by means of corvée labour (labour exacted in lieu of taxes). At the village level this was probably more like a series of days of cooperative community effort than unrewarded drudgery. This unofficial body of village headman and landlord was also responsible for morality in the community. They often helped to work out a compromise in a dispute or even acted as judges. During the Ch'ing dynasty they were responsible for bimonthly public readings of the emperor K'ang-hsi's "Sacred Edict" (*Sheng yü*):

The "Sacred Edict" of K'ang-hsi

> Be filial and respect the social relationships, be frugal and diligent, esteem scholarship and eschew unorthodoxy, be law abiding and pay your taxes.

Sometimes local children were taught along with the children of the "big house" by a tutor, or a school was established elsewhere for the neighbourhood, usually at the Classical (Confucian) temple. These illustrations indicate only some of the functions of the client–patron system of government.

The burdens of this system lay initially upon the tenant but tended to drain the resources of the landlord. It is difficult to imagine a landlord maintaining good *kan-ch'ing* and making a great deal of profit from his landholdings. But, by investing in urban-based enterprises, such as pawnshops, grain shops (in which a surplus could be sold at times of greatest demand), and small industries, the gentry could recoup its losses.

It was difficult to become wealthy through the ownership of land alone and thereby to acquire gentry status because of the rules of equal inheritance by all the male heirs. Wealth was often initially acquired in the cities or, if acquired in the countryside, frequently by such dubious means as banditry. Later this wealth was transferred to land both for economic security and for social prestige.

Families and lineages. The extended families of the "big house" were complex organizations including often only remotely related individuals seeking patronage. The peasant families, on the other hand, tended to be small, approximating a nuclear family (parents and children), although the large, extended pattern was the ideal. For many the cyclical joint (nuclear-extended) family corresponded closest with reality. This pattern, which is a basic feature of East Asian social organization, assumes the dissolution of the extended family after the death of the parents. To the individual the pattern is cyclical, as he is born into a household composed of grandparents and their married sons, which breaks up into nuclear households upon the death of the grandparents. As sons mature it is recreated, only to dissolve with the father's death.

In South China, especially in the provinces of Kwangtung, Kwangsi, and Fukien, the client–patron relationship was solidified by the continued power of segmentary lineages, the *tsu*. In Taiwan large lineage organizations were rare. Most individuals of Chinese descent resident on the island maintained only shallow lineage organizations and remembrances of ancestors.

The kindred form of kinship organization was an element within Chinese and thus Taiwanese kinship organization, but it was not as important for the Chinese as it was for the Koreans and certainly not as important as it was for the Japanese.

Korean Yi dynasty society developed a gentry similar to that of China, but, while in China the importance of the gentry lay in their function as community leaders in villages where they owned only a small percentage of the land, in Korea the gentry (known in Korea as *yangban*) owned most of the land. Koreans possessed a stronger class consciousness, and land ownership was limited mostly to the bureaucratic class.

In recent times at least, the Koreans have tended to practice primogeniture rather than equal inheritance by all sons. The joint family described for the Chinese has thus not been common for the Koreans, although the eldest son, after receiving his inheritance, is responsible to some extent for the welfare of his younger male siblings. The result is a weakly united extended family, a family that does not live together but still feels responsible for its members. The system of kinship terminology indicated by Korean terms of address is much like that of western Europe. It indicates the importance of the nuclear family and bilateral (paternal and maternal) kindred, while the Chinese terms of address indicate an emphasis on patrilocal residence. Other systems of kinship terminology in Korea show the importance of patrilineal ties, as well as the former presence of segmentary lineages.

Tokugawa Japan (1603–1867) was a highly bureaucratic state working within a feudal framework. During this time there were probably as many civil servants in Japan as there were in all of China at the same period, despite the difference in size and population. Admittance

to the bureaucracy was limited to the former warrior class, Samurai or *bushi*, who were placed on stipends in times of peace and whose caste position was protected by law.

Because of the exercise in Japan at this time of tighter political controls, especially vigilance by the Tokugawa and daimyo (hereditary feudal lord) for signs of disloyalty, which might be punished by confiscation of land, there was less need for a client–patron system to emerge, although the Japanese form of the segmentary lineage, the *dozuku*, only gradually declined in importance. A client–patron relationship existed naturally between the daimyo-executive and *bushi*-bureaucrat, but this relation varied from area to area. It tended to be thought of more in terms of formal obligations and efficiency rather than as one based on personalities. While in China there had been a personalization of contractual relationships, in Japan there was a ritualization and formalization, if not to say contractualization, of personal relations.

Socio-economic change in Tokugawa Japan

The Tokugawa period was one of increasing urbanization and commercialization. There was a development of small industries and an improvement of agricultural techniques, based in part upon specialization and the partial commercialization of agriculture. It became more profitable to work the land in small nuclear family groups now that farm management demanded more care and a growing demand for labour existed in the urban centres. As a consequence, the nuclear family came to be the dominant form in rural Japan. The position of the nuclear family in traditional Japanese society was also strengthened by the gradual emergence of a pattern of primogeniture. Primogeniture is relatively less painful in Japan because of the respect given adopted sons-in-law, who often are the younger brothers who did not inherit the estate into which they were born but whose children inherit their wives' estates.

The other groups covered in this article (except the independent Lolo) were dominated by or integrated into the sociopolitical organizations of other East Asian groups or cultures during the traditional period.

The Hui-hui, for example, are very similar in social and familial form to the Chinese, differing only in those aspects in which religion has been influential. One finds in China Chinese-Islāmic communities existing as compact segregated enclaves, focussing on a single mosque. There is more community spirit than among the Chinese, and the Islāmic minister, *chiao-chang*, is also the leader of the community.

The situation of the independent Lolo of the Ta-liang Shan is complex because of the interrelation between caste, class, and lineage affiliation. Only about one-tenth of the population of the Ta-liang Shan were Mosu or Black Lolos—that is, the upper caste of "pure" Lolos. The sociopolitical organization of the Mosu was the segmentary lineage with confederations, the latter continually rearranging because of feuding over "subjects," household attendants, and slaves. The tripartite grouping into subject or White Lolo, attendant, and slave was the result of contact with Chinese. Slaves were recently captured Chinese who after several generations came to regard themselves as Lolo and became household retainers. Their master then would permit them to become White Lolo with the status of subject. Members of the lower caste were apparently not integrated as kin with the Black Lolo but formed their own patrilineages. The relations between Black Lolo and subjects and retainers were superficially relaxed and egalitarian. They ate the same food, slept in the same house, and observed the same rules of hospitality. One might say that the greatest concern of a Black Lolo was for the welfare of his subject, because, if he acted wisely, kindly, and generously, he could attract followers. This became, of course, a point of discord between himself and other Black Lolos, as the men he might attract already had ties as subjects, attendants, or slaves with other Black Lolo.

Socialization and education. East Asian traditional society was dominated by middle-aged men and women. Within the joint family it was the wife—a former daughter-in-law and later a mother-in-law—whose voice was paramount. In the outside world the middle-aged man directed affairs after years of apprenticeship. The transitions from one stage of life to the next were all gradual, except for retirement. Adulthood was not achieved by passing some initiation rite, taking a head, or getting married—and certainly not by reaching a certain age—but only by acting responsibly. Retirement marked a sharp break: within a short time a powerful and authoritative adult might be converted into a gentle grandparent.

Children were closely integrated into the household. In China, Korea, and Japan especially it was believed that they were never able to repay the debt to their parents, and thus they strove to succeed and by this means win parental approval. This was especially true of brothers who had to cooperate in the joint family but who were also competing individually for parental approval.

TRADITIONAL ECONOMIC LIFE

Settlement patterns and housing. There are basically two house types in East Asia. One type is exemplified in its simplest form by the Ainu hut. Uprights support a ridgepole connected by rafters to the corners. Reed matting is used both for walls and roofing, as well as for flooring. The essential point is that the uprights, not the walls, support the roof. The other type may be exemplified by the Lolo house, which is constructed of wooden planks that, together with uprights, support the roof. The roof is made of planks held down by stones.

The North Chinese house is fundamentally the Ainu hut with elaborations. Large pillars may replace the simple uprights of the Ainu, fired bricks the mat walls, and tiles the mat roof, but the plan of construction remains the same. It is possible for an entire wall of a Chinese house to fall away without affecting the stability of the remaining sections. This is also the traditional house type of the Manchus, Hui-hui, Minchia, Koreans, Taiwanese, Miao, and Yao. Significant elaborations include the construction of a raised sleeping platform heated by flues from the stove. The Koreans prefer a heated floor instead. The house may be enclosed within a courtyard among Manchus, Chinese, Minchia, Hui-hui, and sometimes Miao and Yao. Or a smooth, hardened earth area, a *matang*, may be found in front of the house, as among the Koreans.

The Japanese house may be seen as a highly original synthesis of basic Ainu hut and Lolo house. Uprights are used, not necessarily pillars, so that the walls carry some of the weight. The mat flooring again suggests the Ainu type.

In the rural areas, houses may be grouped in tight clusters or as isolated farmsteads. By and large the Chinese, Manchus, Minchia, Koreans, Ainu, and Hui-hui live in tightly clustered settlements, while the other peoples have both. On Taiwan the Chinese-derived population tends to live more often in isolated farmsteads reminiscent of the settlement patterns of the United States Middle West.

Agriculture and trade. Every one of these societies is to some extent dependent upon agriculture. The Ainu have perhaps been mostly dependent upon fishing, gathering, capturing falcons for sale, trapping, and hunting—especially to obtain trade goods from the Japanese.

Most Chinese, Minchia, Koreans, Taiwanese, Hui-hui, and Japanese practiced a labour-intensive/land-scarce type of agriculture during the traditional period in which the primary beast of burden was man.

For South China, Korea, Japan, Taiwan, and among the Minchia rice is the dominant crop. The traditional method of growing rice requires the use of a water buffalo for preparing the soil, in addition to tremendous amounts of human energy for transplanting, weeding, and harvesting. In order to grow paddy, or wet rice, the farmer must have more control over his water supply than nature generally assures. To satisfy this need, countless small irrigation systems have been constructed throughout East Asia. Only some of the marginally arable land of the Hui-hui and adjacent regions in China's Northwest are totally dependent on irrigation. Here wa-

terwheels, placed along the riverbank and canals, are used.

The Miao, Yao, Manchus, and Lolo made more use of animals and undomesticated plants. The Lolo depended as much on pastoralism as on agriculture. The Manchus definitely depended primarily upon agriculture, but at nearly every point a beast of burden assisted man in working the soil. Although the Manchu today grow the same crops as the Chinese farmers of the North China Plain (maize, kaoliang, wheat, oats, millet, and buckwheat), their fields are worked by oxen or horses. Goods are transported in carts pulled by animals. Animals are fed grain rather than left entirely to scavenge. Only in the kitchen garden does Manchu agriculture resemble the intensive cultivation of the Chinese.

In all of East Asia during the traditional period a commercial and monetary economy was manifest. Farmers produced more than was necessary for subsistence, selling their surplus to buy goods at a market or to pay taxes and rents or repay loans. It is difficult to arrange the traditional economies of East Asia along a scale, but certainly by the beginning of the 19th century Japan was the most commercialized and industrialized. The power of the Chinese gentry and the centralized state, always jealous of their prerogatives, stifled industrial development in China. The same was apparently the case in Korea, whereas Japan's postfeudal society was in this respect more like that of Europe.

Degrees of economic development prior to modernization

TRADITIONAL BELIEF AND AESTHETIC PATTERNS

The syncretic religions found throughout the Far East developed through the interweaving of several belief systems. Ancient contributions were shamanism—that is, belief in an unseen world of gods, demons, and ancestral spirits responsive only to the shaman, or priest—and a belief in an impersonal supernatural force. In anthropological literature this force is often called *mana*, but in the Orient the term used is *tao*, the "way." This force has two aspects, an active and a passive, in Chinese known as *yang* and *yin*, respectively (see CHINESE PHILOSOPHY; TAOISM). Perhaps such beliefs derive from the areas to the south, but such a conceptualization of nature could also have derived from the philosophical underpinning of shamanism. A third ancient element is ancestor worship. Important later ingredients have been Buddhism and, in Korea, Christianity, which has been closely associated with Korean nationalism.

Features of East Asian shamanism

Twentieth-century Tungus (including Manchu) shamanism is probably as close to the ancient shamanistic foundation of East Asian beliefs as can still be found. There is a tendency for a shaman to represent his own lineage, although a shaman does not have to limit his services to his kin group. A feature of shamanistic practice is the séance, in which the shaman sends one of his souls out to meet with spirit helpers whose job it is to induce the lost soul, or fraction of a soul, of a patient to return. The séance demands both control and ecstasy. It is a dangerous occupation, psychologically demanding. In one variation, the shaman, instead of sending out one of his souls, becomes "possessed" by the spirit he has contacted. Much of the popular religious Taoism of China uses shamanistic séances of the latter type. Sometimes, a priest—that is, a person primarily concerned with ritual and doctrine rather than with personalized contact with spirits—is prominent, as in Japanese Shintō.

In religious terms *yang* is decidedly better than *yin*. Chinese saints such as Kuang Ti (supposedly a general who lived during the period of the Three Kingdoms, about the 3rd century after Christ) had an abundance of *yang*. Graveyard trees, the ghosts of those who died in childbirth, or those without male heirs have too much *yin*. The latter, called *kuei* in Chinese, supposedly caused much of the misfortune in traditional China. It therefore behooves people to propitiate the *kuei* at the same time that they take care of their own ancestral spirits. If the *kuei* cannot be placated by offerings made at crossroads or by a special festival celebrated at midsummer, then they may hopefully be frightened away by "hot noise," a literal translation of the Chinese word for happiness, with firecrackers, red-coloured paper, and by chanting Classical literature.

The three-part soul of one's own ancestor requires care after death lest it become a *kuei*. Generally, one's own ancestors are remembered by burning incense at the family altar. Ancestor worship is very ancient among the Chinese and probably among some of the other peoples in Asia, but it apparently became important in Japan only after the importation of Chinese culture. Before that time the Japanese worshipped only the deceased founder of the lineage. The reverence bestowed upon the emperor may be seen as deriving from this custom.

The social and intellectual life of the peoples of East Asia has affected their interpretation of the nature of religion. The Chinese learned gentleman has tended to be an agnostic, whereas his Japanese counterpart is generally a Buddhist of the Zen sect. In China and Korea there has been both a Confucian activism and a Taoist passivism. In Japan, on the other hand, Zen encouraged its *bushi* adherents to find their way within the world rather than by withdrawing to some lonely and beautiful spot, as in the Taoist ideal.

Aesthetically the Chinese gentry and their Japanese counterparts expressed themselves along similar lines. Fundamental to their visual art is the skill perfected in practicing calligraphy with brush and ink stone. The novel, a literary form dependent on a "middle class" having leisure time and an interest in details of personality rather than simply in characters of epic proportion, developed much earlier in both China and Japan than it did in the West. Stylized operatic theatrical productions, the *Ching-hsi* of China and the Nō of Japan, were enjoyed by the literati but were in form and content simply the most polished of a series of musical theatrical productions found throughout the area in the form of shadow plays, puppet shows, and tales of itinerant storytellers. It is in pornography that a difference may be found, the Japanese celebrating an exuberant phallicism, the Chinese being more sensual and playful.

MODERN DEVELOPMENTS

For the peoples of East Asia the 20th century has been a time of quickening change. The Taiwanese have emerged as a distinct ethnic group and nation, whereas the Manchus, once the rulers of China, have become eclipsed. By the time of the fall of the Ch'ing dynasty in 1911, most Manchus had become so Sinified that they were simply absorbed into the Chinese population, though some pockets of Manchus (the Bannermen) who had developed a strong sense of national identity and who resisted absorption by the Chinese remained. The establishment and sponsorship in Manchuria (Northeastern Provinces) of Manchukuo by the Japanese in 1932 gave such groups a respite. Now three generations removed from the fall of the Ch'ing dynasty and one from the unification of China under the Communists, there is probably little left of the Manchus as a distinct people.

The Ainu of Hokkaido have in a sense been fighting a losing battle against the better organized Japanese since the 1st millennium of the Christian Era. The Ainu, until Russian penetration of the Far East, were valued by the Japanese as a market. The Russian advance led to closer Japanese control over the Ainu and, after the Meiji Restoration (1868), to much closer supervision of Ainu life. The Ainu were encouraged to become sedentary farmers rather than moving on from plot to plot. This often led to their neglect of fishing and other traditional occupations. Large numbers of Japanese have since settled on Hokkaido, which has led to intermarriage, and for over 100 years now the Ainu have been encouraged to become successful within the framework of modern Japan. Despite discrimination by some Japanese, especially up to the time of World War II, the Ainu have accepted assimilation into Japanese society. To what extent the old ways will continue to be observed by some families cannot be ascertained, but as a distinct people the Ainu may soon disappear except in the ethnographic record.

The fate of the Minchia was sealed much earlier, in the Yüan dynasty (1279–1368), when Kublai Khan con-

quered the Nan Chao state. The Minchia were encouraged to become Chinese literati, and they have since then competed for Chinese honours with even greater fervour than the Chinese themselves. They achieved a high measure of success in Chinese society so that over the centuries they have become very Sinified. At the present time no political, social, or cultural situation of a nature that would slow the Minchia's rapid assimilation into the Chinese population seems likely to come into being.

Ethnic awareness of the Muslim Hui-hui in China

One might assume that the Hui-hui are also on the way to losing their distinct identity, but in fact the victory of the atheistic Communist Chinese has heightened the ethnic awareness of the Islāmic Hui-hui, who have maintained their religious bonds. During the last quarter of the 19th century they were the victims of extreme religious persecution, in the course of which millions of them were killed. Now only tolerance and a real end of discrimination by the Chinese will lessen the Hui-hui's resistance to being assimilated. They have apparently been allowed to play a part as a nominally autonomous group by the Chinese Communists.

In one sense the Ta-liang Shan (independent) Lolo have created one of the more successful counter-cultures to the Chinese. They have maintained their independence and identity during 2,000 years of Chinese pressure. During this time they were able to elaborate a social and ceremonial organization of their own and to develop a non-Chinese and non-alphabetic writing system. But also during the years, other Lolo groups, closely related to the independent Lolo in speech and culture yet living outside the mountain fastness of the Ta-liang Shan and under Chinese jurisdiction, have increased in number until there are about 3,000,000 of them, not including the Nakhi, Lisu, Lahu, and Akha. But there are only some 500,000 independent Lolo left: Chinese control is obviously gaining.

The so-called acephalous political organization (*i.e.*, lacking a governing head or chief) of the Lolo, Miao, Yao, and other groups in southern and southwestern China has given them certain advantages. The primary advantage lay in the inability of the Chinese to be done with them by one decisive battle or campaign. But it has proved a disadvantage as well, as it has left them unable to deal with the gradual encroachment of the Chinese onto their lands, first as individuals, later as officials. Only by enslaving the Chinese and destroying their settlements were the Lolo along the Szechwan–Yunnan border able to hold their own. Once deprived of the strength to carry out such drastic measures and forced to deal with the powerful government of the People's Republic of China, the Lolo, like the other non-Chinese peoples of southern and southwestern China, will probably find within a generation or two that all that remains of their ethnic uniqueness are some folk-art motifs and a few officially sponsored folk festivals.

Paradoxically, the 30,000,000 non-Chinese of southern and southwestern China have been able to maintain their own culture and language for thousands of years; yet, now that a government of China exists that pays special attention to them and that is sometimes sympathetic to them, they are likely to become totally acculturated. The reason for this is that they live scattered in small enclaves among the Chinese population.

Relations between the Taiwanese and the "mainlanders"

As pointed out already, the Taiwanese control most aspects of their lives and, through their domination of the provincial government, the careers of many "mainlanders" on the island. The latter are attempting to re-Sinify the Taiwanese by teaching Mandarin Chinese in the schools, by the cultural-restorations movement exemplified by the building of a beautiful new national museum, and by other government services and programs. But the fact that the Taiwanese own the land and control the resources, plus their greater numbers, suggests a different direction of development. If Taiwan is allowed to settle its own affairs, it is more likely that the Taiwanese will absorb the "mainlanders" than vice versa.

China and the West. Traditional East Asian culture, based on the tribute system and the respect for Chinese intellectual values, ended when it proved impossible to ignore the West and when the West would not agree to be just a somewhat different type of barbarian to be incorporated into the established order. Had the demands of the West been only those of imperialistic barbarians, they could have been dealt with in the classical manner. As it was, the dominant powers of the Far East had to see themselves as students instead of teachers. To accept this meant the destruction of the mythology underpinning the tribute system. The Westerners refused to kowtow and insisted on recognition at the national level. They defeated the Chinese militarily but did not thereupon attempt to conquer China and set up a new dynasty. Within a short time China lost suzerainty over Korea, over the Tungus on the northern watershed of the Amur, over the Khalkha Mongols, the Tibetans, the Vietnamese, the Ryukyuians, the people of Tanna Tuva, and the majority of Kazakhs and Kirgiz.

The response of the Manchu court to this situation was an attempt at antagonistic acculturation; that is, an attempt at learning the technology of the West without accepting Western values or allowing any changes in Chinese social, political, or economic organization. This response allowed the court to survive during the initial stages of the revolutionary period, which began in China partly because of the weakness of the court and its foreign origin but also because of a growing awareness that the gentry system was inadequate to supply services felt to be essential to a modern state. The use of modern arms at first enabled the court to put down rebellion, but by the 20th century the buildup of foreign mercantile and industrial centres in the major coastal cities and the activity of Western missionaries in the interior had awakened a growing desire for modernization and a reassertion of China's identity. It was only natural that the original leaders in the anti-Ch'ing movement should have come from those areas in closest contact with the West. The government tried to meet their demands by changing from the Classical system of education to one stressing Western thought, by allowing the development of industry, and even by accepting the idea of a constitutional monarchy—in fact, abandoning the earlier attempt at antagonistic acculturation. But the anti-Manchu sentiment that had been released was too strong to halt the fall of the Ch'ing dynasty in 1911.

China from the fall of the Manchus until 1950 found itself bound by a number of severe contradictions. Those who had given the revolution intellectual force stressed individualism, Western intellectualism, science, universal education, democracy, national reconstruction, and an improvement in the peoples' livelihood, as well as other liberal and popular measures. These measures were all suited to the needs of the times. During previous periods of Chinese history, such as the Ch'in (221–206 BC) and the Sui (AD 581–618), the Chinese had been quite able to rework their institutions drastically. The new reformers, however, had no real power base in the countryside. Their strength lay in the Western-influenced treaty ports and was suspect to the majority of Chinese. Those who soon came to have real power, the warlords who from 1912 to 1949 controlled most of China independently of the central government, did not understand the nature of the changes that were needed. Most of the warlords had abandoned Classical ideals, it is true, but they had not replaced them with those of the West. In fact, they were primarily opportunists or chauvinistic nationalists. The head of the Kuomintang, Chiang Kai-shek, reflected the desires of the nation to reconcile the different factions, and he might have been successful in doing so had the Japanese not sought to prevent the emergence of a modern China.

In some areas modernization did occur. There was limited industrialization; a socially conscious literary movement exemplified especially by the *pai-hua* (vernacular-language) movement led by Hu Shih and the bitter essayist Lu Hsün; the building of railroads and motor roads; the organization of one of the more efficient postal services in the world; and a nationwide, although not compulsory, school system. But the basic problem was that the leaders lacked the power to deal with the patron–

Breakdown of traditional relationships

client relationship that had held the traditional society together. As the warlords of China increased services, as industrialization continued, as classical principles well adapted to the maintenance of good *kan-ch'ing* were abandoned in favour of individualism, and, most importantly, as the old-fashioned gentry lost their ability to intercede for the welfare of the community, the traditional institutions became means of exploiting the farmers. This was the case because, as the power of the government increased, the restraint on the landlord that had made it essential that he keep the respect and goodwill of his neighbours and tenants was removed. The government gave the landlord protection from bandits of whatever derivation and assured the collection of rents. It became no longer necessary for the landlord to maintain a good standing in the community in which he owned land. In the area near major cities in which this process had evolved furthest, landlords, now often urban industrialists, divorced themselves entirely from the problems of the land and appointed an estate manager who collected a set amount of rent and kept whatever else he could squeeze out of the farmers.

The client–patron system had functioned amazingly well in traditional China. Few such holistic systems, encompassing political, educational, artistic, social, ethical, and economic goals, have lasted so long and with such vigour. Its major weakness—its lack of consistent service (in regard to the provision of regular admittance to schools, for example) and justice—was also a major virtue, as it accompanied a government that lay lightly upon the citizenry. But it became impossible to retain only a part of the system—namely, privileges for the gentry and nepotism in government. These were not destroyed until the Chinese Communist Party had fled to the countryside to lead a revolution in the classical Chinese manner.

The Communists' main thrust was a reorganization of society from the village level upward, with the cooperation of all "progressive" elements in an attempt to do away with "feudalism." This involved the elimination of vestiges of the client–patron system, along with many other associated traditional values, such as the hierarchical position of men in Chinese family life. Emphasis was placed upon the services individuals could offer the community and state regardless of the status of the person and his connections (*kuan-hsi*) or how those of higher rank felt about him (*kan-ch'ing*). In summary, the personalized, traditional society was considered exploitive and inefficient. Thus, the old lines of structural personalization were discarded and new ones established, such as farm cooperatives, women's clubs, literary movements, and anti-Japanese militia and crop improvement committees.

Non-Communist East Asia. Korea became gradually weaker under the Yi dynasty (1392–1910). The Yangban (gentry) took over once public lands and then did not pay taxes. To make up for the deficit the freeholders were taxed more heavily. They could not afford the taxes and lost the land, which thereupon passed into the hands of the bureaucratic landlord class. When famine occurred the central government had no reserves to alleviate the situation. Meanwhile, the intelligentsia, unaware of the real peril, indulged themselves in factional disputes more serious perhaps than those that had contributed to the fall of the Ming dynasty in China (1368–1644). The West largely bypassed Korea, which remained oriented along traditional lines after China had long been opened up to Western trade, missionizing, and interference and after the Meiji Restoration had set Japan on a path of modernization and industrialization. The Yi dynasty would probably have been swept aside in due time by a Christian- or Tonghak- (a syncretic revitalistic religious response to the West including elements of both Eastern religion and Christianity) led rebellion if Japan had not invaded Korea first and conquered it.

The Japanese conquest increased the Korean sense of national unity and in many ways modernized the country. South Korea is one of the more industrialized states in Asia, and a land-reform program is in effect.

Japanese society at the beginning of the 19th century had the strengths of feudal values interwoven into a bureaucratic system of government. Its people were ready, however, to discard the remnants of the feudal class system. It seemed better to be a prosperous businessman than a penniless samurai.

Recent social changes in Japan

In recent times some novel social forms have emerged in Japan, in the development of which the Japanese have been able to draw upon a number of value systems. One of these has been Bushido, the hierarchical and authoritarian conception of the role and status of the *bushi* service bureaucrat, based on a mixture of classical Chinese, feudal Japanese, and bureaucratic values. With its strong emphasis on loyalty in a vertically ordered system of two-person reciprocity, this way of the samurai is one of the important values that have shaped modern Japan.

Another hierarchically structured model dominant in some parts of Japan has been the *dōzoku*, the combination of kith and kin with landlord and tenant. In other areas, relations between landlord and tenant, although influenced by the *dōzoku* model, tended toward abuse by the landlord after a strong central government had assured his rights. This led to the growth of unrest and militancy among the rural population, finally suppressed in the 1930s.

The paternalistic model of organization in Japanese industry, known as *oyabun-kobun*, is probably inspired by both the Bushido system of values and the *dōzoku* (segmentary lineage). The result has been security for the worker, full employment, and ordered advancement with a minimum of division along class lines and in the professions.

The third organizational model has been drawn from the local kindred kinship group and the cooperating local group, the *buraku* (hamlet). These local units have become especially important since a land reform that occurred after World War II brought an end to landlordism in rural Japan. These two sources have provided the inspiration for egalitarianism and a desire for consensus found among the Japanese.

The populace is also divided into associational frames of reference such as industrial firms and villages. Differences in rank and status are important, but not class, except for former outcaste groups such as the Eta. Economically, the system probably works as well if not better than that of the West, as it does not tend to cause social alienation.

The interaction of these various concepts in the political arena has been a mixed blessing. Perhaps some unique adjustment—allowing for efficiency in government, the desire for consensus, and homage to rank while furthering egalitarianism—will be achieved in the future.

In the 20th century East Asia as a cultural entity, as defined and described at the beginning of this article, was badly shaken by the collapse of traditional China. At times it has seemed that Japan was destined to take China's place, but Japan's defeat in 1945 changed this. Up to the early 1970s China has directed its attention to its Inner Asian frontiers and to internal social, cultural, and economic change, while Japan, South Korea, and Taiwan have concentrated more on industrialization and on developing their international trade. Given a continuation of present circumstances, one might predict the end of East Asia as a psychological reality within another generation.

BIBLIOGRAPHY. GEORGE B. CRESSEY, *Asia's Lands and Peoples*, 3rd ed. (1963); and JOSEPH E. SPENCER, *Asia East by South*, 2nd ed. (1971), are comprehensive cultural geographies. The best balanced and most thorough history of East Asia is EDWIN O. REISCHAUER and JOHN K. FAIRBANK, *A History of East Asian Civilization*, vol. 1, *East Asia: The Great Tradition* (1961) and vol. 2, *East Asia: The Modern Transformation* (1965). The archaeological report *Kwang-chih Chang, The Archaeology of Ancient China* (1963), is comprehensive but unduly polemical concerning the possibility of Chinese borrowing from the West during the Bronze Age. HERRLEE G. CREEL, *The Birth of China* (1954), depicts the early civilization in China. FREDERICK W. MOTE, *Intellectual Foundations of China* (1971), is a concise, audacious, scholarly book on the classical period of East

Asian philosophy. The short pocketbook by PETER DUUS, *Feudalism in Japan* (1969), is a good introduction to the distinctive features of Japanese sociopolitical development. There are no works equivalent to this for other areas in East Asia, but KARL A. WITTFOGEL, *Oriental Despotism* (1957), and WOLFRAM EBERHARD, *Conquerors and Rulers*, 2nd rev. ed. (1965), are of interest. JOHN K. FAIRBANK (ed.), *The Chinese World Order: Traditional China's Foreign Relations* (1968), contains a discussion of China's tribute system. Opinions concerning Japanese prehistory are found in ROBERT J. SMITH and RICHARD K. BEARDSLEY (eds.), *Japanese Culture: Its Development and Characteristics* (1962). FRANK M. LEBAR, GERALD C. HICKEY, and JOHN K. MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia* (1964), is a comprehensive work concerning not only the peoples of southern and southwestern China but also those of Southeast Asia. The statements contained within it are based upon an evaluation of various primary sources. The contradictory opinions are passed on to the reader along with an annotated working bibliography. The monograph does not cover the non-Chinese peoples of most of Szechwan Province. For this area the reader should consult HENRI M. D'OLLONE, *Mission d'Ollone 1906–1909*, 2nd ed. (1911; Eng. trans., *In Forbidden China: The d'Ollone Mission, 1906–1909*, 1912); DAVID C. GRAHAM, "The Customs of the Ch'uan Miao," and "The Ceremonies of the Ch'uan Miao," *Journal of the West China Border Research Society*, Series A, 9:13–119 (1937), and "The Customs of the Ch'iang," *ibid.*, 14:65–99 (1942). For Taiwan GEORGE W. BARCLAY, *Colonial Development and Population in Taiwan* (1954, reprinted 1972), is good; also FRANK B. BESSAC, "Cult-unit and Ethnic Unit: Processes and Symbolism," in the American Ethnological Society *Proceedings of the 1967 Annual Spring Meeting*, pp. 58–71 (1968). A sociological and ethnographic approach to the question of gentry society in China is treated in HSIAO-T'UNG FEI and CHIH-I CHANG, *Earthbound China: A Study of Rural Economy in Yunnan*, rev. ed. (1945); and HSIAO-T'UNG FEI, "Peasantry and Gentry: An Interpretation of Chinese Social Structure and Its Changes," *Am. J. Sociology*, 52:1–17 (1946); FRANCIS L.K. HSU, *Under the Ancestors' Shadow: Kinship, Personality, and Social Mobility in China* (1948, reprinted 1971); YUNG-TEH CHOW, *Social Mobility in China* (1966); MORTON H. FRIED, *Fabric of Chinese Society* (1953, reprinted 1969); and FRANK B. BESSAC, "Change in the Meaning of Kanch'ing in Taiwan Due to the Land Reform," *Proceedings of the VIIIth International Congress of Anthropological and Ethnological Science*, vol. 2, pp. 245–247 (1968). The influence of the segmentary lineage upon Chinese life is analyzed by MAURICE FREEDMAN in *Lineage Organization in Southeastern China* (1958, corr. reprint 1965). RICHARD K. BEARDSLEY, JOHN W. HALL, and ROBERT E. WARD, *Village Japan* (1959), is a well-balanced discussion of rural Japan. HARUMI BEFU, *Japan: An Anthropological Introduction* (1971), attempts to sketch the totality of Japanese life including introductory chapters on prehistory and the traditional period. Ethnographies or reports that reflect the revolution in China are C.S. CHEN, *Rural People's Communes in Lien-Chiang* (1969); ISABEL and DAVID CROOK, *Revolution in a Chinese Village: Ten Mile Inn* (1959); JAN MYRDAL, *Rapport från kinesisk* (1963; Eng. trans., *Report from a Chinese Village*, 1965); CH'ING-K'UN YANG, *A Chinese Village in Early Communist Transition* (1959). Chen's work is an analysis of captured documents kept by local cadre men. Myrdal's report is largely the presentation of notes from interviews with cadres and those committed to the revolution. The Crooks attempt to reconstruct the revolution on the basis of interviews. Yang writes as a trained observer. Concerning religion and the arts, S.M. SHIROKOGOROV, *Psychomental Complex of the Tungus* (1935), is a classic discussion of Siberian and East Asian shamanism; and *Mai-mai Sze, The Tao of Painting: A Study of the Ritual Disposition of Chinese Painting, with a Translation of the Chieh Tzu Yüan Hua Chuan; or, Mustard Seed Garden Manual of Painting, 1679–1701*, 2nd ed., 2 vol. (1963), is an excellent basic text on traditional East Asian art. RYUSAKU TSUNODA, WILLIAM THEODORE DE BARY, and DONALD KEENE (eds.), *Sources of the Japanese Tradition* (1958; 2-vol. text ed., 1964), is a comprehensive introduction to Japanese humanism. EVELYN MCCUNE, *The Arts of Korea* (1962), provides a collection of prints in part portraying not simply Korean but also East Asian styles in general. MICHAEL SULLIVAN, *Chinese Art in the Twentieth Century* (1959), brings the subject up to date. CH'ING-K'UN YANG, *Religion in Chinese Society: A Study of Contemporary Social Functions of Religion and Some of Their Historical Factors* (1961), combines good factual knowledge with sociological interpretations. Finally, JAMES B. CROWLEY (ed.), *Modern East Asia: Essays in Interpretation* (1970), is an eclectic but helpful and thoughtful collection.

(F.B.B.)

Easter Island

Easter Island (Isla de Pascua), a Chilean dependency in the eastern Pacific Ocean, famous for its giant stone statues, represents the easternmost outpost of the Polynesian island world. It stands in isolation 1,200 miles (1,900 kilometres) east of lonely Pitcairn Island and 2,300 miles (3,700 kilometres) west of Chile. Forming a triangle 14 miles long by 7 miles wide, the island has an area of 45 square miles (117 square kilometres); its highest point is 1,969 feet (600 metres) above sea level. To its original inhabitants the island is known as Rapanui (Great Rapa) or Te Pito te Henua (Navel of the World); its Dutch discoverers named it Paaseiland (Easter Island) in memory of their own day of arrival. Its mixed population, predominantly of Polynesian blood, numbers slightly over 1,000; almost all live in the village of Hanga-Roa on the sheltered west coast. The island is administered as part of the province of Valparaíso, Chile (for coverage of the region as a whole, see PACIFIC ISLANDS; and PACIFIC OCEAN; for its history, see OCEANIA, HISTORY OF).

The landscape. The small and hilly island is not part of a sunken landmass but is a typical oceanic high island formed by a series of separate underwater volcanic eruptions. Geological and oceanographic evidence shows that no perceptible emergence or submergence of the island's coastline has taken place since the last fall in sea level, which occurred less than 10,000 years ago. Three extinct volcanoes chiefly composed of tuff (a rock formed of compacted volcanic fragments) and joined by their own lava flows give the island its characteristic triangular shape. Parasitic tuff craters and cones (*i.e.*, craters and cones formed on the side of, or near, volcanoes after the original vent has become plugged up) are interspersed in the landscape, which is otherwise dominated by eroded lava fields. Most of these fields are rugged and are often thickly packed with both large and small lumps of cellular and tuffaceous lava that is either black or rusty in colour. Soil is sparse and suitable for extensive cultivation mainly in the Hanga-Roa and Mataveri area in the southwest, at Vaihu and on the plain southwest of the volcano Rano Raraku, and on the prehistorically cleared Poike peninsula in the eastern corner of the island. The arid landscape has no rivers or streams, but rain collects in the partly bog-covered crater lakes of the volcanoes Rano Kao, Rano Raraku, and Rano Aroi. Water from the extremely deep crater of Rano Kao, which is about 1,000 yards (one kilometre) wide, is piped to Hanga-Roa. The coast is formed by soft, eroded, ashy cliffs, with a vertical drop of about 500 to 1,000 feet; the cliffs are intercepted by long stretches of low, hard, and rugged lava formations. There is no natural harbour, but anchorage is found off Hanga-Roa on the west coast; off Vinapu and Hotu-Iti on the south coast; and off Anakena and in the Bahía la Perouse on the north coast. Notable among the few small offshore islets are Motu-Nui, Motu-Iti, and Motu-Kaokao (which figured in a local bird cult) near the southwest cape. The only real sandy beach is at Anakena; most other beaches are of gravel. Caves abound, many consisting of subterranean rooms joined by narrow tunnels extending far into the lava beds.

Anchorages

Climate. The climate is subtropical, with an average annual temperature of about 72° F (22° C), varying between about 77° F (25° C) in January and 70° F (21° C) in August. Average annual precipitation is about 40 inches, but with considerable annual variation. September is the driest month, and the heaviest rainfall occurs in June and July. Winds in June and August are irregular; during the rest of the year trade winds from the east and southeast are dominant. From September through March the Peru (or Humboldt) Current, which has an average temperature of about 70° F (21° C), flows against the island.

A subtropical climate

Vegetation. Indigenous plants and animals are few. The toromiro tree is the only wild tree and the Carolina wolfberry the only shrub; the vegetation is predominantly herbaceous. Analysis of pollen deposits has revealed that other trees and shrubs, among them a palm, were formerly present on the island until exterminated by extensive fires occurring at the time of aboriginal human

settlement. Today only 31 wild flowering plants, 14 ferns, and 14 mosses are reported. Grass and small ferns dominate the barren landscape, whereas the boggy crater lakes are thickly covered by two imported American species, the totora reed (an important building material) and *Polygonum acuminatum* (a medicinal plant). A number of cultivated species of plants was also introduced partly from America and partly from Polynesia before the arrival of Europeans; of these the principal species was the sweet potato, which was cultivated in extensive plantations and formed the staple diet. Bottle gourds, sugarcane, bananas, taro, yams, chili peppers, and two useful trees (*i.e.*, the Asiatic paper mulberry, with bark used for cloth manufacture, and the American *Triumfetta semitriloba*, with bark used for rope making) were of aboriginal importation, as also probably were the husk-tomato, a small variety of pineapple, and the coconut. The Chilean government has begun a program of reforestation.

Animal life. Before the arrival of human beings, the only vertebrates were either fish or seabirds capable of long flights. The animal life on land was otherwise restricted to a very few species of isopods (an order of crustaceans, which are aquatic animals with shells), spiders, insects, worms, a snail, and a centipede. Vast quantities of flies, large cockroaches, and a small scorpion were introduced recently. A small, long-legged chicken reported to lay blue eggs was introduced in pre-European times but later interbred with European varieties. The aboriginal edible Polynesian rat was subsequently replaced by larger European species. Sheep, horses, cattle, and pigs were introduced into the islanders' economy by the missionaries who established themselves ashore in 1866. Sheep have been especially numerous since a French sheep rancher began commercial ranching in 1870. A large wild cat, living in caves, is of unknown introduction. A Chilean partridge, a quail, and a small hawk have been added to the wildlife since 1880. Sea turtles and seals that formerly visited the island are now rare curiosities, but crayfish and various coastal and deep-sea fishes abound around the coast.

The inhabitants. The island population represents the easternmost settlement of a basically Polynesian subgroup, although non-Polynesian physical and linguistic traits have been noted among them since the time of European discovery. The original Rapanui vocabulary has been lost except for some mixed Polynesian and non-Polynesian words recorded before the Tahitian dialect was introduced to the decimated population by missionaries in 1864. Today Spanish is generally spoken. The islanders consistently divide themselves into descendants of two distinct ethnic groups, the "Long-Ears" from South America, and the "Short-Ears" from Polynesia (see below); intermarriage is nevertheless common, and an influx of foreign blood has become increasingly dominant in recent years. Whereas the aboriginal economy was based on the cultivation of sweet potato, chicken raising, and coastal fishing, the islanders are at present in the process of shifting to a cash economy following their employment at hourly wages after an airfield was opened at Mataveri, where a large settlement of Chilean labourers, technicians, and civil servants has been established.

History. The first European to discover Easter Island was the Dutch admiral Jacob Roggeveen, who paid it a single day's visit in 1722. He and his crew found a racially mixed population worshipping the sun in front of huge standing statues. An expedition dispatched by the Spanish viceroy of Peru rediscovered the island in 1770. The Spanish spent four days ashore and reported a population of 3,000 persons of extremely mixed physical appearance. A civil war seems to have raged on the island before the arrival of the English navigator Captain Cook in 1774; a decimated, poverty-stricken Polynesian population of only about 600 or 700 men and less than 30 women was found by the Englishmen, who also observed that the large statues were no longer venerated, most of them having been deliberately overthrown. In 1786 the French navigator Jean-François de Galaup, comte de La Pérouse arrived and tried in vain to introduce domestic animals. At the beginning of the 19th century a number of slave raiders, attempting to use slave labour in the establishment of seal-hunting colonies elsewhere in the Pacific, visited Easter Island, causing subsequent visitors to receive a hostile reception. A major slave raid launched from Peru in 1862, followed by smallpox epidemics, reduced the population to 111 before, at the end of the last century, it began to increase once more. In 1864 Brother Eugène Eyraud, a French Catholic missionary, became the first foreigner to settle on the island; as a result, the population became converted to Christianity by 1868. Settlers from Tahiti began to raise sheep in 1870. In 1888 the island was annexed by Chile, which leased nearly all its territory for sheep raising; in 1954 the Chilean navy administration took over the sheep range. Recently a civilian governor was appointed by the Chilean government. The island is now experiencing rapid change. There is a commercial airport, and a tourist hotel has been opened near the Hanga-Piko anchorage on the west coast. There is also a school, a hospital, and a leper station.

The stone statues

Archaeology. The lonely island is famous for its gigantic stone statues, of which there are more than 600, and for the ruins of giant stone walls, which show masterly construction. Archaeological surveys were carried out in 1914 and in 1934; archaeological excavations were undertaken in 1955. The excavations revealed that three distinct cultural periods are identifiable on the island. The early period is characterized by a solar observatory, megalithic walls of exquisitely fitted stone masonry astronomically oriented to the annual movements of the sun, and a heterogeneous variety of small- to medium-sized stone statues in human form. The middle period is characterized by a single conventionalized type of long-eared and legless busts erected on top of stepped stone platforms known as *ahu*. Existing early-period temple terraces were destroyed and rebuilt with no regard to solar orientation but for the sole purpose of supporting the tremendous weight of the giant stone statues of the new type. The sizes of the statues made were increased until they reached stupendous dimensions; the slim and lofty busts also had huge cylindrical *pukao* (topknots) of red tuff placed on top of their slender heads. Most middle period statues range from about 10 to 20 feet in height, but the biggest among those formerly standing on top of an *ahu* was about 32 feet tall, consisted of a single block weighing about 82 tons, and had a *pukao* of about 11 tons balanced on its apex. The largest statue still standing partly buried in the deep silt below the quarries is about 37 feet tall, and the largest unfinished one with its back attached to the rock is about 68 feet tall. Early-period sculptures were variously made of basalt (a fine-grained, dark-coloured rock solidified from a molten state), scoria (volcanic slag), and tuff, but the middle-period sculptures were all quarried from the special yellow-gray tuff found in the crater walls of Rano Raraku. Inside and outside the crater bowl numerous unfinished statues and thousands of crude stone picks are scattered about, bearing witness to a sudden interruption of the sculptors' work. The unfinished images show that each statue had its front and sides completed to a polish before the back was detached from the bedrock. The image was then slid away to be raised at random in the rubble below the quarries to have the back finished before being moved to some distant *ahu*. Deep oval eye cavities and red topknots were added only after the monument had been finally erected. From one to a dozen statues would stand in row on a single *ahu*, always facing inland. Experiments made by the islanders in 1955 to 1956 showed that 180 persons could pull a medium-sized statue over the ground. Twelve islanders were able to lift a 25-ton statue about ten feet off the ground and to tilt it on end on top of an *ahu;* this work took 18 days with no tools other than two wooden logs that were used as levers. Stones of all sizes were wedged under the statue one by one to form a slowly rising cairn in order to lift the giant monoliths upright. Traditions, supported by archaeology, suggest that the images represented important personalities who were deified after death. The three early period types of statue lack counterparts in Oceania, but have close affinities to

Unfinished carvings

the three specialized forms of monuments erected by an unknown people in the highlands of South America during the Tiahuanaco Period, which occurred in the 1st millennium AD. The middle-period busts, however, clearly evolved from a local prototype and have no counterpart elsewhere. Also peculiar to the middle period was a non-Polynesian bird cult with attendant birdman rites that survived into the third, or late, period. Its ceremonial centre was the village of Orongo on top of Rano Kao, which consisted of stone houses of a non-Polynesian type with roof vaults built as false arches. These houses and contiguous circular masonry dwellings with roof entrances are characteristic of the two earlier archaeological periods on the island; while unknown elsewhere in Polynesia they are common in the adjacent area of South America. The late-period Easter Islanders dwelt in boat-shaped pole-and-thatch houses, or in caves. This period was marked by internal wars, general destruction, and cultural decadence. The *mataa*, or obsidian spearpoint, which was mass-produced, is the characteristic artifact of this period. Wood carving and small crude stone figurines replaced monumental art. Written wooden tablets covered with incised signs (called *rongo-rongo*) placed in boustrophedon (method of writing in which the lines run alternately from right to left and from left to right) were copied from earlier specimens merely for ritual purposes; their proper reading was forgotten, and—despite many claims—modern attempts at deciphering them have failed. During this period art treasures were hidden in secret family caves, while the upright *ahu* images were successively overthrown. Silt from the abandoned quarries descended to the chests of the blind and unfinished busts standing at the foot of the volcano, rendering their overthrow impossible and thus securing for posterity the eyeless heads that have given the island its fame.

The wars of the late period

Tradition insists that destruction began after a period of peaceful coexistence between two people of different culture and language—the Long-Ears and the Short-Ears. The latter, tired of toiling for the former, all but exterminated them in a pyre along an ancient defense ditch at Poike on the far northeastern coast. Carbon dating and genealogies concur in placing this event and the beginning of the late period at about AD 1680. The middle period is estimated to have begun in about AD 1100. The original construction of the artificial Poike ditch, according to carbon dating, took place in about AD 380, but original settlement may well have occurred earlier.

Prospects for the future. The non-Polynesian vestiges on Easter Island have given rise to much speculation. A unanimous resolution adopted by 3,000 scientists at the Tenth Pacific Science Congress in 1961 affirmed that South America, as well as Southeast Asia, is a major source area for the peoples and cultures of the Pacific islands; Easter Island is the most apparent example of this cultural amalgamation. Today island life is increasingly assimilating Chilean influences. Easter Island is also gaining a new importance as a stepping-stone in modern aviation routes east and west across the South Pacific. The introduction of tourism, as well as the replanting of the forests, is likely to signal the opening of an entirely new era in local life.

BIBLIOGRAPHY. M.J. ROGGEVEEN, *Extract from . . .* [his] *Official Log of his Discovery of and Visit to Easter Island in 1722*, Hakluyt Society, 2nd ser., vol. 13 (1908), and F. GONZALEZ, *The Voyage of Captain Don Felipe Gonzalez in the Ship of the San Lorenzo . . . to Easter Island in 1770–71, ibid.*, two annotated syntheses of logs and documents; W.J. THOMSON, *Te Pito te Henua, or Easter Island* (1889), an early survey of the archaeology and culture; K. ROUTLEDGE, *The Mystery of Easter Island* (1919), a popular account of the first major expedition; A. METRAUX, *Ethnology of Easter Island* (1940), a comprehensive ethnological survey; P.S. ENGLERT, *La Tierra de Hotu Matua* (1948), observations by a resident missionary; THOR HEYERDAHL, *Aku-Aku: The Secret of Easter Island* (1958), an account of the adventures of an archaeological expedition, "The Statues of Easter Island" and "The Composite Culture of Easter Island," in *Sea Routes to Polynesia* (1968), a popular report on present knowledge of pre-European activities, and edited with E.N. FERDON, JR., *Report of the Norwegian Archaeological Expedition to Easter Island and the East Pacific*, 2 vol. (1961–65), a general introduction to the island, followed by illustrated reports by a team of American and Norwegian archaeologists.

(Th.H.)

Eastern Arctic Cultures

In the eastern, or Eurasian, Arctic there is a diversity of ethnic groups distributed across the great expanses of land from northern Scandinavia to the Asian shores of the Bering Sea. All occupy either one or both of the principal ecological zones of the north, the far northern tundra (cold, desert lowland) and the taiga (swampy, coniferous forest). This bizonal distribution of people stands in marked contrast to the ethnic distribution in North America, where traditionally the tundra was occupied only by Eskimo (including Aleut) and the taiga only by Indians. (For a map of the distribution of eastern and western Arctic ethnic groups, see ARCTIC PEOPLES AND CULTURES.)

PEOPLES AND LANGUAGES

The major cultural groups of the eastern Arctic may conveniently be classed into five categories: (1) the Lapp and Komi, living in far northern Europe; (2) the Samoyed, Ostyak, and Mansi, living west and east of the Ural Mountains; (3) the Tungus, living across much of Northern Siberia; (4) the Yakut, of Northeastern Siberia; and (5) the Paleosiberians (Yukaghir, Koryak, Chukchi, Kamchadal, Asiatic Eskimo, Aleut, and Ket), most of whom live in Far Eastern Siberia and on the Bering shores. It should be noted that some of these peoples, living in the Soviet Union, have adopted new names (cited below).

Lapp, Komi, and Samoyed

Of northern Finno-Ugrian-speaking peoples west of the Urals, the Lapp and the Komi are the only recognizable entities remaining today. The Lapp comprise small minorities in northern Norway, Sweden, Finland, and on the Kola Peninsula in Russia; their total population is 35,000; their principal livelihoods are fishing and reindeer breeding. The Komi, by contrast, are concentrated in northern Russia between the Pechora and Vychegda rivers (southeast of the White Sea); most are agriculturalists, but in the northern part of their area, some are hunters and reindeer breeders. They number 280,000. The Komi Autonomous Soviet Socialist Republic has been in existence since 1936.

East of the Komi and on both sides of the Urals lies the traditional area of the Samoyed peoples. Their languages and those of the Finno-Ugrians make up the Uralian language stock. The majority of the Samoyed are located along the coast and lower reaches of rivers between the White Sea and Taymyr, where they have adapted to a tundra ecology based principally on reindeer breeding, hunting, and fishing. These groups are now known in the Soviet Union as the Nentsy (20,000 persons), Nganasan (1,000), and Entsy (500). A fourth group of Samoyed, the Selkup (4,000 persons), is settled well to the south of the others and along the upper reaches of the Ob River. Their nearest non-Russian neighbours are two Finno-Ugrian-speaking peoples who moved to the east of the Urals: the Ostyak, now called the Khanty, who number 20,000, and the Vogul, now called the Mansi, who number 6,000. Earlier, the Selkup, Khanty, and Mansi were hunters and fishers only; later, they adopted reindeer breeding (Nentsy influence), fur trapping (tsarist influence), and fur farming (Soviet influence).

Tungus and Yakut

The Siberian home of most of the Tungus people, who speak what is called a Tunguso-Manchurian language, is in the vast area between the Yenisey and Lena rivers, to the east and south of the Samoyed. These Tungus are now known in the Soviet Union as Evenki; they number 25,000 persons. Like the Nentsy, they are traditionally reindeer breeders, hunters, and fishers principally; Evenki culture, however, is adapted to the taiga and not to the tundra. There are also the Dolgan, a small Tungus group of 4,000 persons, living north of the Evenki and in the southern part of Taymyr in ecologic conditions approaching those of Samoyed groups. The Dolgan are now Yakut speakers. One group of Siberian Tungus lives well to the east of the Yenisey–Lena area; they are the 9,000 Lamut,

now called Even, whose culture and language are held to be closely similar to the Evenki. The dispersed Tungus, or Evenki, population continues east and south beyond the Lena River to the Sea of Okhotsk and the Amur River; it is particularly in these areas that Evenki culture becomes blended with Yakut.

The Yakut, who number 230,000 persons, are distributed on both sides of the middle basin of the Lena and on to the Kolyma River in Northeastern Siberia. The Yakut speak a Turkic language, and it and the Tungus-Manchurian language of the Tungus combine to form the so-called Altaic linguistic family. Only a minority of the Yakut are reindeer breeders, most of them having kept cattle and horses; agriculture and hunting also are traditional practices. The Yakut may have moved northward into their present locations not long before the Russians arrived in that part of Siberia in the 17th century; local groups of other ethnic populations (besides Dolgan and Evenki, also Yukaghir and even Russian settlers) were assimilated by the Yakut, who have always had an important role as traders. Since 1922 there has been a Yakut Autonomous Soviet Socialist Republic.

Paleosiberians

Far Eastern Siberia is the ancestral home of some of the different peoples known collectively as Paleosiberians. They do not form a linguistic group of their own, and, unlike all other Arctic peoples of the Eastern Hemisphere, they do not have linguistic affinities with peoples to the south. They appear to have been in Siberia much longer than the other peoples, who are sometimes therefore designated collectively as Neosiberians. The Yukaghir, immediately to the northeast of the Yakut, are Paleosiberians who used to practice a primitive hunting economy over wide areas of tundra. Privation, disease, and assimilation (particularly to the Even, Yakut, and Russians) have since taken a severe toll, and there are now under 500 persons who call themselves Yukaghir. Many of them are reindeer breeders, and a third of them regard Russian as their native tongue. East of the Yukaghir are the Koryak and the Chukchi, both of whom are traditionally adapted to the tundra (reindeer breeding and hunting) as well as to the sea (hunting and fishing). Both offered relatively severe opposition to the Russians, the Chukchi not becoming Russian subjects until as late as 1789. Only the Chukchi expanded their territory and numbers in historical times, however, and this was at the expense of the Koryak as well as of the Yukaghir. The Koryak number about 6,000 and the Chukchi over 11,000.

Along the Bering Sea are 1,000 Asiatic Eskimo interspersed with Chukchi settlements. There are also about 1,000 Kamchadal, now called Itelmen, along the west coast of the Kamchatka Peninsula; the economies of both peoples are marine. The languages of the Koryak, Chukchi, and Itelmen are related. Finally, there are a few hundred Russian Aleut on Commander Island off the Kamchatka coast, and far to the west there are 1,000 Ket, who are hunters, trappers, and fur farmers living along the middle reaches of the Yenisey. Both of these peoples are included among the Paleosiberians; the language of the Ket is the sole survivor of a linguistic group earlier present in the area.

TRADITIONAL CULTURE PATTERNS

Ecological and economic systems. From earliest times to the present day, the way of life of many of these peoples has centred upon the reindeer; indeed, they have had to adjust their lives to the peculiarities of the species. Important in this regard are the seasonal migrations of reindeer between tundra and taiga; their movement in herds; their seasonal preferences for certain grasses, mosses, and other fungi; their dependence upon salt; their relative helplessness—and panic—when attacked by mosquitoes and other flying insects in the summer and such predators as the wolf and wolverine in the winter. The situation has always given rise to a number of similarities in all types of reindeer economy, among the more important of which are the seasonal movements of camps and the seasonal differences in their size and composition.

There are, however, distinctive ecological and economic differences among the Arctic peoples of the Old World, relating to each of the three principal ways in which they utilize reindeer. The traditional Paleosiberian Yukaghir, for example, were simply reindeer hunters. Although they "monitored" the seasonal movements of the reindeer, they made no effort to herd the animals or to use them for transportation (the Yukaghir adopted both these practices later, however, after contact with neighbouring peoples, particularly the Even). They used decoy reindeer in the hunt and travelled on foot or ski, by dogsled (though this was never general), and by canoe or raft. Reindeer were by far their most important source of food and of skins for clothing and tent covering, so that their lack of control over the condition and movement of the herds left the Yukaghir susceptible to unpredictable periods of privation. Fishing, the hunting of other species, and berry gathering were also essential to the economy.

Reindeer hunting, breeding, and herding

Another basic way of using reindeer is represented by the many Tungus groups that partly controlled reindeer breeding as an adjunct to a basic economy of hunting and fishing in the taiga. Geldings were ridden, used as pack animals, and sometimes harnessed to sleds; females were milked, making possible various dairy products; dead reindeer were an important source of skins, bone, antler, and sinew for clothing, shelter, and simple utensils. These Tungus secured their meat, however, not from their domesticated reindeer but from wild reindeer, Asiatic red deer, and elk that they hunted. The domesticated herd was always small (herds of up to 25 reindeer per household were usual); the animals usually pastured in the vicinity of the camp, securing protection from wolves there in the winter and finding relief from the summer mosquitoes in the smudge fires provided by the camp for that purpose. These hunting and herding Tungus, it may be noted, also used dogs—as camp sentinels to warn against predators or strangers and as "pointers" in hunting small game such as sable or squirrel.

The third basic method of using reindeer was controlled breeding and herding. This was the principal or sole livelihood of several other Tungus groups (many Nentsy and Entsy and some Nganasan), of the tundra Chukchi and Koryak in Northeastern Siberia, of the northern Samoyed in Western Siberia, and of many Lapp. All these groups depended upon the herded reindeer for food, and many used dogs to shepherd the herds, which sometimes were large in number. Prestige and social status, indeed, were measured by the size of a household's herd; and competition for prestige sometimes had damaging ecological consequences: pastures would be overgrazed or aggressively expanded, or epidemics would decimate large, closely packed herds. As a result, impoverished reindeer owners would be driven to find alternative livelihoods in villages or elsewhere; some would revert to reindeer nomadism. (Today there are state-imposed regulations in both the Soviet Union and Scandinavia ameliorating such problems.) As a supplement to herding, most of these peoples pursued fishing in the summer, and they bartered extensively with urban, village, and agricultural populations.

Social systems. The peoples of the eastern Arctic have been studied and written about far less than the Eskimo of the western Arctic. In tsarist Russia and the Soviet Union, in particular, there has always been some tendency to play down any attention to national differences. Most investigations have thus depended on inferential sources. They have tried, that is, to reconstruct the social relations of certain eastern Arctic peoples by piecing together information derived from myths, legends, and other folklore that have been published. They have also tried to infer the social systems of Arctic peoples by reference to the social structures of their Turkic and Mongol neighbours and kinsmen farther south, who have been more thoroughly studied—the assumption being that the systems are probably similar.

Tribes and clans

Among all the eastern Arctic peoples, there appear to have been large territorial groups that may be called tribes—*i.e.*, clusters of communities that could act together for such common purposes as warfare and trade. One tribe could make alliances and establish systems of exchange with other tribes. Below the level of the tribe,

at least among the Neosiberians, there were probably clans, groups of people having a stricter sense of kinship. Normally, clans would be exogamous; that is, marriage within the clan would be forbidden. Further, a person's lineage within the clan would be important in determining his residence, role, and inheritance.

From the 17th century onward, however, clans—and even tribes—had to leave their traditional territories and disperse in smaller units and "mixed" kin units to search for new territories. This happened to the Tungus when they were faced with the exhaustion of reserves of fur that had to be supplied to the Russians and when they were faced with the continual encroachment of Russian settlement. The movement of Tungus groups created pressures on the territories of other neighbouring peoples; the northern movement of the Yakut had similar effects in Northeastern Siberia. The competitive expansion of grazing lands could also disrupt traditional social relations, as it apparently did among the northern Samoyed. Various demographic and economic changes, in short, could have serious social consequences. Finally, there was the political influence. The tsarist Russian government tried to consolidate the tribe–clan system and use it to administer all northern peoples, but the administrative use of tribe and clan, being a foreign idea, tended paradoxically to sever the already weakening association between kinship and territorial organization. Among most groups, therefore, the village or camp emerged as the relevant territorial and social unit in the 19th and early 20th century.

Dominance of village and camp

Village or camp enterprises of a seasonal nature generally involved each of the three principal livelihoods: reindeer breeding, hunting, and fishing. When hunting and fishing was pursued collectively, the game or catch was distributed collectively. When production was not collective, responsibilities for distribution were limited, and they were minimal or absent when furbearing animals were caught for trade with foreigners. Villages and camps (and subdivisions made within these units) claimed rights to the produce from hunting and fishing grounds usually on an annual or lifetime basis. Among reindeer breeders, herding was undertaken collectively, and rights to pasture were claimed collectively, also. The animals, however, were owned individually. The size and composition of herding camps could be adjusted at each season of the year.

In all ethnic groups, close interpersonal bonds of friendship arose between non-kinsmen within the village or camp. Kinship nevertheless remained in many ways the most important principle of interpersonal and group relations, and many privileges and obligations of kinship were recognized. Among the more onerous was the prosecution of the blood feud, a retaliatory killing and counterkilling between different kin groups that might continue for generations. Blood feuds seem to have been widespread even at the end of the 19th century, indicating that clanship was still a recognizable principle but without any territorial meaning, for villages and camps were then composed of persons from different clans.

The incidence of cross-cousin marriage, in which certain categories of first cousins were obliged to marry, remained notable among many of the eastern Arctic peoples. Practiced in conjunction with the general rule of patrilocal residence, this marriage preference had earlier supported the territorial groupings of persons in clans; the preference remained even after the development of villages and camps in which cousins remained anyway, without the compulsion of marriage. The matching of several siblings of one family with those of another and a close bond between husbands of sisters—sometimes denoted as closer than that between brothers—were prevalent among the northern Eurasians. So, too, was the injunction that a bridegroom provide several years' "service" to his father-in-law before being allowed to set up his own marital household. Common among the reindeer breeders only was the establishment of marriages through contributions of wealth (including reindeer) from both the parties (bride price from the groom, dowry from the bride). Having two or more wives was a privilege of personal rank and occurred most frequently among preeminent hunters or warriors, shamans and smiths, and well-to-do reindeer owners. These marriage practices (except those prohibited by law) persist among some contemporary Lapp groups in Scandinavia. Among the Lapp, building a herd was probably the most influential factor governing marital patterns.

In all these societies there have been strains of social and economic inequality. At the very least, the need for authority in organizing work teams promoted village or camp elders, local headmen, lead hunters, and so on; among some peoples—the Chukchi are notable—an ethic of success and of male reputation licensed the presence of "strong men" and aggressive behaviour. Intertribal warfare, fairly general though sporadic throughout northern Eurasia, led to the use of captives as slaves. Among the reindeer breeders, differential accumulation of herd capital placed some herdsmen in employer–employee relationships.

Animism and shamanism

Belief systems. Traditionally, there were broad similarities among the spiritual cultures of all the northern Eurasians. All tended toward animism—*i.e.*, a belief that everything in nature, every beast and bird, stone and earth, wind and snow, is alive with soul or spirit—and all believed in shamanism; indeed, the word shaman is of Tungus origin (*šaman*). The shaman was priest, medicine man, and prophet; that is, he directed communal sacrifices and ceremonies; he cured illnesses, and he foretold the future. He acquired his profession either by inheritance or by what might be called spontaneous calling or "election," but in any event he ordinarily had to go through initiations or trials of self-abnegation and a kind of hysteria before he acquired the ability to "see" spirits, behave like a spirit, and leave his body and travel in all cosmic regions like a spirit. It was these presumed abilities that were thought to qualify him as a healer, for the cause of illness was thought to be the straying of one's soul, and only the shaman was deemed able to search out the soul and return it to the patient's body. Siberian shamans were frequently women; male and female transvestism in shamans was common.

The shaman's power was often associated with the properties of iron, and there was a corresponding association of the profession of shaman with that of smith. The role of the smith had particular prestige among the Tungus and Yakut.

Christianity reached these peoples at different historical periods; some Lapp were converted as early as the 16th century, whereas some Chukchi probably were never converted, even before the Russian Revolution discouraged religious faith. Christianity, when it was adopted, usually came in the wake of trade with the Western world and with the West's imposition of taxes. In any event, animism and shamanism rarely died completely but became blended with Christian beliefs, rituals, and liturgy. Christian motifs, for instance, appeared in the drawings on Lapp shamans' drums; Christian crosses and crucifixes were worn together with amulets and other talismen of animistic belief.

CONTEMPORARY TRENDS

Collectivization and russification

Ethnic changes in the Soviet Union. The abolition of serfdom in 1861 brought millions of Russian peasants to Siberia in a comparatively short time; Soviet settlement, development, and industrialization during the 20th century has affected the demographic structure of Northern Siberia even more. Although it must be recognized that no non-Soviet observer has been permitted to study any of the northern peoples since the 1930s or even earlier, some historical trends and events are clearly evident. First, it can be noted that attempts to deal administratively with northern peoples began as early as 1922. Then, between 1924 and 1935, a Committee for Assistance to the Peoples of the Northern Regions (the Committee of the North) sought determinedly to bring the native peoples into the Soviet administrative structure and to establish among them "integral cooperatives," collective farms, and educational, medical, technical, and social programs. In 1925 Leningrad State University accepted the first stu-

dents from among the northern peoples, and in 1930 the Institute of Northern Peoples was established at the university with 195 pupils in attendance from 19 ethnic groups. The institute closed in 1941, but its work was continued after the war at several regional institutions. Among its principal objectives has been the training of native persons as teachers and administrators.

The editors of the monumental Soviet work *Narody Sibiri* (published 1956; Eng. trans., *The Peoples of Siberia*, 1964) stated that, "By the present time, the village soviet has become the principal local unit of Soviet administration among the peoples of the North, while the collective farm now is universally the fundamental economic unit." The progress toward this state of affairs, however, was beset by difficulties and controversies. The Committee for Assistance to the Peoples of the Northern Regions, for instance, ruled, in 1926, in favour of "clan-based native soviets," but this arrangement proved unworkable because the clan elders were not too willing to support, explain, and enforce Soviet policies. Accordingly, the native soviets were redesigned between 1929 and 1932 as territorial units at the bottom rung of the standard Soviet administrative structure; above the soviet was established the region (*rayon*) and then the district (*okrug*). The northern *okruga* were named after the dominant minority people of the region; by the end of 1932 there were nine such *okruga* of northern peoples, 82 *rayony*, and 462 soviets. Native membership in the executive committees of these *okruga* and *rayony* has varied from 33 to 97 percent.

The progress of collectivization among the northern peoples was recorded as 12 percent of households in 1931, 36 percent in 1934, 75 percent in 1940, and 97 percent in 1948. The program met its chief resistance from reindeer breeders. The problem at heart is a conflict between ideology and ecology or economics. Official Soviet ideology holds that nomadism is an unworthy way of life in a Socialist society, but this is not easily reconciled with the ecological fact that reindeer pastoralism can be conducted successfully only on a nomadic basis, for reindeer must move seasonally. These difficulties have adversely affected not only production in the Soviet reindeer industry but also recruitment to it. Lately, however, ethnographers have tried to persuade administrators and other officials that "production nomadism" is necessary and is compatible with an otherwise stable or sedentary way of life. Probably of greater significance for the future is the issue of whether northern peoples should be encouraged to join the Siberian industrial force or, on the contrary (as Soviet ethnographers seem to urge), be encouraged to remain in their traditional pursuits and to retain their traditional skills.

In education, a number of native grammars were published and used in schools in the early 1930s. The change from a Roman-based to a Cyrillic-based alphabet in 1937, however, presaged greater Russian influence. This seems to have continued, and Russian is now the lingua franca of much of the north, although the Komi and Yakut languages still appear to retain a respected status.

One of the most notable aspects of ethnic relations in the Soviet north today is its complexity. Even at the level of the village soviet, there can be polyethnic membership, and this mixture of representatives from various ethnic groups increases at the levels of the *rayon* and the *okrug*; there are polyethnic collectives such as the Nentsy–Entsy–Nganasan–Dolgan collective in the Taymyr and the Yukaghir–Lamut–Chukchi one in the northeast; Russians also sometimes belong to northern peoples' soviets and collectives. Some ethnic groups—the Yakut and the Komi—are deemed to have reached the level of "Socialist nation"; others, having national *okruga* and written languages, are perhaps progressing in that direction. But the majority are unlikely ever to achieve nationhood. It seems clear that they will either remain as minorities or disappear. Russification is the dominant trend.

The Lapp in Scandinavia. The Lapp of Norway, Sweden, and Finland today form small minorities even in the northern regions of these countries, but they are receiving greater recognition than ever before. The explanation may lie more in the changing attitudes of Scandinavians, who are becoming more tolerant of diversity, than it does in any activism on the part of the Lapp themselves. In any event, the Lapp of Sweden, for instance, have their own ombudsman, whose work is particularly concerned with the legal position of Lapp groups faced by public or private encroachments on their reindeer pastures. There are Lapp language primers, and school readers now available in each of the three countries (though by no means are all schools with Lappish-speaking children using them); and there are Lappish high schools, Lapp-language newspapers and radio programs, and Lapp-language courses in several Scandinavian universities.

Combined ethnic and national trends among the Lapp

The reindeer-breeding industry is, in general, in good condition. In Norway it is predominantly a Lappish livelihood, and in Sweden it is exclusively so (in Finland, however, the Finns predominate in most reindeer breeding localities). The nomadic way of life associated with reindeer pastoralism is becoming seriously modified, but most of the modifications either are initiated by individual households or camps (and hence vary locally) or emanate from the regional and national associations that have been formed by the reindeer owners themselves. Direct legislation by the state is less important in this process, though the indirect influence of the state—as the principal source of capital loans, subsidized housing, and infrastructure amenities—should not be minimized. Cooperatives for the purchase and marketing of reindeer products are established in many areas.

Lappish individuals and communities today, without exception, have a heightened sense of national citizenship, and all persons except the very old and perhaps the very young are conversant in the language of the country of their citizenship. Indeed, individuals who leave reindeer breeding are likely to assume a non-Lappish identity: in Sweden such persons usually join the northern industrial labour force or move to Stockholm. In Norway, where the majority of Lapp reside, most families of Lappish descent have never been involved in reindeer herding; most of them have been and are fishermen. Among these people, Lappish culture, in all its respects, is fast being replaced by Norwegian culture.

BIBLIOGRAPHY. The outstanding work of reference compiled by Soviet ethnographers is *The Peoples of Siberia*, ed. by M.G. LEVIN and L.P. POTAPOV (1964; orig. pub. in Russian, 1956). Two earlier general works that retain their importance are MARIE ANTOINETTE CZAPLICKA, *Aboriginal Siberia: A Study in Social Anthropology* (1914); and WALDEMAR JOCHELSON, *Peoples of Asiatic Russia* (1928). Siberian ethnographies are few and, for the most part, date from the turn of the century; outstanding are *The Chukchee* (1904–09) and *The Eskimo of Siberia* (1913), both by WALDEMAR BOGORAS; *The Koryak* (1905–08), *The Yukaghir and the Yukaghirized Tungus* (1910–26), and *The Yakut* (1933), all by WALDEMAR JOCHELSON. These five titles are publications of the American Museum of Natural History. An early report from the Samoyeds is KAI DONNER, *Bei den Samojeden in Sibirien* (1926; Eng. trans., *Among the Samoyed in Siberia*, 1954). The indispensable work on the Tungus is S.M. SHIROKOGOROFF, *Social Organization of the Northern Tungus* (1933, reprinted 1966); the author, however, is concerned with Tungus of the Manchurian–Chinese–Russian border area. Important reviews of aspects of the contemporary situation are TERENCE ARMSTRONG, "The Administration of Northern Peoples: The USSR," in R.ST.J. MACDONALD (ed.), *The Arctic Frontier* (1966); STEPHEN and ETHEL DUNN, "The Transformation of Economy and Culture in the Soviet North," *Arctic Anthrop.*, 1:1–28 (1963); and ETHEL DUNN, "Educating the Small Peoples of the Soviet North: The Limits of Cultural Change," *ibid.*, 5:1–31 (1968). A great deal of ethnography has been, and continues to be, written about the Lapp of Scandinavia. Two outstanding general works are BJORN COLLINDER, *The Lapps* (1949), particularly strong on the culture of the Lapps of Sweden; and ORNULV VORREN and ERNST MANKER, *Samekulturen*, 2nd ed. (1958; Eng. trans., *Lapp Life and Customs: A Survey*, 1962), an overview of all the Lapp groups of Scandinavia with particular attention to the various ecologic divisions. The contemporary situation of the Lapp in Finland is outlined in KARL NICKUL, *Report on Lapp Affairs* (1952). HARALD EIDHEIM, *Aspects of the Lappish Situation* (1971), provides a social anthropological analysis of the Lapp minority situation in Norway.

(R.P.B.P.)

Eastern Christianity, Independent Churches of

Historical and cultural importance

The Eastern Christian churches that do not recognize the authority of the pope in Rome or the ecumenical patriarch in Istanbul, such as the Syrian, Armenian, Coptic, and Ethiopian churches, are of great historical and religious importance (see also EASTERN ORTHODOXY, HISTORY OF; CHRISTIANITY BEFORE THE SCHISM OF 1054; EASTERN RITE CHURCHES). Most of them are of very ancient origin, reaching back to the subapostolic (late 1st to early 2nd centuries AD) or even to the apostolic age (1st century AD)—to the time of the first Christian communities. Syrian Christianity possesses special significance, since primitive Syrian Christianity derived its traditions not via Antioch (in Syria) and the Hellenistic (post-Alexandrian Greek) Christianity of St. Paul but directly from Palestine, taking over tenets from Palestinian (Aramaic-speaking) Christianity.

A very important access to the understanding of early Palestinian Christianity is thus opened to scholars that would otherwise be closed. Moreover, with the emergence of the Christian faith in the early Christian centuries among the Eastern peoples from Egypt to India, the Eastern churches contributed to a transformation of popular and national life and thought—by inventing alphabets, translating biblical and Christian writings, and inaugurating an era marked by national literatures and cultures. The development of many of the Eastern nations is bound up with the development of the independent Eastern churches. The profound spirituality of these churches, which was expressed in asceticism, mysticism, a deep vision of the Christian life, and a burning zeal for the missionary enterprise, reached out beyond regional boundaries and enriched the lives of other nations. The influence of their intellectual creativity also extended to Europe and to non-Christian cultures.

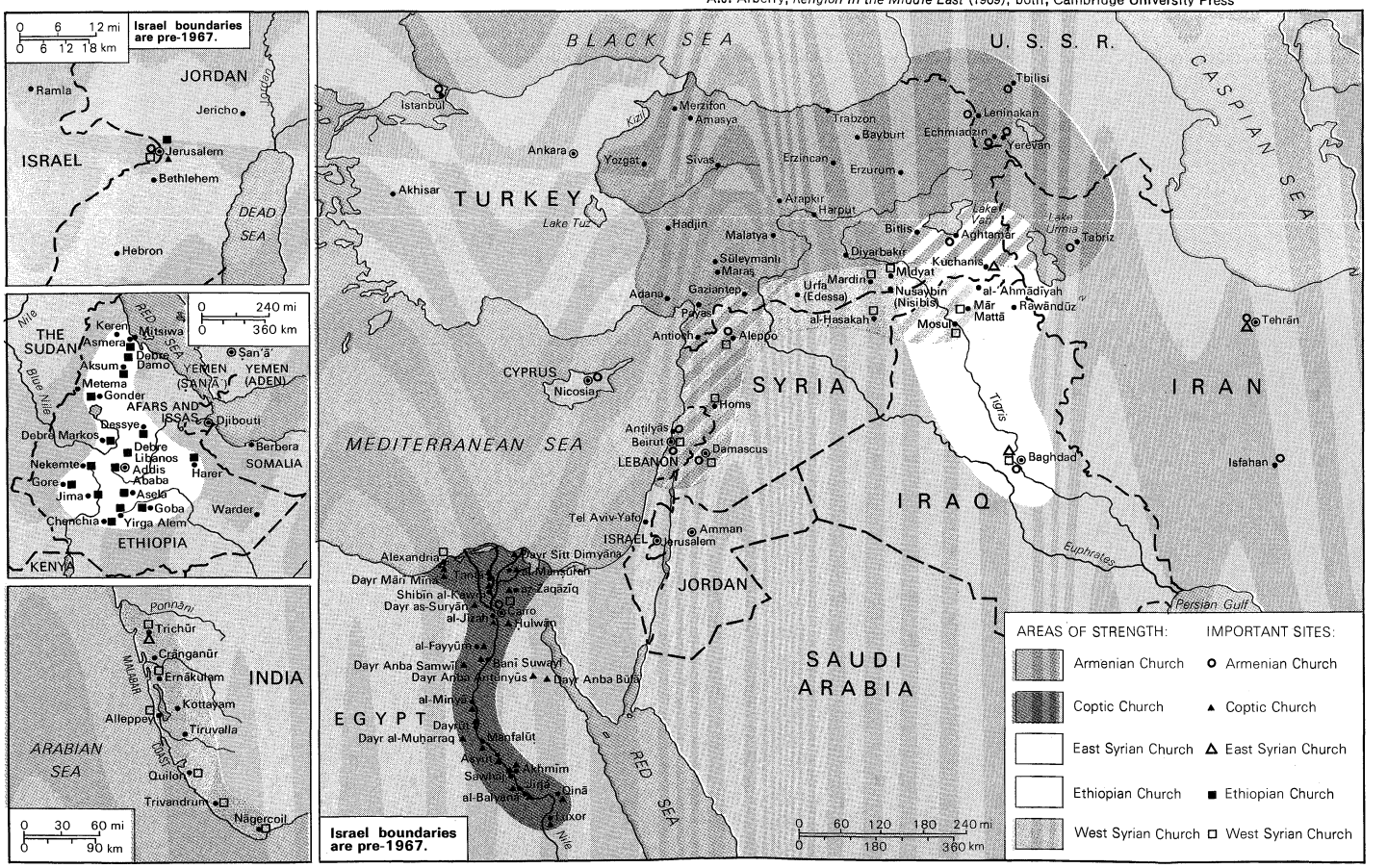
Yet, despite their indubitable historical and cultural significance, the history of the Eastern Christian independent churches has been relatively neglected by Western scholars. Various interests, biases, and technical problems have contributed to a dearth of knowledge of the Eastern independent churches. Western-centred views of the history of culture, such as the view of the 19th-century German philosopher G.W.F. Hegel—that the spiritual sun, unlike the physical sun, rises in the West and then radiates its light to the Eastern sphere—have contributed to this neglect, as have the difficulties in mastering the Eastern languages (Syriac, Armenian, and others). The result is that for most of these churches there is no real history beyond that of a general, sketchy survey. Historical research concerning these churches is still in a preparatory stage, since manuscript sources must first be unearthed and gathered before historical judgments based on documentary evidence can be made.

The Western approach to church history has been dominated by the study of the Western Church and hence lacks the perspective and breadth necessary to grasp all of Christianity in its fullness in thought and action. This deficiency can, in part, be rectified by increased knowledge of the Eastern branches of Christianity through more research and study.

THE WEST SYRIAN CHURCH: THE SYRIAN ORTHODOX CHURCH OF ANTIOCH

Origins in the doctrines of Monophysitism. The 5th-century Monophysite doctrine—that Christ is not in two separate natures (divine and human) but is one nature out of two natures—found a zealous following among the Syrians. The decision of the ecumenical Council of Chalcedon (451) against Monophysitism rallied a large body of Christians in Syria and adjacent areas who were against Greek dominance to support the Monophysite movement, although the towns, because of their Hellenistic elements, remained loyal to the orthodox two-nature doctrine of Christ formulated at Chalcedon. During

Adapted from (bottom left) L.W. Brown, *The Indian Christians of St. Thomas* (1956), (others) A.J. Arberry, *Religion in the Middle East* (1969); both, Cambridge University Press



Distribution of independent Eastern Christian churches.

the Acacian Schism (484–519), created by the doctrinal conflict over the decision of the Council of Chalcedon between the patriarch Acacius of Constantinople and the Roman pope Felix III (483–492), the Syrian Church became Monophysite.

Severus and John bar Qursos (John of Tella)

The founder of Syrian Monophysitism was Severus, patriarch of Antioch (512–518). After a temporary victory of Severus over the exponents of orthodoxy, the situation became very critical for the Syrian Church. During the reign (518–527) of Emperor Justin I, a synod in July 518 made a decision to condemn Severus. In the following year, intensive persecution of the Monophysites swept through the Syrian patriarchate, and diocese after diocese lost its bishop. The acute shortage of clergy became a life-and-death issue for virtually every Monophysite community. At conclaves of bishops, they all refrained from ordaining priests for fear of official governmental and ecclesiastical reprisals. At this juncture, John Bar Qursos, bishop of Tella, volunteered to undertake the risk of ordaining priests without support from Constantinople. From the Persian frontiers to Armenia, Cappadocia, and Phoenicia, he encouraged, instructed, and examined candidates for ordination and performed mass ordinations. His heroic efforts soon began to show results. Depressed communities felt a quickening spirit. Candidates for ordination came to him "like a flood that is produced in a river by thick clouds" (John of Ephesus) wherever he appeared—in monasteries, on the roads, and even in the desert. The Syrian Oriental churches thus successfully withstood this first harsh test that threatened their existence.

When, under Emperor Justinian I (527–565), Pope Agapetus personally took the problem of the Syrian churches into his expert hands, the outlook for the Monophysites became bleak. The renewed persecutions brought consequences of the utmost gravity—the Syrian flock found itself once again without shepherds. When the Monophysite cause was at its lowest ebb, Ḥārith ibn Jabalah, king of the Arabs, appeared in Constantinople in 542–543, determined to create a closely knit Monophysite realm. He demanded two or three bishops from Justinian's wife, the empress Theodora, who favoured the Monophysite cause, and she complied.

The reorganization under Jacob Būrd'ānā (Baradaeus). One of the most decisive events in the history of the Monophysite movement then occurred when Patriarch Theodosius of Alexandria (536–567), a follower of the views of Severus, intoned the ceremony of episcopal consecration for two monks: Theodoros of Arabia as bishop for the Arabs (542/543–571/572) and Jacob Būrd'ānā ("garbed in a patchwork garment") as bishop of Edessa (542/543–578), which actually included all the area of the Syrians.

The number of ordinations Jacob Būrd'ānā (latinized as Baradaeus) performed in a necessarily clandestine fashion is reported to have been extremely large (100,000). Though he was pursued and had a price on his head, he was never caught by Byzantine imperial or church agents. His tireless work forged the reorganized communities into a church body, creating the national Syrian Church.

Role of monasticism in the organization of the church

In this process, a large share of credit belongs to the monastic movement that was fostered by such leaders as Rabbula of Edessa (411/412–435), John bar Qursos (519–538), and particularly by Philoxenos of Mabbug (485–523). This Syrian type of monasticism was characterized by the enthusiasm, zeal, and rigour manifested by such personages as the stylites (ascetics living on a pillar) and recluses. The huge, enthusiastic army of ascetics and monks rendered extremely important services to the work of the organization. Jacob Būrd'ānā then established the hierarchy: two metropolitans each were appointed in Asia Minor (Seleucia, in Isauria, and Taurus) and in Syria (Laodicea, Seleucia), and one for Mesopotamia (Amida). In consecrating the patriarch Sergios of Tella, Jacob filled the highest post in the ecclesiastical hierarchical structure of the Syrian Monophysites. Because of the merits of Jacob, the Syrian Monophysites after the 8th or 9th century began to call themselves Jacobites, but the designation is no longer preferred.

Developments after the Arab conquest. The Arab conquest in the 7th century was hailed by the Syrians because it freed them from the oppression and tyranny of Byzantium. Since the new situation was a blow to the Hellenistic elements in the towns, the indigenous Syrian Church benefitted. The native church began to develop its own way of life without restrictions. It expanded into northern Mesopotamia, with a centre in Mosul, where its efforts could be pursued under new conditions. The seat of the primate (called maphrian in Syriac and also catholicos) of the Orient was established in Tagrit, with Marutha (629–649) as the first maphrian. The spiritual life of the area also benefitted from the new situation. Literary works began to be produced freely, especially works on the Bible, asceticism, spirituality, and the liturgy. The liturgy of the Syrian Church is one of the oldest and richest in Christendom. The oldest collection of liturgical manuscripts in the monastery of Za'farān contains many Anaphoras (prayers of consecration in the liturgy). The Syrian Liturgy of St. James (which is still one of the forms of worship of the Syrian Church) has its roots deep in antiquity. In addition to liturgical renewal, a number of monastic communities became involved in education, literary life, and scholarly work. Scholars such as Severus Sebokht (died 666/7), Jacob of Edessa (died *c.* 708), Dionysios Tell Mahre (died 845), and later Michael the Great (died 1199) and Bar'Ebrāyā (died 1286) enabled the Syrian Church to assume leadership in theological thought in Monophysitism, a role that it retained until the end of the Middle Ages. All of this was fostered by conditions which, under the Arab rule during the dynasty of the Umayyads (661–750), were liberal and which remained tolerable under the 'Abbāsids (750–1258). During the dynasty of the 'Abbāsids, the Jacobites enjoyed an influential position at the court of the caliphs at Baghdad.

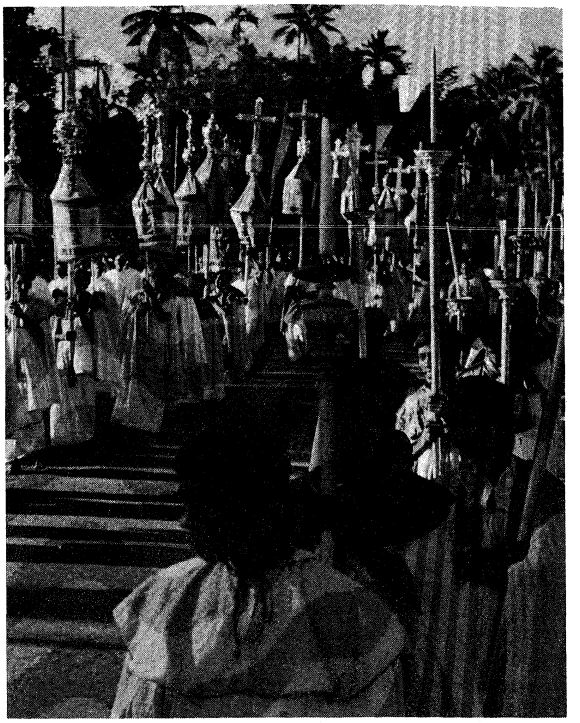
Liturgical and theological developments

In the 12th century the Jacobite Church reached its widest expansion: 20 metropolitan seats and 103 eparchies (bishoprics) in Syria, Mesopotamia, and Cyprus, with about 2,000,000 believers. The decline in importance and numbers, however, came quite rapidly. Islām took its toll by conversions that were fostered by internal disputes and strifes, including a schism (1364–94), and the catastrophes (the Mongol invasions) in the 13th and 14th centuries reduced the number of adherents; by the 17th century there were only 20 bishoprics. In the same century the Syrian Church succeeded in winning a part of the Christians of St. Thomas in Malabār (Kerala), South India, namely those who were not incorporated under the jurisdiction of Rome when the church in Malabār, which had been under the Catholicos of Seleucia–Ctesiphon, was thrown into conflict because the Roman Catholic Curia had given to the Portuguese jurisdiction over all the Christians of that area. This gain, however, was quickly balanced by the loss to Rome of a part of the Jacobites, who became Syrian Catholics. The remaining body has been known as the Syrian Orthodox church.

Modern developments. Until World War I, most members of the church lived in northern Mesopotamia, in the region of Mardin, Midyat, Diyarbakır, and Urfa (all in Turkey). These compact communities were subjected to savage persecution and massacre in 1914–18 by the Turks. Those who escaped from the persecution established communities in Syria, Lebanon, and Iraq. Today, about 80,000–90,000 Syrian Orthodox are scattered in communities in Iraq, Lebanon, Syria, Jordan, Egypt, Turkey, and the United States. Of the once flourishing communities in north Mesopotamia, Midyat is the only town that has retained its Christian character. Other Syrian Orthodox communities, except the one in Istanbul, are very small.

Modern persecutions

The patriarchal see that was located in Mardin was moved to the monastery of Za'farān, then to Mosul, Iraq, to Homs, Syria (1921), and finally established in Damascus, Syria (1959). Since 1293 all the patriarchs bear the surname Ignatios in honour of Ignatius of Antioch (the early martyred bishop). The present patriarch, Ignatios XXXIX Jacob III, rules from Damascus over 11 metropolitans and three bishops and over the Syrian Orthodox Church in Malabār. His actions toward healing a 50-



Procession of Christians of St. Thomas (Syrian Church) bearing silver crosses, Kerala, India.
Bernhard Moosbrugger

year-old rift with the church in India reached success in 1958. He has fostered education (for example, the St. Aphrem theological school at Atshaneh, Lebanon), and through his visits to the congregations in several continents strengthened the life and unity of the church.

THE EAST SYRIAN CHURCH

Origins in the doctrines of Nestorianism. The belief that there were two separate natures in Christ—the one divine and the other human—was a deep-rooted conviction in the Antiochian (Syrian) school of biblical interpretation. The movement that rallied behind this doctrine is connected with the name of Nestorius, patriarch of Constantinople (428–431), who was involved in a violent theological controversy that ended with his deposition by the Council of Ephesus (431). His adversaries were so successful that five years later Nestorius—the least besmirched figure in the whole net of ecclesiastical and political intrigues—was banished to Upper Egypt to end his life in obscurity (*c.* 451). The decrees of the Council of Ephesus, however, were not able to wipe out the opposition movement (later called Nestorianism).

After the theological settlement of 433 between the patriarchs Cyril of Alexandria (reigned 412–444) and John of Antioch (died 441/2), the bishops who could not accept this agreement were driven closer to the Christians outside the Byzantine (Orthodox) territory, namely, in Persia. This party found a centre in the Christian School of Edessa that had been founded by Syrian refugees from Persia and was under the leadership of Ibas (or Hiba; died 457), who had supported Nestorius at Ephesus. Renewed theological disputes flared up at Edessa, and due to agitation by the Monophysites, the teachers and a number of students were compelled to flee and to seek refuge in Persian territory, mainly at Nisibis. This gave another impetus to various movements to form a church independent of Antioch.

Organization of the Nestorian Church. The Persian Church in 486, at the synod of Seleucia–Ctesiphon, officially accepted the Nestorian creed, and the Nestorian Church was established under the leadership of the catholicos at Seleucia–Ctesiphon. This move was motivated by ecclesiastical and political considerations in order to dissipate Persian suspicions regarding links with Byzantium. At a synod in 612, the dogmatic position of the Nestorian Church was definitely fixed according to the formulations of Babai the Great (died after 627/8). The church rejected the condemnation of Nestorius at Ephesus and elevated him, together with Diodorus of Tarsus (died *c.* 390) and Theodore of Mopsuestia (*c.* 350–428/429), to the position of its major authorities. A vigorous church emerged and flourished, in spite of periods of persecutions under the Zoroastrian Sāsānids (226–657), until 643, developing its own way of life.

Developments after the Arab conquest. These conditions continued after the Arab conquest as long as the Umayyad dynasty lasted. Conditions changed to some degree, however, with the ʿAbbāsids. There were occasional outbursts of oppression, but the church was able to continue without any serious danger to its basic structure and concerns. Under the Islāmic caliphs, the Nestorians became spokesmen for all Christians in their empire; and under some caliphs, church leaders enjoyed a favoured position. Catholicos Timotheos I (780–823) was regarded as a protector of a considerable number of Nestorian elite: physicians, pharmacists, merchants, and scholars. The Nestorians utilized the favourable conditions to their advantage.

Ancient liturgical traditions

The Nestorian Church has retained many of its ancient traditions. The canon of the New Testament reflects an early stage in development by its exclusion of all the Catholic Epistles (those attributed to James, Peter, John, and Jude) and the Apocalypse (Revelation). The liturgy, due to the Nestorians' isolation, has preserved its very primitive character. The chief liturgies are those of Mār Addai and Mār Mari, Theodore of Mopsuestia, and Nestorius. The celebration of the Eucharist begins with the ritual of making the eucharistic bread for which the "holy leaven" has been used. The Nestorians believe that the holy leaven literally can be traced back to apostolic times. Their avoidance of pictures and icons in the churches puzzled the first visitors from the West after the Reformation and left them with the impression that the Nestorians were "Protestants of the East."

The establishment of the Nestorian Church as a reaction against the West also included a reaction against the predominance of ascetic forces, so that marriage was required even of the ecclesiastical hierarchy, thereby alienating the ascetic circles and driving them into the Monophysite fold. But in 544 the church revived its stand in favor of celibacy, thereby paving the way for reconciliation with monasticism.

Abraham of Kashkar (died 588) founded the monastery of Izlā (near Nisibis), which became a centre of reform monasticism. Under its aegis, a network of monastic communities was established. Spiritual, ascetic, and mystical literature was produced in these monastic centres that nurtured piety, education, charity, and intellectual culture.

Missionary activity

The religious zeal of the Nestorians is indicated by their missionary outreach. Through the centuries, this church became the most widespread of the Eastern Christian churches. Not only Syria and Persia but also Arabia, Media (Holvan), Khorāsān (Merv), Herat, Samarkand, and Turkistan were included in its mission field. Nestorian missionaries early reached India. The use of Syriac as the ecclesiastical language of the Christians in South India, known as the Christians of St. Thomas, indicates their origin, and for centuries they were under the jurisdiction of the catholicos in Seleucia–Ctesiphon. Remnants of Nestorian manuscripts testify to their missionary work in Chinese Turkistan and Turfan, and literary data to that in Tibet. The monument of Hsi-an (Ch'ang-an) of 781 gives a summary report of the success of Nestorianism in China. In the 13th century, various tribes of Mongols were converted to Christianity. At the beginning of the 14th century, there were ten metropolitan sees in Central Asia, developing an indigenous clergy. The missionary activity of the Nestorians indicates its tremendous spread: the catholicos in Seleucia–Ctesiphon presided over an area that covered an immense territory from Cairo to Peking.

The Nestorian churches also produced important centres of learning. The School of Nisibis supplanted the

School of Edessa as the main centre of Nestorian culture; it influenced the founding of many other schools and the development of higher education in Europe. These schools became the focuses for the cultivation of a literary life that was marked by its manifoldness and its creation of new genres. Its amplitude was marked by such names as Narsai (died 503), Babai (died after 627/8), Ḥenānā (died *c.* 610), and Ḥunain ibn Isḥāq (died 873). The Nestorians thus have secured an important position in the history of culture by transmitting to the Arabs the heritage of late antiquity in philosophy, medicine, and natural sciences.

After the 10th and 11th centuries, the Nestorian Church suffered increasing losses to Islām; the Syrian renaissance, which, however, produced such a manysided Nestorian scholar as ʿAbdīšōʿ (died 1318), was very influential before its extinction. The Mongol invasion, with its massacres and destruction under Timur (1336–1405) in 1380, nearly annihilated the Nestorian Church. A remnant escaped the destruction under Timur and found a place of refuge in the mountains of Kurdistan, where its descendants, known as Assyrian Christians, continued their existence. A union with the Roman Catholic Church in the 16th century threw their communities into an internal crisis that subsequently weakened the position of the Christians with the Muslims. As a result, the Chaldean Church was constituted, embracing the Nestorian faithful in Iraq, Iran, and Syria. (Its membership today is about 96,000.) The efforts of the present patriarch, Paul II Cheikho, are marked by lively fostering of church life, monasticism, and education as well as by endeavours toward healing of deep wounds, which have destroyed several bishoprics in the north, caused by the war between the Iraqi government and the Kurds. The vigour of the church is manifested by the expansion of church work and the establishment of several new eparchies. The Nestorians in Malabār entered into a union with Rome in 1599 and today number about 1,800,000.

The Chaldean Church

Modern developments. In the 19th century, Anglican missionaries attempted to revive the Assyrian Christians, and the Nestorian Church came under Anglo-Saxon influence. In World War I, a new tragedy struck the Assyrian Christians with their expulsion from the territory of the Kurds in the old Ottoman (Turkish) Empire. They were persecuted by the Kurds and, according to some estimates, more than a half of them died. Their hopes to found their own state collapsed, and the remnants gathered in Caucasia, Iran, Syria, and particularly in Iraq. After 1933 new bloody conflicts brought new sufferings to the Assyrian refugees. Attempts to practice full autonomy in matters of civil authority evoked most serious conflicts with the new state of Iraq, which could not tolerate a state within a state. This caused an exodus to America. Since 1940 even the patriarch, Simon XXI (from a family in which the office has been hereditary for centuries), has made his residence in the United States (San Francisco). The church has more than 80,000 believers in Iraq, in Syria about 30,000, in Iran about 20,000, in Malabār about 8,000, and in America about 25,000.

THE ARMENIAN CHURCH

Origins under Gregory the Illuminator. The Armenians look to Gregory the Illuminator (240–*c.* 325) as their apostle. Consecrated by the metropolitan of Caesarea, Cappadocia, Gregory converted King Tiridates (Trdat) III and through him introduced Christianity to Armenia (before 300). Though Christianity barely penetrated to the lower levels of Armenian society, the reputation for having established the first Christian state belongs to the Armenians. In his work, Gregory was supported by Armenians under Hellenistic as well as Syrian influence and by a contingent of Syrians. The Syrians were particularly instrumental because they introduced a zeal for monasticism as well as for evangelistic work that inspired the Armenians to carry the missionary work over into Georgia. At the beginning, all worked peacefully together, and Greek and Syriac became the ecclesiastical languages.

In 374 the Armenians rejected their dependence on the Church of Caesarea. Repudiation of Greek and Syriac, adoption of their own idiom, creation of their own alphabet, and translation of the Scriptures under Catholicos Sahak (390–441) and Mesrob (Mashtots; died 439/440) all created conditions to weld together a religion, culture, and nationality. The Armenians were not represented at the Council of Ephesus (431), but when the acts were sent to Armenia, they were approved by Sahak. They also were not represented at the Council of Chalcedon (451) because at that time they were involved in mortal struggle with the Persians, and it was not before the end of these battles in 484 that they could resume contact with the Byzantine Church.

Rise of the indigenous Armenian Church

At this time, the Byzantine emperor Zeno (died 491) propagandized his letter to the Egyptian Church, the *Henotikon*, which was similar to the Monophysite formulations. During the Synod of Dvin (505/6), relations were established with the Monophysites in Syria and Persia. The formula, *mia physis tou theou sesarkomenē* (one incarnate nature of the godhead), coined by Cyril of Alexandria, was accepted; but it is not clear how Monophysitism was understood at that time, except that the positions of the early proponents of the Monophysite view, Eutyches, the patriarch of Antioch, and Severos, were rejected. Ancient Armenian Christology (doctrines about Christ) apparently was under the influence of Julian of Halicarnassus (died before 538). Most of the theologians, however, were not interested in Christological problems, and those who were were not in agreement as to whether Monophysitism should be interpreted according to its definition or according to its content. The formulations in the Armenian liturgical texts indicate, however, that the Armenians were more tolerant with the diophysite (two natures) terms than the West Syrians.

Developments after the Arab conquest. After the end of the persecutions under the Persians, the existence and development of the Armenian Church were guaranteed under the Arab rule, owing to the more tolerant position of Arabs toward the Christians—regardless of conflicts and occasional persecutions under the ʿAbbāsids in which churches and monasteries were destroyed. Frequent insurrections aggravated the condition of the Armenians, but in spite of their sufferings, their faith and nationality grew into an inseparable unity that became a protective wall preventing apostasy to Islām.

In the 9th century, when the Muslim rule became weaker, the Armenians succeeded in gaining some independence; and under the native dynasty of the Bagratids (885–1046), a cultural renaissance was initiated, and many exiled Armenians returned to their country. The Armenians devoted themselves to the cultivation of their own national life, manners, and customs. In their liturgical traditions, they were in debt to Greek and Syrian sources; but in their ritual, they developed their own idiosyncratic forms that became objects of heated controversies with the Syrians. In their Eucharist rite, the Armenians, alone among the Oriental churches, had adopted the use of unleavened bread. They did not mix water with wine in the Eucharist or use olive oil for the myron (oil of chrism or confirmation). They brought sacrificial animals into the church at Easter time—explaining this custom as a heritage from the days of Gregory—and celebrated the Nativity of Christ and the Epiphany together on January 6 according to an archaic Christian tradition. Disregarding the practice of such idiosyncratic customs, a not insignificant number of Armenians were inclined toward union with the Orthodox Church, and one overanxious catholicos (Vahan I [967–969]) even lost his office in the attempt. The great masses, however, remained loyal to their own church and its ways.

Liturgical practices

After this period of national efflorescence, Armenia experienced a long period of sufferings and trials under Byzantium and the Turks, and many were compelled to leave the country. In the 14th century, Armenia was subjected to a horrible holocaust of mass destruction under Timur, and only a remnant of the nation survived. Flight and exile remained the lot of the Armenians in subsequent centuries, so that the communities in dispersion in

Syria, Egypt, and Iran grew steadily. Even the catholicate was transferred to Isfahan, Persia, between 1614 and 1638.

In the 16th and 17th centuries, the catholicos of Echmiadzin (now Armenian S.S.R.) showed inclinations to union with the Roman Catholic Church. These attempts, which had been given an earlier expression in the 11th century, finally resulted in the celebration of a uniate service (the local rite with approval from Rome) in 1740 in Turkish territory. A uniate patriarchate (under Roman Catholic jurisdiction) was established in Sis (now in Beirut), and in developing communities in Cilicia, Syria, Palestine, Mesopotamia, and Egypt. Today the uniate church has about 100,000 believers.

Modern developments. In the late 18th century, when Russia conquered Armenia, new sufferings occurred because of the policy to coerce the Armenians into the Russian state church. The massacres in Turkey in 1895–96, 1909, and 1914–15 by Turks and Kurds exceeded all other persecutions. In the first massacre, about 200,000 Armenians were killed; in the second, in the district of Adana, about 30,000; and in the third about 600,000, together with those who were deported. Those who escaped (about 150,000) and were not dragged into harems (about 100,000 wives and girls) were compelled to leave the country.

Present organization of the Armenian church

There are thought to be about 2,500,000–3,500,000 Armenians today scattered all over the world (in the Middle East, Europe, North and South America) under two catholicates (Echmiadzin and Sis) and two patriarchates (Jerusalem and Constantinople). The catholicate of Echmiadzin was vacant from 1938 until 1945, when a catholicos, approved by the Soviet government for propaganda purposes (according to many ecclesiastical authorities), filled the position. The catholicos of Sis has his residence in Anṭilyās, Lebanon (since 1921). There are two classes of clergy. The ordinary priests, who must marry before ordination into the diaconate, are recruited from monks or lay people, mainly from the cantors and teachers, and thus the influence of the lay people on the affairs of the church is great. Another class of priests is represented by the *vartabeds*, or doctors, who must remain unmarried. From this class of elite are produced the bishops. These celibates are generally members of a clerical community (*ucht*) who live together in a religious establishment (*vanq*). Monasteries exist only as seminaries for preparing the clergy.

THE COPTIC CHURCH

Origins and early development of Coptic Christianity. Remnants of many apocryphal gospels and papyri discovered in Egypt testify to an early form of Egyptian Christianity (first part of the 2nd century) that must have developed quite rapidly.

During the Council of Chalcedon (451), Dioscurus, patriarch of Alexandria (444–451), defended the Monophysite position and was deposed and banished. This act aroused the Egyptians, who were mindful of their national and ecclesiastical autonomy. The attempts to impose the Chalcedonian decrees on the Egyptian Church fanned the anti-Byzantine sentiments among the native population, and thus the thesis of Cyril of Alexandria ("one nature of the incarnate Word") became a slogan for their fight against Byzantine rule. Proterius, the new Orthodox patriarch, was recruited from the Egyptian clergy; but because he had been consecrated in Constantinople, he could find few adherents (branded as Melkites, "emperor's men").

The great majority remained faithful to Dioscurus and looked upon him as a martyr. Bishops loyal to Dioscurus elected a rival patriarch, Timotheus Aelurus. The riots that followed his consecration precipitated various atrocities, including the mob murder of Proterius on Good Friday, 457. As champions of the interests of the indigenous people, the monks took over the leadership of the masses. Christian monasticism was not generated in Alexandria, a centre of Greek culture, but in the indigenous Coptic (Arab Ḳibṭ, denoting the descendants of the original inhabitants of Egypt) Christian settlements of Upper Egypt. It developed rapidly into a significant movement, able to produce such figures as Shenute (*c.* 348–451/466), a reformer of monastic life and a celebrated leader of the people. The role of the monks is attested by the many Coptic "Lives of the Saints" and the large number of ruins of their monasteries. Under their leadership, almost the entire indigenous population embraced the Monophysite belief. The definitive acts in separating the Egyptian Church from the jurisdiction of Byzantine-oriented ecclesiastics were carried out by Patriarch Timotheos Aelurus (died 477) and Petrus III Mongus (477/482–490). The doctrine that the Egyptian Church adopted represented a modified position, rejecting the concept of the fusion of the divine and human nature of Christ similar to the mixing of water and wine (Eutyches) but understanding the union as similar to the union between the soul and the body. Thereby the absolute form of Monophysitism of Eutyches was rejected in favour of the pre-Chalcedonian Christology of Cyril. Theological writings, however, left room for rapprochement in one or another direction.

Influence of monasticism and doctrinal position of Coptic Christianity

Developments after the Arab conquest. Embittered by the attempts to force them to accept the Chalcedonian decrees, the native Egyptians tried to free themselves from the Byzantine yoke. When the Arabs invaded Egypt in 642, they were greeted at first as liberators. The Umayyad period (661–750) gradually brought restrictions, but what became intolerable was the taxation. In order to simplify control, marks were burned on the hands of the Christians, including the clergy and monks. The heavy burdens caused many native Christians to become apostates. The sufferings increased under the ʿAbbāsids (750–868). In 837 Christians were ordered to wear five-pound crosses around their necks, all new churches were demolished, and Christian education, as well as solemn celebration of festivals, was forbidden. These oppressions seldom led, however, to bodily harm. The Synaxaria (collections of extracts from biographies of saints) show that the number of martyrs under the Byzantines was much greater than that under the Arabs.

Repeated insurrections, however, caused bloody reprisals that in turn caused losses through conversions to Islām. But the church drew courage from the spiritual heritage that had found its expression in monasticism, and in this way preserved its national character, traditions, and idiosyncrasies. Because of the disappearance of the Greek elements during these centuries, however, independent theological thought in the Coptic Church became increasingly barren. A comparison with Syrian thought makes this manifest. There is nothing in Coptic Christianity comparable to the enrichment of life through the extensive translations from the Greek theological works into Syriac and their numerous original literary accomplishments. As a result, the Coptic Church became a borrower of Syrian theological works.

Preservation of the national character of the Coptic Church

With the Shīʿite Fāṭimids (969–1171), a flourishing period for the Egyptian Church occurred—interrupted, however, by the persecutions of al-Ḥākim (996–1021), whose name is connected with the destruction of thousands of churches and chapels and forced conversions to Islām. Many monasteries were also destroyed under the Ayyūbids (1171–1250), and in 1321, under the Mamlūks (1257–1517), most of the remaining churches and monasteries were destroyed. The Copts took vengeance by burning down the mosques, which in turn resulted in mass executions of the Copts. The Turkish period (16th to 19th centuries) brought new persecutions and massacres. All this intermittent bloodletting decimated the number of the Christians. As against 168 bishops in the 6th century, ruling over four ecclesiastical provinces (Pentapolis, Libya, Aegyptus Inferior, and Thebais), in the 10th century the number of bishops had fallen to 110, and dropped to 47 within a century; and in the 17th century, there were only 17 left. Only a very small number (4,000) of the Coptic Christians accepted the jurisdiction of the Roman pope in 1741 under the leadership of the bishop of Jerusalem.

Modern developments. More peaceful times did not come until the reign of the tolerant Muḥammad ʿAlī (1805–48) and the occupation of Egypt by the French and

the British. Some reform movements were initiated, particularly by Patriarch Cyril IV (1854–61), who was celebrated as a promoter of education and literature and a restorer of the churches.

Coptic liturgy and other religious practices

The liturgy, namely, that of Basil, Gregory Nazianzus, and St. Mark, has remained the same, even though the centuries have brought about numerous changes in the life of the church. The Coptic language lives only in the liturgical books. In the manuscripts and printed materials, a Coptic column is furnished with a parallel column in Arabic. Confession, regarding which a reform movement in the 12th century had fought a bitter battle, today is hardly required. Besides the anointing of the sick, there is also an anointing of the healthy as a kind of strengthening of repentance. Monasticism has declined drastically. Today there are eight monasteries with 300 monks and five convents with 80 to 100 nuns.

According to various estimates, today there are about 1,300,000–3,000,000 Coptic Christians, with 19 archbishoprics in Egypt, headed by the patriarch. Outside Egypt there are archbishops in Jerusalem (one), in The Sudan (two), in Uganda (two), in South Africa (one), and in Kuwait (one). Of the Coptic Christians who now accept the jurisdiction of the Roman pontiff, there are about 80,000.

THE ETHIOPIAN CHURCH

Origins and early developments. The earliest phases in the Christianization of Ethiopia are unknown. The first definite historical figure was Frumentius, bishop of Axum (*c.* 356). ʿĒzānā, the king of Aksum, was converted to the Christian faith, and this event was reflected in numismatic and epigraphic sources. The same sources indicate, however, that a large number of the people apparently resisted the new religion so that the king had to follow a policy of opportunism. As a consequence, the endeavours of Frumentius apparently slowed down almost to the point of stagnation. The renewal of Christian missionary activity began with the arrival of the Syrian monks (6th century) whose names have been immortalized in the "Chronicle of the Kings" and the Synaxarium. This revival also introduced Monophysitism. When the patriarchate of Alexandria was transferred to Cairo (*c.* 640), the Ethiopian Church became entirely dependent on it.

The "Dark Ages" of Ethiopian Christianity

Developments after the Arab conquest. From the Arab invasion until the 13th century, the history of Ethiopian Christianity is almost entirely unknown. Conditions probably were deplorable for Christianity since pagan traditions, polygamy, and low standards plagued rulers and clergy alike, as well as the people. Some rulers of the 12th century, however, are celebrated as saints; *i.e.*, they were regarded as champions of Christian faith still wrestling with grave dangers. One of them, Lālibelā (1190–1225), is regarded as the founder of the large rock churches. The common type of church building was the round church, the basic form of which was derived from the huts that were the dwellings of the people. Ties with the Coptic Church were intimate, and in the 12th century the head of the Ethiopian Church, Abuna, was chosen by the Coptic patriarch from the Coptic hierarchy. This connection was particularly valuable for the Copts since after the Arab conquest of the Nile Valley and after the gradual disappearance of Christianity in Nubia (14th century), the Christians in Ethiopia could serve as leverage for the Egyptian Christians. Destruction of churches in the Nile Valley encouraged repercussions in the form of destruction of mosques in Ethiopia.

With the emergence of the Solomonid dynasty (1270), the centre of the state was moved to the south, and an intensified struggle with Islām was initiated. A period of ecclesiastical efflorescence associated with the work of the vigorous and able patriarch Takla Hāymanōt was inaugurated, especially in the areas of Coptic–Arabic literature and monasticism. The position of the head of the monks became strengthened so that it even exceeded that of the Coptic Abuna. In the 14th century, two monastic orders devoted to missionary work were created, and monasticism increased in importance. During the reign of Zarʿa-Yaʿcōb (1434–68), the Ethiopian Church reached its zenith. He was celebrated as a religious reformer and promoter of schools and monasteries.

Contacts with the Roman Catholic Church

Muslim invasions (1520–51) encouraged new attempts aimed at communion with the Roman Catholic Church in order to obtain military help from the Portuguese. Though the 13th-century attempts had ended with the murder of Dominican missionaries, the 16th-century endeavours brought success. Jesuit missionaries arrived in Ethiopia, Negus Susenyos (Malak Sagad III) accepted the Catholic (Chalcedonian) creed in 1621 or 1622, a Catholic patriarch, Atonso Mendes, was sent from Rome in 1624 or 1625, and in 1626 the Catholic religion was elevated to official status. All this, however, proved to be ephemeral. In 1632 Susenyos was succeeded by his son Fasiladas, who reversed his father's policies. The Jesuits were expelled, and attempts made by the Franciscans ended tragically—most of them were martyred. Since the early 19th century Roman Catholic Lazarist and Capuchin missionaries have been more successful, and the number of the uniates (those who accept the jurisdiction of Rome) is now about 76,000.

Modern developments. Changes in the church administration came in the 20th century. The emperor Haile Selassie reorganized the church in 1929, and the number of bishops was increased to 12. The quest for independence from the Coptic Church was achieved in 1948 when the Ethiopian archbishopric received recognition of autocephalic (independent) status from the Coptic patriarch. Though the Coptic patriarch retains a primacy of honour, the patriarch of Ethiopia has primacy of jurisdiction. After 1955 the Ethiopian Church became the state church. It has about 8,000,000 members, including 500,000 in Eritrea, and therefore is the largest among the Eastern independent churches.

Customs and practices. Throughout the centuries, the Ethiopian Church has retained its own traditions. Its canon of the Bible is the largest of Christian biblical canons, containing such apocryphal books as the Jubilees, the Ethiopic Enoch, IV Esdras, the Rest of the Words of Baruch, the Ascension of Isaiah, the Shepherd of Hermas, and others. The Apostles' Creed is not used. These differences, however, are overshadowed by idiosyncrasies that come from Jewish influences, which probably arrived in the kingdom of Aksum from South Arabia and from archaic Semitic elements. Jewish customs and traditions include the Zion (holy hill) of the Ethiopians in Aksum, the seat of the Ark of the Covenant—depicted in the national saga *Kebra Nagast* ("Glory of the Kings"), a replica of which is an essential cult object in every church—the keeping of the sabbath, circumcision, distinction between ritually clean and unclean meats, purification laws, and the architectural design of churches that reflects the threefold division of the Hebrew temple. Pagan customs also survive, such as redemption ceremonies, including the sacrifice of a goat or a lamb for sick persons, and magical customs and prayers that retain their thinly veiled pagan substratum.

The clergy of the Ethiopian Church

There are two kinds of clergy in the Ethiopian Church. The priests who administer the sacraments receive little preparation, and their standards for ordination are low. Little is expected from them by the masses, among whom polygamy, free love, and even infanticide are not uncommon. They are expected to be able to manage the ancient Geʿez language (not understood by the people who speak Amharic) enough to read the liturgy. The *dabtaras* are educated lay clerks entrusted with the chant in the churches and teaching in the schools. Until about 1900, church schools were the only places for education.

Monasticism in the Ethiopian Church is numerically the strongest among the Eastern independent churches. Due to a very low level of achievement of the priesthood, the monks exercise the major religious, moral, and intellectual influence on the faithful.

BIBLIOGRAPHY. General works include: R. JANIN, *Les Églises orientales et les rites orientaux*, 2nd rev. ed. (1926); D. ATTWATER, *The Christian Churches of the East*, rev. ed., 2 vol. (1948–61); B. SPULER, *Die morgenländischen Kirchen* (1964); and A.S. ATIYA, *A History of Eastern Christianity* (1968). Special studies for each branch are: J. STEWART, *Nestorian Mis-*

sionary Enterprise (1928); P.Y. SAEKI, *Nestorian Documents and Relics in China* (1937); A. VOOBUS, *History of Asceticism in the Syrian Orient*, 2 vol. (1958–60), *History of the School of Nisibis* (1965), *New Important Manuscript Discoveries* (1971); M. ORMANIAN, *L'Église arménienne* (1910; Eng. trans., *The Church of Armenia*, 2nd rev. ed., 1955); E.R. HARDY, *Christian Egypt* (1952); and J.S. TRIMINGHAM, *The Christian Church and Missions in Ethiopia* (1950).

(A.V.)

Eastern Orthodoxy

Eastern Orthodoxy, characterized by its continuity with the apostolic church, its liturgy, and its territorial churches, is one of the three major branches of Christianity.

This article is divided into the following sections:

I. Nature and significance
- Significance
- The norm of church organization

II. History
- Causes of the East-West schism
- The Schism of 1054 and its aftermath

III. Doctrine
- Councils and confessions
- God and man
- Christ
- The Holy Spirit
- The Holy Trinity
- The transcendence of God
- Modern theological developments

IV. The structure of the church
- The canons
- The episcopate
- Clergy and laity
- Monasticism

V. Worship and sacraments
- The role of the liturgy
- The eucharistic liturgies
- The liturgical cycles
- The sacraments
- Architecture and iconography

VI. The church and the world
- Missions: ancient and modern
- Orthodoxy and other Christians
- Church, state, and society
- The Orthodox Church today

I. Nature and significance

The term Eastern Orthodoxy designates the large body of Christians who follow the faith and practices that were defined by the first seven ecumenical councils. The word orthodox, meaning "right believing," has traditionally been used, in the Greek-speaking Christian world, to designate those communities, or individuals, who preserved the true faith (as defined by those councils), as opposed to those who were declared heretical. The official designation of the church in Eastern Orthodox liturgical or canonical texts is "the Orthodox Catholic Church." Because of the historical links of Eastern Orthodoxy with the Eastern Roman Empire and Byzantium (Constantinople), however, in English usage it is referred to as the "Eastern" or "Greek Orthodox" Church. These terms are sometimes misleading, especially when applied to Russian or Slavic churches and to the Orthodox communities in western Europe and America.

SIGNIFICANCE

The Schism of 1054

The schism between the churches of the East and the West (1054) was the culmination of a gradual process of estrangement that began in the very first centuries of the Christian era and continued through the Middle Ages. Linguistic and cultural differences, as well as political events, contributed to the estrangement in various degrees. From the 4th to the 11th centuries, Constantinople, the centre of Eastern Christianity, was also the capital of the Eastern Roman, or Byzantine Empire, while Rome, after the barbarian invasions, fell under the influence of the Holy Roman Empire of the West, a political rival of Byzantium. In the West theology remained under the prevailing influence of St. Augustine of Hippo (354–430), while in the East doctrinal thought was shaped by the Greek Fathers. Theological differences could have been settled if the two areas of Christendom had not simultaneously developed different concepts of church authority. The growth of Roman primacy, based on the concept of the apostolic origin of the Church of Rome, was incompatible with the Eastern idea that the importance of certain local churches—Rome, Alexandria, Antioch, and later, Constantinople,—could be determined only by their numerical and political significance and not by apostolic foundation. For the East, the highest authority in settling doctrinal disputes was the ecumenical council.

At the time of the Schism of 1054 between Rome and Constantinople, the membership of the Eastern Orthodox Church was spread throughout the Middle East, the Balkans, and Russia, with its centre in Constantinople, also called "New Rome" and a recognized metropolis of Christian civilization. The vicissitudes of history have greatly modified the internal structures of the Orthodox Church, but, even today, the bulk of its members live in the same geographic areas. Missionary expansion toward Asia and emigration toward the West, however, have maintained the importance of Orthodoxy worldwide.

The total membership of the Orthodox Church today is, in varying estimations, between 100,000,000 and 150,000,000.

THE NORM OF CHURCH ORGANIZATION

Autocephalous and autonomous churches

The Orthodox Church is a fellowship of "autocephalous" churches (governed by their own head bishops), with the Ecumenical Patriarch of Constantinople holding titular or honorary primacy. The number of autocephalous churches has varied in history. Today there are 15: the Church of Constantinople (Istanbul, Turkey), the Church of Alexandria (Egypt), the Church of Antioch (with headquarters in Damascus, Syria), the Church of Jerusalem, the Church of Russia, the Church of Georgia, the Church of Serbia, the Church of Romania, the Church of Bulgaria, the Church of Cyprus, the Church of Greece, the Church of Albania, the Church of Poland, the Church of Czechoslovakia, and the Church of America.

There are also "autonomous" churches (retaining a token canonical dependence upon a mother see) in Crete, Finland, and Japan. Among the autocephalous churches, the first nine are headed by "patriarchs," the others by archbishops or metropolitans. These primatial titles are strictly honorary.

The order of precedence in which the autocephalous churches are listed does not include any reflection on their actual influence or numerical importance. The ancient patriarchates of Constantinople, Alexandria, and Antioch, for example, present only shadows of their past glory. There remains, however, a consensus within the Orthodox Church that Constantinople's primacy of honour, recognized by the ancient canons because it was the capital of the ancient empire, should remain as a symbol and tool of church unity and cooperation. The recent pan-Orthodox conferences were thus convoked by the ecumenical patriarch of Constantinople. Several of the autocephalous churches are de facto national churches, by far the largest being the Russian Church; however, it is not the criterion of nationality but rather the territorial principle that is the norm of church organization.

In recent times, and especially since the Russian Revolution, there has been much turmoil and administrative conflict within the Orthodox Church. In western Europe and in the Americas, in particular, overlapping jurisdictions have been set up and political passions have led to the formation of ecclesiastical organizations without clear canonical status. Though it has provoked controversy, the recent establishment (1970) of the new autocephalous Orthodox Church in America by the patriarch of Moscow has as its stated goal the resumption of normal territorial unity in the Western Hemisphere.

II. History

CAUSES OF THE EAST–WEST SCHISM

The underlying issue in the schism between East and West was, and is, whether in New Testament times and in

the following centuries the Church of Rome possessed a God-given primacy, based on the words of Christ to Peter (Matt. 16:18), or whether each Christian Church was fundamentally identical in status with all the others, provided it kept with them the unity of faith and love. Clear polarization of thought on this issue became evident after the transfer of the Roman Empire's capital from Rome to Constantinople (dedicated in 330) by Constantine I. The Trinitarian and Christological controversies that raged from the 4th to the 8th century were resolved by ecumenical councils meeting in the East and convoked by the emperors. The second ecumenical Council (Constantinople, 381) issued the following ruling: "The bishop of Constantinople shall have the prerogative of honor after the bishop of Rome, because Constantinople is New Rome" (canon 3). In 451 the fourth ecumenical Council of Chalcedon expanded the same view, explaining the concept of the primacy of the "Old Rome" in terms of its having been the old imperial capital and reaffirming the patriarchal authority of the "New Rome" (canon 28). In 599 the archbishop of Constantinople, on the basis of the same theory, assumed the title of ecumenical patriarch. The accepted criterion of primacy in the East was the de facto authority assumed by some churches because of historical circumstances. In order to be fully legitimate, however, this authority must be sanctioned by church councils.

The ecumenical patriarch

Political and cultural factors also contributed to the schism. The Persian wars of Emperor Heraclius (610–641) and the subsequent challenge of Islām forced the Byzantines to concentrate their entire attention on problems of survival in the East. Deprived of the Byzantine emperor's protection against the Lombard invasions, Pope Stephen II (reigned 752–757) asked for the support of the Frankish king Pepin (754); and a later successor, Pope Leo III (reigned 795–816), crowned Pepin's son, Charles, as emperor in Rome. The Frankish influence at the papal court greatly contributed to the breaking of the last bridges with the East.

THE SCHISM OF 1054 AND ITS AFTERMATH

The conflict between Patriarch Photius and Pope Nicholas I ended with a reconciliation under Nicholas' successor, John VIII (872–882). The estrangement, however, grew more intense in the 10th and 11th centuries, due to the growing German influence in Rome and, also, the insertion of the word Filioque (and the Son) into the Nicene Creed (probably in 1014). The Schism of 1054 was nothing more than an unsuccessful attempt at reconciliation between Roman legates and Patriarch Michael Cerularius. The attempt ended in mutual anathemas.

The mistreatment of Eastern Christians by the Crusaders deepened the rift even more, so that the numerous attempts at reunion, and especially the councils of Lyons (1274) and Florence (1439), ended in failure.

Almost simultaneously with the rift with Rome, the Byzantine Church greatly expanded its missionary activity. The mission to the Slavs, initiated by St. Cyril and St. Methodius (863), led to the conversion of Bulgaria (864), Kievan Russia (988), and the establishment of the Serbian Church (1220).

The fall of Constantinople under the Turks (1453) did not put an end to the prestige of the patriarchate. The Ottomans entrusted to it the civil administration of all the Christians of the Turkish Empire. The Russian Church, however, developed full independence and viewed itself as the "Third Rome," successor of Byzantium.

Moscow, the Third Rome

With the gradual disintegration of the Turkish Empire in the 19th and the 20th centuries, independent autocephalous churches were established in the Balkans. Meanwhile, the Russian Church developed a system of education, expanded through its missionaries to Asia, Japan, and Alaska, and played a leading role in the Orthodox world.

The Russian Revolution of 1917, the temporary eclipse and the modern re-emergence of the Russian Church, the institution of Communist regimes in the Balkans after World War II, the establishment of relatively strong Orthodox communities in western Europe and the Americas, as well as the involvement of all Orthodox Churches in the ecumenical movement are the main factors that influence the life of the Orthodox Church today (see also EASTERN ORTHODOXY, HISTORY OF).

III. Doctrine

COUNCILS AND CONFESSIONS

All Orthodox credal formulas, liturgical texts, and doctrinal statements affirm the claim that the Orthodox Church has preserved the original apostolic faith, which was also expressed in the common Christian tradition of the first centuries. The Orthodox Church recognizes as ecumenical the seven councils of Nicaea I (325), Constantinople I (381), Ephesus (431), Chalcedon (451), Constantinople II (553), Constantinople III (681), and Nicaea II (787) but considers that the decrees of several other later councils also reflect the same original faith (*e.g.*, the councils of Constantinople that endorsed the theology of St. Gregory Palamas in the 14th century). Finally, it recognizes itself as the bearer of an uninterrupted living tradition of true Christianity that is expressed in its worship, in the lives of the saints, and in the faith of the whole people of God.

The first seven ecumenical councils

In the 17th century, as a counterpart to the various "confessions" of the Reformation, there appeared several "Orthodox confessions," endorsed by local councils but, in fact, associated with individual authors (*e.g.*, Metrophanes Critopoulos, 1625; Peter Mogila, 1638; Dosítheos of Jerusalem, 1672). None of these confessions would be recognized today as having anything but historical importance. When expressing the beliefs of his church, the Orthodox theologian, rather than seeking literal conformity with any of these particular confessions, will rather look for consistency with Scripture and tradition, as it has been expressed in the ancient councils, the early Fathers, and the uninterrupted life of the liturgy. He will not shy away from new formulations if consistency and continuity of tradition are preserved.

What is particularly characteristic of this attitude toward the faith is the absence of any great concern for establishing external *criteria* of truth—a concern that has dominated Western Christian thought since the Middle Ages. Truth appears as a living experience accessible in the communion of the church and of which the Scriptures, the councils, and theology are the normal expressions. Even ecumenical councils, in the Orthodox perspective, need subsequent "reception" by the body of the church in order to be recognized as truly ecumenical. Ultimately, therefore, truth is viewed as its own criterion: there are signs that point to it, but none of these signs is a substitute for a free and personal experience of truth, which is made accessible in the sacramental fellowship of the church.

The criterion of truth

Because of this view of truth, the Orthodox have traditionally been reluctant to involve church authority in defining matters of faith with too much precision and detail. This reluctance is not due to relativism or indifference but rather to the belief that truth needs no definition to be the object of experience and that legitimate definition, when it occurs, should aim mainly at excluding error and not at pretending to reveal the truth itself that is believed to be ever present in the church.

GOD AND MAN

The development of the doctrines concerning the Trinity and the incarnation, as it took place during the first eight centuries of Christian history, was related to the concept of man's *participation* in divine life.

The Greek Fathers of the church always implied that the phrase found in the biblical story of the creation of man (Gen. 1: 26), according to "the image and likeness of God," meant that man is not an autonomous being and that his ultimate nature is defined by his relation to God, his "prototype." In paradise Adam and Eve were called to participate in God's life and to find in him the natural growth of their humanity "from glory to glory." To be

"in God" is, therefore, the *natural* state of man. This doctrine is particularly important in connection with the Fathers' view of human freedom. For theologians such as Gregory of Nyssa (4th century) and Maximus the Confessor (7th century) man is truly free only when he is in communion with God; otherwise he is only a slave to his body or to "the world," over which, originally and by God's command, he was destined to rule.

Human freedom and sin

Thus, the concept of *sin* implies separation from God and the reduction of man to a separate and autonomous existence, in which he is deprived of both his natural glory and his freedom. He becomes an element subject to cosmic determinism, and the image of God is thus blurred within him.

Freedom in God, as enjoyed by Adam, implied the possibility of falling away from God. This is the unfortunate choice made by man, which led Adam to a subhuman and unnatural existence. The most unnatural aspect of his new state was death. In this perspective, "original sin" is understood not so much as a state of guilt inherited from Adam but as an unnatural condition of human life that ends in death. Mortality is what each man now inherits at his birth and this is what leads him to struggle for existence, to self-affirmation at the expense of others, and ultimately to subjection to the laws of animal life. The "prince of this world" (*i.e.*, Satan), who is also the "murderer from the beginning," has dominion over man. From this vicious circle of death and sin, man is understood to be liberated by the death and Resurrection of Christ, which is actualized in Baptism and the sacramental life in the church.

The general framework of this understanding of the God–man relationship is clearly different from the view that became dominant in the Christian West—*i.e.*, the view that conceived of "nature" as distinct from "grace" and that understood original sin as an inherited guilt rather than as a deprivation of freedom. In the East, man is regarded as fully man when he participates in God; in the West, man's nature is believed to be autonomous, sin is viewed as a punishable crime, and grace is understood to grant forgiveness. Hence, in the West, the aim of the Christian is justification, but in the East, it is rather communion with God and deification. In the West, the church is viewed in terms of mediation (for the bestowing of grace) and authority (for guaranteeing security in doctrine); in the East, the church is regarded as a communion in which God and man meet once again and a personal experience of divine life becomes possible.

Deification

CHRIST

The Orthodox Church is formally committed to the Christology (doctrine of Christ) that was defined by the councils of the first eight centuries. Together with the Latin Church of the West, it has rejected Arianism (a belief in the subordination of the Son to the Father) at Nicaea (325), Nestorianism (a belief that stresses the independence of the divine and human natures of Christ) at Ephesus (431), and Monophysitism (a belief that Christ had only one divine nature) at Chalcedon (451). The Eastern and Western churches still formally share the tradition of subsequent Christological developments, even though the famous formula of Chalcedon, "one person in two natures," is given different emphases in the East and West. The stress on Christ's identity with the pre-existent Son of God, the Logos (Word) of the Gospel According to John, characterizes Orthodox Christology. On Byzantine icons, around the face of Jesus, the Greek letters ὁ ὤν—the equivalent of the Jewish Tetragrammaton YHWH, the name of God in the Old Testament—are often depicted. Jesus is thus always seen in his divine identity. Similarly, the liturgy consistently addresses the Virgin Mary as Theotokos (the "one who gave birth to God"), and this term, formally admitted as a criterion of orthodoxy at Ephesus, is actually the only "Mariological" (doctrine of Mary) dogma accepted in the Orthodox Church. It reflects the doctrine of Christ's unique divine Person, and Mary is thus venerated only because she is his mother "according to the flesh."

Logos as the Christ

This emphasis on the personal divine identity of Christ, based on the doctrine of St. Cyril of Alexandria (5th century), does not imply the denial of his humanity. The anthropology (doctrine of man) of the Eastern Fathers does not view man as an autonomous being but rather implies that communion with God makes man fully human. Thus the human nature of Jesus Christ, fully assumed by the divine Word, is indeed the "new Adam" in whom the whole of humanity receives again its original glory. Christ's humanity is fully "ours"; it possessed all the characteristics of the human being—"each nature (of Christ) acts according to its properties," Chalcedon proclaimed, following Pope Leo—without separating itself from the divine Word. Thus, in death itself—for Jesus' death was indeed a fully human death—the Son of God was the "subject" of the Passion. The theopaschite formula ("God suffered in the flesh") became, together with the Theotokos formula, a standard of orthodoxy in the Eastern Church, especially after the second Council of Constantinople (553). It implied that Christ's humanity was indeed real not only in itself but also for God, since it brought him to death on the cross, and that the salvation and redemption of humanity can be accomplished by God alone—hence the necessity for him to condescend to death, which held humanity captive.

Theology of salvation and redemption

This theology of redemption and salvation is best expressed in the Byzantine liturgical hymns of Holy Week and Easter: Christ is the one who "tramples down death by death," and, on the evening of Good Friday, the hymns already exalt his victory. Salvation is conceived not in terms of satisfaction of divine justice, through paying the debt for the sin of Adam—as the medieval West understood it—but in terms of uniting the human and the divine with the divine overcoming human mortality and weakness and, finally, exalting man to divine life.

What Christ accomplished once and for all must be appropriated freely by those who are "in Christ"; their goal is "deification," which does not mean dehumanization but the exaltation of man to the dignity prepared for him at creation. Such feasts as the Transfiguration or the Ascension are extremely popular in the East precisely because they celebrate humanity glorified in Christ—a glorification that anticipates the coming of the Kingdom of God, when God will be "all in all."

Participation in the already deified humanity of Christ is the true goal of Christian life, and it is accomplished through the Holy Spirit.

THE HOLY SPIRIT

The gift of the Holy Spirit at Pentecost "called all men into unity," according to the Byzantine liturgical hymn of the day; into this new unity, which St. Paul called the "body of Christ," each individual Christian enters through Baptism and "chrismation" (the Eastern form of the Western "confirmation") when the priest anoints him saying "the seal of the gift of the Holy Spirit."

This gift, however, requires man's free response. Orthodox saints such as Seraphim of Sarov (died 1833) described the entire content of Christian life as a "collection of the Holy Spirit." The Holy Spirit is thus conceived as the main agent of man's restoration to his original natural state through Communion in Christ's body. This role of the Spirit is reflected, very richly, in a variety of liturgical and sacramental acts. Every act of worship usually starts with a prayer addressed to the Spirit, and all major sacraments begin with an invocation to the Spirit. The eucharistic liturgies of the East attribute the ultimate mystery of Christ's Presence to a descent of the Spirit upon the worshipping congregation and upon the eucharistic bread and wine. The significance of this invocation (in Greek *epiklēsis*) was violently debated between Greek and Latin Christians in the Middle Ages because the Roman canon of the mass lacked any reference to the Spirit and was thus considered as deficient by the Orthodox Greeks.

The Holy Spirit as the agent of restoration

Since the Council of Constantinople (381), which condemned the Pneumatomachians ("fighters against the Spirit"), no one in the Orthodox East has ever denied that

the Spirit is not only a "gift" but also the giver—*i.e.*, that he is the third Person of the holy Trinity. The Greek Fathers saw in Gen. 1:2 a reference to the Spirit's cooperation in the divine act of creation; the Spirit was also viewed as active in the "new creation" that occurred in the womb of the Virgin Mary when she became the mother of Christ (Luke 1:35); and finally, Pentecost was understood to be an anticipation of the "last days" (Acts 2:17) when, at the end of history, a universal communion with God will be achieved. Thus, all the decisive acts of God are accomplished "by the Father in the Son, through the Holy Spirit."

THE HOLY TRINITY

By the 4th century a polarity developed between the Eastern and Western Christians in their respective understandings of the Trinity. In the West God was understood primarily in terms of one essence (the Trinity of Persons being conceived as an irrational truth found in revelation); in the East the tri-personality of God was understood as the *primary* fact of Christian experience. For most of the Greek Fathers, it was not the Trinity that needed theological proof but rather God's essential unity. The Cappadocian Fathers (Gregory of Nyssa, Gregory of Nazianzus, and Basil of Caesarea) were even accused of being tri-theists because of the personalistic emphasis of their conception of God as one essence in three hypostases (the Greek term *hypostasis* was the equivalent of the Latin *substantia* and designated a concrete reality). For Greek theologians, this terminology was intended to designate the concrete New Testamental revelation of the Son and the Spirit, as distinct from the Father.

Modern Orthodox theologians tend to emphasize this personalistic approach to God; they claim that they discover in it the original biblical personalism, unadulterated in its content by later philosophical speculation.

The Filioque controversy

Polarization of the Eastern and the Western concepts of the Trinity is at the root of the *Filioque* dispute. The Latin word *Filioque* ("and from the Son") was added to the Nicene Creed in Spain in the 6th century. By affirming that the Holy Spirit proceeds not only "from the Father" (as the original creed proclaimed) but also "from the Son," the Spanish councils intended to condemn Arianism by reaffirming the Son's divinity. Later, however, the addition became an anti-Greek battle cry, especially after Charlemagne (9th century) made his claim to rule the revived Roman Empire. The addition was finally accepted in Rome under German pressure. It found justification in the framework of Western conceptions of the Trinity; the Father and the Son were viewed as one God in the act of "spiration" of the Spirit.

The Byzantine theologians opposed the addition, first on the ground that the Western Church had no right to change the text of an ecumenical creed unilaterally and, second, because the *Filioque* clause implied the reduction of the divine persons to mere relations ("the Father and the Son are two in relation to each other, but one in relation to the Spirit"). For the Greeks the Father alone is the origin of both the Son and the Spirit. Patriarch Photius (9th century) was the first Orthodox theologian to explicitly spell out the Greek opposition to the *Filioque* concept, but the debate continued throughout the Middle Ages.

THE TRANSCENDENCE OF GOD

Apophatic theology

An important element in the Eastern Christian understanding of God is the notion that God, in his essence, is totally transcendent and unknowable and that, strictly speaking, God can only be designated by negative attributes: it is possible to say what God is not, but it is impossible to say what he is.

A purely negative, or "apophatic" theology—the only one applicable to the *essence* of God in the Orthodox view—does not lead to agnosticism, however, because God reveals himself personally—as Father, Son, and Holy Spirit—and also in his *acts*, or "energies." Thus, true knowledge of God always includes three elements: religious awe; personal encounter; and participation in the acts, or energies, which God freely bestows on creation.

This conception of God is connected with the personalistic understanding of the Trinity. It also led to the official confirmation by the Orthodox Church of the theology of St. Gregory Palamas, the leader of Byzantine hesychasts (monks devoted to divine quietness through prayer), at the councils of 1341 and 1351 in Constantinople. The councils confirmed a real distinction in God, between the unknowable essence and the acts, or "energies," which make possible a real communion with God. The deification of man, realized in Christ once and for all, is thus accomplished by a communion of divine energy with humanity in Christ's glorified manhood.

MODERN THEOLOGICAL DEVELOPMENTS

Until the conquest of Constantinople by the Turks (1453), Byzantium was the unquestioned intellectual centre of the Orthodox Church. Far from being monolithic, Byzantine theological thought was often polarized by a Humanistic trend, favouring the use of Greek philosophy in theological thinking, and the more austere and mystical theology of the monastic circles. The concern for preservation of Greek culture and for the the political salvation of the empire led several prominent Humanists to adopt a position favourable to union with the West. The most creative theologians (*e.g.*, Symeon the New Theologian, died 1033; Gregory Palamas, died 1359; Nicholas Cabasilas, died *c.* 1390), however, were found rather in the monastic party that continued the tradition of patristic spirituality based upon the theology of deification.

The dark age of Orthodox theology

The 16th, 17th, and 18th centuries were the dark age of Orthodox theology. Neither in the Middle East nor in the Balkans nor in Russia was there any opportunity for independent theological creativity. Since no formal theological education was accessible, except in Western Roman Catholic or Protestant schools, the Orthodox tradition was preserved primarily through the liturgy, which retained all its richness and often served as a valid substitute for formal schooling. Most doctrinal statements of this period, issued by councils or by individual theologians, were polemical documents directed against Western missionaries.

After the reforms of Peter the Great (died 1725), a theological school system was organized in Russia. Shaped originally in accordance with Western Latin models and staffed with Jesuit-trained Ukrainian personnel, this system developed, in the 19th century, into a fully independent and powerful tool of theological education. The Russian theological efflorescence of the 19th and 20th centuries produced many scholars, especially in the historical field (*e.g.*, Philaret Drozdov, died 1867; V.O. Klyuchevsky, died 1913; V.V. Bolotov, died 1900; E.E. Golubinsky, died 1912; N.N. Glubokovsky, died 1937). Independently of the official theological schools, a number of laymen with secular training developed theological and philosophical traditions of their own and exercised a great influence on modern Orthodox theology (*e.g.*, A.S. Khomyakov, died 1860; V.S. Solovyev, died 1900; N. Berdyayev, died 1948), and some became priests (P. Florensky, died 1943; S. Bulgakov, died 1944). A large number of the Russian theological intelligentsia (*e.g.*, S. Bulgakov, G. Florovsky) emigrated to western Europe after the Russian Revolution (1917) and played a leading role in the ecumenical movement.

With the independence of the Balkans, theological schools were also created in Greece, Serbia, Bulgaria, and Romania. Modern Greek scholars contributed to the publication of important Byzantine ecclesiastical texts and produced standard theological textbooks.

The Orthodox diaspora

The Orthodox diaspora—the emigration from eastern Europe and the Middle East—in the 20th century has contributed to modern theological development through their establishment of theological centres in western Europe and America.

Orthodox theologians reacted negatively to the new dogmas proclaimed by Pope Pius IX: the Immaculate Con-

ception of Mary (1854) and papal infallibility (1870). In connection with the dogma of the Assumption of Mary, proclaimed by Pope Pius XII (1950), the objections mainly concerned the presentation of such a tradition in the form of a dogma.

In contrast to the recent general trend of Western Christian thought toward social concerns, Orthodox theologians generally emphasize that the Christian faith is primarily a direct experience of the Kingdom of God, sacramentally present in the church. Without denying that Christians have a social responsibility to the world, they consider this responsibility as an outcome of the life in Christ. This traditional position accounts for the remarkable survival of the Orthodox Churches under the most contradictory and unfavourable of social conditions, but, to Western eyes, it often appears as a form of passive fatalism.

IV. The structure of the church

THE CANONS

The permanent criteria of church structure for the Orthodox Church today, outside of the New Testament writings, are found in the canons (regulations and decrees) of the first seven ecumenical councils; the canons of several local or provincial councils, whose authority was recognized by the whole church; the so-called *Apostolic Canons* (actually some regulations of the church in Syria, dating from the 4th century); and the "canons of the Fathers," or selected extracts from prominent church leaders having canonical importance.

The nomocanon

A collection of these texts was made in the Byzantine nomocanon, attributed, in its final form, to the patriarch Photius (9th century). The Byzantine Church, as well as the modern Orthodox Churches, has adapted the general principles of this collection to its particular situation, and the local autocephalous churches govern themselves according to their own particular statutes, although all accept the ancient canons as their common canonical reference.

The canons themselves do not represent a system or a code. They do, however, reflect a consistent view of the church, of its mission, and of its various ministries; they also reflect an evolution of ecclesiastical structure—*i.e.*, the growth of centralization in the framework of the Christian Roman Empire. For the Orthodox Church today, only the original self-understanding of the church has a theologically normative value. Thus, those canons that reflect the nature of the church as the body of Christ have an unchanging validity today; other canons, if they can be recognized as conditioned by the historical situation in which they were issued, are subject to change by conciliar authority; others have simply fallen out of practice. The use and interpretation of the canons is therefore possible only in the light of some understanding of the church's nature. This theological dimension is the ultimate criterion through which it is possible to distinguish what is permanent in the canons from that which represents no more than a historical value.

THE EPISCOPATE

The Orthodox understanding of the church is based on the principle, attested to in the canons and in early Christian tradition, that each local community of Christians, gathered around its bishop and celebrating the Eucharist, is the local realization of the *whole* body of Christ. "Where Christ is, there is the Catholic church," wrote Ignatius of Antioch (*c.* AD 100). Modern Orthodox theology also emphasizes that the office of the bishop is the highest among the sacramental ministries and that there is therefore no divinely established authority *over* that of the bishop in his own community, or diocese. Neither the local churches nor the bishops, however, can or should live in isolation. The wholeness of church life, realized in each local community, is regarded as identical with that of the other local churches in the present and in the past. This identity and continuity is manifested in the act of the ordination of bishops, an act that requires the presence of several other bishops in order to constitute a conciliar act and to witness to the continuity of apostolic succession and tradition.

The bishop is primarily the guardian of the faith and, as such, the centre of the sacramental life of the community. The Orthodox Church maintains the doctrine of apostolic succession—*i.e.*, the idea that the ministry of the bishop must be in direct continuity with that of the Apostles of Jesus. Orthodox tradition—as expressed especially in its medieval opposition to the Roman papacy—distinguishes the office of the "Apostle" from that of the bishop, however, in that the first is viewed as a universal witness to the historic Jesus and his Resurrection, while the latter is understood in terms of the pastoral and sacramental responsibility for a local community, or church. The continuity between the two is, therefore, a continuity in faith rather than in function. This Orthodox concept of the doctrine of apostolic succession has received wider exposure in Western churches recently because of increased encounters and consultations between Orthodox and Anglican churchmen, the Orthodox always emphasizing unity of faith as a prerequisite for recognition, on their part, of the "validity" of Anglican orders.

Apostolic succession

No bishop can be consecrated or exercise his ministry without being in unity with his colleagues—*i.e.*, be a member of an episcopal council, or "synod." After the Council of Nicaea (325), whose canons are still effective in the Orthodox Church, each province of the Roman Empire had its own synod of bishops that acted as a fully independent unit for the consecration of new bishops and also as a high ecclesiastical tribunal. In the contemporary Orthodox Church these functions are fulfilled by the synod of each autocephalous church. In the early church the bishop of the provincial capital acted as chairman of the synod and was generally called "metropolitan." Today, this function is fulfilled by the local primate who is sometimes called "patriarch" (in the autocephalous churches of Constantinople, Alexandria, Antioch, Jerusalem, Russia, Georgia, Serbia, Romania, and Bulgaria), but he may also carry the title of archbishop (Cyprus, Greece) or metropolitan (Poland, Czechoslovakia, America). The titles of archbishop and metropolitan are also widely used as honorific distinctions.

Generally, but not always, the jurisdiction of each autocephalous synod coincides with national borders—the exceptions are numerous in the Middle East (*e.g.*, jurisdiction of Constantinople over the Greek islands, jurisdiction of Antioch over several Arab states, etc.)—and concerns also the national dioceses of the Orthodox diaspora (*e.g.*, western Europe, Australia, America), which frequently remain under the authority of their mother churches. The latter situation led to an uncanonical overlapping of Orthodox jurisdictions, all based on ethnic origins. Several factors, going back to the Middle Ages, have contributed to modern ecclesiastical nationalism in the Orthodox Church. These factors include the use of the vernacular in the liturgy and the subsequent identification of religion with national culture; this identification sometimes helps the survival of the church under adverse political conditions, but it also hampers missionary expansion and the sense of a specifically Christian identity of the faithful.

CLERGY AND LAITY

The emphasis on communion and fellowship, as the basic principle of church life, inhibited the development of clericalism. The early Christian practice of having the laity participate in episcopal elections never disappeared completely in the East. In modern times, it has been restored in several churches. The Moscow Council of 1917–1918 introduced it in Russia, even if the events of the Revolution prevented its full implementation. Bishops are also elected by clergy-laity conventions in America and in other areas of the Orthodox world.

Clerical marriage

The lower orders of the clergy—*i.e.*, priests and deacons—are generally married men. The present canonical legislation allows the ordination of married men to the diaconate and the priesthood, provided that they were married only once and that their wives are neither wid-

ows nor divorcees. These stipulations reflect the general principle of absolute monogamy, which the Eastern Church considered as a Christian norm to which candidates for the priesthood are to comply strictly. Deacons and priests cannot marry after their ordination.

Bishops, however, are selected among the unmarried clergy or widowed priests. The rule defining the requirement for an unmarried episcopate was issued at a time (6th century) when monks represented the elite of the clergy. The contemporary decrease in the number of monks in the Orthodox Church has created a serious problem in some territorial churches, in that new candidates for the episcopacy are difficult to find.

Besides being admitted, at least in some areas, to participation in episcopal elections, Orthodox laymen often occupy positions in church administration and in theological education. In Greece almost all professional theologians are laymen. Laymen also frequently serve as preachers.

MONASTICISM

The tradition of Eastern Christian monasticism goes back to the 3rd and 4th centuries of the Christian Era. From its beginning it was essentially a contemplative movement seeking the experience of God in a life of permanent prayer. This contemplative character has remained its essential feature throughout the centuries. Eastern Christianity never experienced the development of religious orders, pursuing particular missionary or educational goals and organized on a universal scale, as did Western Christianity.

Eremitic and cenobitic monasticism

Concern for prayer, as the central and principal function of monasticism, does not mean that the Eastern Christian monastic movement was of a single uniform character. Eremitic (solitary) monasticism, favouring the personal and individual practice of prayer and asceticism, often competed with "cenobitic" (communal) monastic life, in which prayer was mainly liturgical and corporate. The two forms of monasticism originated in Egypt and coexisted in Byzantium, as well as throughout eastern Europe.

In Byzantium the great monastery of Studius became the model of numerous cenobitic communities, including that of the Caves in Kiev in the 11th century. It is in the framework of the eremitic, or hesychast tradition, however, that the most noted Byzantine mystical theologians (*e.g.*, Symeon the New Theologian, Gregory Palamas, etc.) received their training. One of the major characteristics of the hesychast tradition is the practice of the "Jesus prayer," or constant invocation of the name of Jesus, sometimes in connection with breathing. This practice won wide acceptance in medieval and modern Russia.

Cenobitic traditions of Byzantium also were important in Slavic lands. The colonization of the Russian north was largely accomplished by monks who acted as pioneers of civilization and as missionaries.

In Byzantium, as well as in other areas of the Orthodox world, the monks were often the only upholders of the moral and spiritual integrity of Christianity, and thus they gained the respect of the masses, as well as that of the intellectuals. The famous Russian *startsy* ("elders") of the 19th century became the spiritual leaders of F. Dostoyevsky, N. Gogol, and L. Tolstoy and inspired many religious philosophers in their quest for religious experience.

Today, the most famous, though declining, centre of Orthodox monasticism is Mount Athos (Greece), where over a thousand monks of different national backgrounds form a variety of communities, grouped into a monastic republic.

V. Worship and sacraments

THE ROLE OF THE LITURGY

By its theological richness, spiritual significance, and variety, the worship of the Orthodox Church represents one of the most significant factors in this church's continuity and identity. It helps to account for the survival of Christianity during the many centuries of Muslim rule in the Middle East and the Balkans when the liturgy was the only source of religious knowledge or experience. Since liturgical practice is practically the only religious expression legally authorized in the Soviet Union today, the continuous existence of Orthodox communities there is also centred almost exclusively around the liturgy.

Worship as the primary experience of Eastern Christianity

The concept that the church is most authentically itself when the congregation of the faithful is gathered together in worship is a basic expression of Eastern Christian experience. Without that concept it is impossible to understand the fundamentals of church structure in Orthodoxy, with the bishop functioning in his essential roles of teacher and high priest in the liturgy. Similarly, the personal experience of man's participation in divine life is understood in the framework of the continuous liturgical action of the community.

According to many authorities, one of the reasons that helps to explain why the Eastern liturgy has made a stronger impact on the Christian Church than has its Western counterpart is that it has always been viewed as a total experience, appealing simultaneously to the emotional, intellectual, and aesthetic faculties of man. The liturgy includes a variety of models, or symbols, using formal theological statements as well as bodily perceptions and gestures (*e.g.*, music, incense, prostrations) or the visual arts. All are meant to convey the content of the Christian faith to the educated and the noneducated alike. Participation in the liturgy implies familiarity with its models, and many of them are conditioned by the historical and cultural past of the church. Thus, the use of such an elaborate and ancient liturgy presupposes catechetical preparation. It may require an updating of the liturgical forms themselves. The Orthodox Church recognizes that liturgical forms are changeable and that, since the early church admitted a variety of liturgical traditions, such a variety is also possible today. Thus, Orthodox communities with Western rites now exist in western Europe and in the Americas.

The Orthodox Church, however, has always been conservative in liturgical matters. This conservatism is due, in particular, to the absence of a central ecclesiastical authority that could enforce reforms and to the firm conviction of the church membership as a whole that the liturgy is the main vehicle and experience of true Christian beliefs. Consequently, reform of the liturgy is often considered as equivalent to a reform of the faith itself. However inconvenient this conservatism may be, the Orthodox liturgy has preserved many essential Christian values transmitted directly from the experience of the early church.

Throughout the centuries, the Orthodox liturgy has been richly embellished with cycles of hymns from a wide variety of sources. Byzantium (where the present Orthodox liturgical rite took shape), while keeping many biblical and early Christian elements, used the lavish resources of patristic theology and Greek poetry, as well as some gestures of imperial court ceremonial, in order to convey the realities of God's kingdom. The results are impressive and the liturgy thus has served as a powerful missionary weapon: *The Russian Primary Chronicle* reports that the envoys of the pagan prince Vladimir, visiting Constantinople in 988 and confronted with the liturgy in Hagia Sophia, did not know any more "whether they were in heaven, or on earth."

Normally, the content of the liturgy is directly accessible to the faithful, because the Byzantine tradition is committed to the use of any vernacular language in the liturgy. Translation of both Scriptures and liturgy into various languages was undertaken by the medieval Byzantines, as well as by modern Russian missionaries. Liturgical conservatism, however, leads de facto to the preservation of antiquated languages. The Byzantine Greek used in church services by the modern Greeks and the Old Slavonic still preserved by all the Slavs are at least as distant from the spoken languages as is the language of the King James Version—used in many Protestant Churches—from modern English.

THE EUCHARISTIC LITURGIES

Two eucharistic liturgies are most generally used in Orthodox worship—*i.e.*, the so-called liturgies of St. John Chrysostom and of St. Basil the Great. Both acquired their present shape by the 9th century, but it is generally recognized that the wording of the eucharistic "canon" of the liturgy of St. Basil goes back to the 4th century—*i.e.*, to St. Basil himself. The so-called Liturgy of St. James is used occasionally, especially in Jerusalem. During the period of Lent, a service of Communion, with elements (bread and wine) reserved from those consecrated on the previous Sunday, is celebrated in connection with the evening service of Vespers; it is called the "Liturgy of the Presanctified" and is attributed to St. Gregory the Great.

The liturgies of St. John Chrysostom and of St. Basil differ only in the text of the eucharistic canon: their overall structure, established in the high Middle Ages, is identical.

These eucharistic liturgies begin with an elaborate rite of preparation (*proskomidē*). A priest on a separate "table of oblation" disposes on a paten (plate) the particles of bread that will symbolize the assembly of the saints, both living and dead, around Christ, the "Lamb of God." Then follows the "Liturgy of the Catechumens," which begins with a processional entrance of the priest into the sanctuary with the Gospel (Little Entrance) and which includes the traditional Christian "liturgy of the word"—*i.e.* the reading from the New Testament letters and the Gospels as well as a sermon. This part of the liturgy ends with the expulsion of the "catechumens," who, until they were baptized, were not admitted to the sacramental part of the service. The "Liturgy of the Faithful" includes another ceremonial procession of the priest into the sanctuary. He carries the bread and wine from the table of oblations to the altar (Great Entrance); the following recitation of the Nicene Creed, the eucharistic canon, the Lord's Prayer, and Communion are—as in the West—the characteristic parts of the Byzantine "Liturgy of the Faithful." The bread used for the Eucharist is ordinary leavened bread; Communion is distributed under both kinds (bread and wine) with a special spoon (*labis*).

THE LITURGICAL CYCLES

Variety of hymnodical texts

One of the major characteristics of the Byzantine liturgical tradition is the wealth and variety of hymnodical texts marking the various cycles of the liturgical year. A special liturgical book contains the hymns for each of the main cycles. The *daily* cycle includes the offices of Vespers (Hesperinos), Compline (Apodeipnon), the midnight prayer, Matins (Orthros), and the four canonical "hours"—*i.e.*, offices to be said at the "First" (6:00 AM), "Third" (9:00 AM), "Sixth" (12:00 noon), and "Ninth" (3:00 PM) hours. The liturgical book covering the daily cycle is called the *Hōrologion* ("The Book of Hours"). The *Paschal* (Passover) cycle is centred on the "Feast of Feasts"—*i.e.*, of Christ's Resurrection; it includes the period of Great Fast (Lent), preceded by three Sundays of preparation and the period of 50 days following Easter. The hymns of the Lenten period are found in the *Triōdion* (Three Odes), and those of the Easter season in the *Pentekostarion* (called the "Flowery Triodion"). The weekly cycle is the continuation of the Resurrection cycle found in the *Triōdion* and the *Pentekostarion;* each week following the Sunday after Pentecost (50 days after Easter) possesses its own musical tone, or mode, in accordance with which all the hymns of the week are sung. There are eight tones whose composition is traditionally attributed to St. John of Damascus (8th century). Each week is centred around Sunday, the day of Christ's Resurrection.

The Easter and weekly cycles clearly dominate all offices of the entire year and illustrate the absolute centrality of the Resurrection in the Eastern understanding of the Christian message. The date of Easter, set at the Council of Nicaea (325), is the first Sunday after the full moon following the spring equinox. Differences between the East and West in computing the date exist because the Orthodox Church uses the Julian calendar for establishing the date of the equinox (hence a delay of 13 days) and also because of the tradition that Easter must necessarily follow the Jewish Passover and must never precede it or coincide with it. The *yearly* cycle includes the hymns for each of the 366 days of the calendar year, with its feasts and daily commemoration of saints. They are found in the 12 volumes of the *Menaion* ("Book of Months").

The golden age of Byzantine hymnody

From the 6th to the 9th centuries the Byzantine Church experienced its golden age of creativity in the writing of hymns by outstanding poets such as John of Damascus. In more recent times hymn writing has generally followed the accepted patterns set by these authors but rarely has it reached the quality of its models. Since the Eastern Orthodox tradition bans instrumental music, or accompaniment, the singing is always a cappella, with only a few exceptions admitted by Westernized parishes in America. The idea behind the ban is based upon the practice of worship in the New Testament; *i.e.*, only the natural aptitudes of the living congregation are viewed as capable of expressing praise that is worthy of God. In many Orthodox churches there is a wealth of new musical compositions for liturgical texts.

THE SACRAMENTS

Contemporary Orthodox catechisms and textbooks all affirm that the church recognizes seven *mystēria*, or "sacraments": Baptism, chrismation, Communion, holy orders, penance, anointing of the sick (the "extreme unction" of the medieval West), and marriage. Neither the liturgical book called *Euchologion* (prayer book), which contains the texts of the sacraments, nor the patristic tradition, however, formally limits the *number* of sacraments; they do not distinguish clearly between the "sacraments" and such acts as the blessing of water on Epiphany day or the burial service or the service for the tonsuring of a monk that in the West are called *sacramentalia.* In fact, no council recognized by the Orthodox Church ever defined the number of sacraments; it is only through the "Orthodox confessions" of the 17th century directed against the Reformation that the number seven has been generally accepted. The underlying sacramental theology of the Orthodox Church is based, however, on the notion that the ecclesiastical community is the unique *mystērion*, of which the various sacraments or *sacramentalia* are the normal expressions.

The Christian community as the *mystērion*

In the West, since the Scholastic period (Middle Ages) and, especially, since the Catholic Reformation (16th century), much emphasis has been placed on the vicarious juridical power of the minister to administer the sacraments validly. The Orthodox East, however, interprets each sacramental act as a prayer of the entire ecclesiastical community, led by the bishop or his representative, and also as a response of God, based upon Christ's promise to send the Holy Spirit upon the church. These two aspects of the sacrament exclude both magic and legalism: they imply that the Holy Spirit is given to free men and call for their responses. In the *mystērion* of the church, the participation of men in God is effected through their "cooperation" or "synergy"; to make this participation possible once more is the goal of the incarnation.

Baptism and confirmation. Baptism is normally performed by triple immersion as a sign of the death and Resurrection of Christ; thus, the rite appears essentially as a gift of new life. It is immediately followed by confirmation, performed by the priest who anoints the newly baptized Christian with "Holy Chrism" (oil) blessed by the bishop. Baptized and confirmed children are admitted to Holy Communion. By admitting children immediately after their Baptism to both confirmation and Communion, the Eastern Christian tradition maintains the positive meaning of Baptism—*i.e.*, as the beginning of a new life nourished by the Eucharist.

The Eucharist. There never has been, in the East, much speculation about the nature of the eucharistic mys-

tery. Both canons presently in use (that of St. Basil and that of St. John Chrysostom) include the "words of institution" ("This is my Body . . .," "This is my Blood . . ."), which are traditionally considered in the West as the formula necessary for the validity of the sacrament. In the East, however, the culminating point of the prayer is not in the remembrance of Christ's act but in the invocation of the Holy Spirit, which immediately follows: "Send down Thy Holy Spirit upon us and upon the Gifts here spread forth, and make this bread to be the precious Body of Thy Christ. . . ." Thus, the central mystery of Christianity is seen as being performed by the prayer of the church and through an invocation of the Spirit. The nature of the mystery that occurs in the bread and wine is signified by the term *metabolē* ("sacramental change"). The Western term transubstantiation occurs only in some confessions of faith after the 17th century.

Orders. The Orthodox Church recognized three major orders: the diaconate, the priesthood, and the episcopate (bishop), as well as the minor orders of the lectorate and the subdiaconate. All the ordinations are performed by a bishop and, normally, during the eucharistic liturgy. The consecration of a bishop requires the participation of at least two or three bishops, as well as an election by a canonical synod.

Penance. The sacrament of penance in the early church was a solemn and public act of reconciliation, through which an excommunicated sinner was readmitted into church membership. Historically it has evolved into a private act of confession through which every Christian's membership in the church is periodically renewed. In the Orthodox Church today there is a certain variety in both the practice and the rite of penance. In the churches of the Balkans and the Middle East, it fell into disuse during the four centuries of Turkish occupation but is gradually being restored today. In Greek-speaking churches only certain priests, especially appointed by the bishop, have the right to hear confessions. In Russia, on the contrary, confessions remained a standard practice that was generally required before communion. General or group confession, introduced by John of Kronstadt, a Russian spiritual leader of the early 20th century, is also occasionally practiced. The rite of confession in the *Euchologion* retains the form of a prayer, or invocation, said by the priest for the remission of the penitent's sins. In the Slavic ritual a Latin-inspired and juridical form of personal absolution was introduced by Peter Mogila, metropolitan of Kiev (17th century). Confession, in Orthodox practice, is generally viewed as a form of spiritual healing rather than as a tribunal. The relative lack of legalism reflects the Eastern patristic approach to sin—*i.e.*, as an internal passion and as an enslavement. The external sinful acts—which alone can be legally tried—are only manifestations of man's internal disease.

Practice of confession

Anointing of the sick. Anointing of the sick is a form of healing by prayer. In the Greek Church it is performed annually in church for the benefit of the entire congregation on the evening of Holy Wednesday.

Marriage. Marriage is celebrated through a rite of crowning, performed with great solemnity and signifying an eternal union, sacramentally "projected" into the Kingdom of God. Orthodox theology of marriage insists on its sacramental eternity rather than its legal indissolubility. Thus, second marriages, in cases of either widowhood or divorce, are celebrated through a subdued penitential rite, and men who have been married more than once are not admitted to the priesthood. Remarriage after divorce is tolerated on the basis of the possibility that the sacrament of marriage was not originally received with the consciousness and responsibility that would have made it fully effective; remarriage can be a second chance.

ARCHITECTURE AND ICONOGRAPHY

Since the time of Constantine I, Eastern Christianity has developed a variety of patterns in church architecture. The chief model was created when Emperor Justinian I completed the "great church" of Hagia Sophia in Constantinople in the 6th century. The architectural conception of that church consisted of erecting a huge round dome on top of the classical early Christian basilica. The dome was meant to symbolize the descent of heaven upon earth—*i.e.*, the ultimate meaning of the eucharistic celebration.

By courtesy of the Greek Orthodox Cathedral, London; photograph, A.C. Cooper Ltd.

The iconostasis or screen, characteristically decorated with icon paintings, separates the sanctuary from the nave in an Orthodox church. In the Greek Orthodox Cathedral, London.

The screen, or iconostasis, which separates the sanctuary from the nave in contemporary Orthodox Churches, is a rather late development. After the triumph of orthodoxy over iconoclasm (destruction of images) in 843, a new emphasis was placed upon the permanent revelatory role of *images*. The incarnation implied that God had become man—*i.e.*, fully visible and thus describable in his human nature. The images of Christ and the saints, who had manifested in their lives the new humanity transfigured by the grace of God, were placed everywhere in full evidence before the congregation. A contrast was thus suggested between the visible manifestation of God through the pictural representation of Christ as man and his more perfect but mysterious and invisible presence in the Eucharist. The iconostasis, together with those parts of the liturgy that involve the closing and opening of the curtain before the altar, emphasizes the mysterious and "eschatological" (consummation of history) character of the eucharistic service. They suggest, however, that this *mystery* is not a "secret" and that the Christian is being introduced through the eucharistic liturgy into the very reality of divine life and of the kingdom to come, which was revealed when God became man.

The iconoclastic controversy

The long iconoclastic controversy (725–843), during which the Orthodox theology of icons was fully developed, concerned itself primarily with the problem of the incarnation; it was the direct continuation of the Christological debates of the 5th, 6th, and 7th centuries. The image of Christ, the incarnated God, became, for the Eastern Christian, a pictorial confession of faith: God was truly visible in the humanity of Jesus of Nazareth, and the saints—whose images surround that of Christ—are witnesses of the fact that the transfigured, "deified" humanity is accessible to those who believe in Christ.

Popular use of icons

Departing from tridimensional images or statues, which were reminiscent of pagan idolatry, the Christian East developed a rich tradition of iconography. Portable icons —often painted on wood but also using mosaics with enamel techniques—are always kept in houses or public places. Among the icon painters, who never signed their work, there appeared several artists of genius. Most of

"The Old Testament Trinity," Russian icon painted by Andrey Rublev, c. 1410. In the State Tretyakov Gallery, Moscow.

By courtesy of the State Tretyakov Gallery, Moscow

them are unknown, but tradition and written documents have revealed the names of some, such as the famous 14th–15th-century Russian painter Andrey Rublev.

VI. The church and the world

The schism between the Greek and the Latin churches coincided chronologically with a surge of Christian missionary activity in northern and eastern Europe. Both sides contributed to the resultant expansion of Christianity but used different methods. The West imposed a Latin liturgy on the new converts and thus made Latin the only vehicle of Christian civilization and a major instrument of ecclesiastical unity. The East, meanwhile, as noted above, accepted from the start the principle of translating both the Scriptures and the liturgy into the spoken tongues of the converted nations. Christianity thus became integrated into the indigenous cultures of the Slavic nations, and the universal Orthodox Church evolved as a fellowship of national churches rather than as a centralized body.

MISSIONS: ANCIENT AND MODERN

The Christian East, in spite of the integrating forces of Christian Hellenism, was always culturally pluralistic: since the first centuries of Christianity, Syrians, Armenians, Georgians, Copts, Ethiopians, and other ethnic groups used their own languages in worship and developed their own liturgical traditions. Even though by the time of the Greek missions to the Slavs the Byzantine Church was almost monolithically Greek, the idea of a liturgy in the vernacular was still quite alive, as is demonstrated by the use of the Slavic language by the missionaries of SS. Cyril and Methodius in the 9th century.

The Turkish conquest of the Middle East and of the Balkans (15th century) interrupted the missionary expansion of the Orthodox Church. Throughout the Middle Ages Islām and Christianity had usually confronted each other only militarily, and the victory of Islām meant that the Christians could survive only in enclaves and were legally excluded from proselytizing among Muslims.

Russian missionary activity

The Russian Church alone was able to continue the tradition of SS. Cyril and Methodius, and it did so almost without interruption until today. In the 14th century St. Stephen of Perm translated the Scriptures and the liturgy into the language of a Finnish tribe of the Russian north and became the first bishop of the Zyrians. The expansion of the Russian Empire in Asia was accompanied by efforts of evangelization that—sometimes in opposition to the avowed policy of Russianization practiced by the government of St. Petersburg—followed the Cyrillo-Methodian pattern of translation. This method was utilized among the Tatars of the Volga in the 16th century and among the various peoples of Siberia throughout the 18th and the 19th centuries. In 1714 a mission was established in China. In 1794 monks of the Valamo Abbey reached Alaska; their spiritual leader, the monk Herman, was canonized by the Orthodox Church in 1970.

The Alaskan mission was under the direction of a modest priest sent to America from eastern Siberia, Ivan Veniaminov. During his long stay in America, first as a priest, then as a bishop (1824–1868), he engaged in the work of translating the Gospels and the liturgy into the languages of the Aleuts, the Tlingit Indians, and the Eskimos of Alaska.

In Japan an indigenous Orthodox Church was established by the recently canonized St. Nikolay Kasatkin (died 1913). The entirely indigenous character of this church enabled it to survive the political trials of the Russo-Japanese War (1904–1905), of the Russian Revolution, and of World War II. The new Church of Japan received its full autonomy from the Russian Church in 1970.

The missionary tradition is also being revived in Greece. Various Greek associations are dedicated to the pursuit of missionary work in East Africa, where sizable indigenous groups have recently joined the Orthodox Church.

ORTHODOXY AND OTHER CHRISTIANS

Since the failure of the unionist Council of Florence (1439) there have been no official attempts to restore unity between the Orthodox Church and Roman Catholicism. In 1484 an Orthodox council defined that Roman Catholics desiring to join the Orthodox Church were to be received through chrismation (or confirmation). In the 17th century, however, the relations deteriorated to the point that the ecumenical patriarchate of Constantinople decreed that all Roman Catholic and Protestant sacraments, including Baptism, were totally unauthentic. A parallel attitude prevailed in Russia until the 18th century, when large numbers of Eastern Rite Roman Catholics ("Uniates") were received back into Orthodoxy by a simple confession of faith, and this practice was adopted in the acceptance of individual Roman Catholics as well.

After the 16th-century Reformation, a lengthy correspondence took place between the Tübingen group of Reformers (German Lutheran) headed by P. Melanchthon and ecumenical Patriarch Jeremiah II. It led to no concrete results, for the East generally considered the Protestants as only a branch of deviation of the altogether erroneous Roman Church.

Various attempts at rapprochement with the Anglican Communion, especially since the 19th century, were generally more fruitful. Several private associations of churchmen and theologians promoted understanding between Eastern Orthodoxy and the "Anglo-Catholic" branch of Anglicanism. The Orthodox, however, were reticent in taking any formal step toward reunion before a satisfactory statement on the content of Anglican faith, taken as a whole, could be obtained.

Ecumenical movement

The contemporary ecumenical movement involved the Orthodox Church from the very beginning. Eastern Orthodox representatives took part in the various Life and Work (practical) and Faith and Order (theological) Conferences from the very beginning of this century. One by one the various independent Orthodox Churches joined the World Council of Churches, created in 1948. Often, and especially at the beginning of their participation, Orthodox delegates had recourse to separate statements, which made clear to the Protestant majorities that, in the Orthodox view, Christian unity was attainable only in the full unity of the primitive apostolic faith from which the Orthodox Church had never departed. This attitude of the Orthodox could be understood only if it made sufficiently clear that the truth—which historic Eastern

Orthodoxy claims to preserve—is maintained by the Holy Spirit in the church as a whole and not by any individual or any group of individuals on their own right and also that the unity of Christians—which is the goal of the ecumenical movement—does not imply cultural, intellectual, or ritual uniformity but rather a mystical fellowship in the fullness of truth as expressed in eucharistic Communion.

The ecumenical movement, especially since the second Vatican Council, is today much wider than the formal membership of the World Council of Churches. The principle of conciliarism and the readiness of the popes to appear publicly as equals of Eastern patriarchs—as in the recent meetings between Pope Paul VI and Patriarch Athenagoras—represent significant moves in the direction of a better understanding between Orthodoxy and Roman Catholicism.

The tendency, however, represented by those Western Christians who apparently identify Christianity with various political or social causes has the effect of again widening the traditional gap that has been, in the past, one of the major causes of the break between East and West.

CHURCH, STATE, AND SOCIETY

In the West, after the fall of the Roman Empire, the church assumed the unifying social function that no other individual or institution was able to fulfill. Eventually, the popes were formally invested with civil authority in Christendom. In the East the empire persisted until 1453 and in Russia until 1917; thus the church had to fulfill its social functions in the political framework of the Christian empire.

This historical contrast coincides with a theological polarization: the Eastern Fathers conceived the God–man relationship in terms of personal experience and Communion culminating in deification; Western theology, meanwhile, understood man as autonomous in the secular sphere, although controlled by the authority of the church, which was conceived as vicariously representing God.

Caesaropapism

The Byzantine and Eastern form of church–state relations has often been labelled as "caesaropapism," because the hierarchy of the church was, most of the time, deprived of the legal possibility of opposing imperial power. But this label is inaccurate in two aspects: first, it presupposes that the emperor possessed a recognizable power to define the content of the faith, comparable to that of the papacy; and, second, it underestimates the power of the church (as a corporate, transfiguring, and deifying power) that is effective without legal guarantees or statutes. The Byzantine ideal of church–state relations was a "symphony" between the civil and the ecclesiastical functions of Christian society. The abuses of imperial power were frequent, but innumerable examples of popular resistance to those imperial decrees that were considered as detrimental to the faith can be cited. Neither the strong emperors of the 7th century, trying to impose Monophysitism, nor the weakened Palaeologans (13th to 15th centuries), attempting reunion with Rome, were able to overcome the corporate opposition of Orthodox clergy and laity.

The Byzantine conception of church–state relations was not, however, without major weaknesses. It often led to a de facto identification of the interests of the church with those of the empire. Conceived when both the church and the empire were supranational and, in principle, universal, it gradually evolved into a system that gave a sacred sanction to *national* states. Modern ecclesiastical nationalism, which inhibits relations between Orthodox Churches today, is the outcome of the medieval alliance between the empire and the church.

Only after the Turkish occupation of the Balkans was civil authority directly assumed by the Orthodox Church hierarchy in the Middle East. It was granted to it by the new Muslim overlords, who chose to administer their Christian subjects as a separate community, or *millet*, ruled by its own religious leaders. The patriarch of Constantinople was thus appointed by the sultan as head (*millet-bachi*) of the entire Christian population of the Ottoman Empire. Understood by some, especially the Greeks, as the heir of Byzantine emperors and by others, especially the Balkan Slavs and Romanians, as an agent of the hated Turks, the patriarch exercised these powers until the secularization of the Turkish republic by Kemal Atatürk in 1921. By that time, however, he had lost most of his jurisdictional powers because of the establishment of autocephalous churches in Greece, Serbia, Bulgaria, and Romania.

The *millet* system of the Ottoman Empire

The *millet* system, however, survived in other areas of the Middle East. In Cyprus, for example, the church assumed a leading role in national liberation, and its prestige made Archbishop Makarios the natural leader of the young republic.

The *millet* system and the active political responsibilities that it implied for the church, it should be noted, originated in the Ottoman period only and is not in the spiritual tradition of the Christian East as such. The Russian Church is the most recent example of religious survival without practical social or political involvement.

The Orthodox attitude toward social responsibility in the world constitutes a distinct contribution to the contemporary ecumenical movement. But it will be meaningful only if it is understood in its proper framework—*i.e.*, as an understanding of the Christian faith as a personal spiritual experience of God, which is self-sufficient knowledge of God and which, as such, can lead to an authentically Christian witness in the secularized world.

The practical forms of that witness have varied greatly in history, and Orthodox tradition has placed among the church's saints both hermits and politicians, hesychast monks as well as emperors. According to Serge Bulgakov, a modern Orthodox theologian, the Orthodox Church accepts "a relativism of means and methods," provided there remains "an absolute and unique goal," which is the Kingdom of God still *to come* but also already *present* in the mystery of the church.

Codification and systematization of practical devices in the fields of personal or social ethics is foreign to Orthodoxy, which rather relies on free human conscience; each Christian, in his behaviour, stands before the judgment of the New Testament and of the great examples of the saints.

THE ORTHODOX CHURCH TODAY

The contemporary Orthodox Church is confronted with major problems of a structural, intellectual, and spiritual nature. On the world level its internal organization, while very decentralized, is still shaped by traditions going back to the Byzantine Empire. The ancient patriarchates of Constantinople, Alexandria, Antioch, and Jerusalem—which occupy the first four places among the autocephalous churches in virtue of their past as major centres of the empire—represent altogether no more than 1,000,000 faithful; and if the patriarch of Antioch still governs a sizable community of Arab Christians in Syria and Lebanon, his three colleagues are dependent upon the human and material support of Greece.

The violent persecutions suffered by the Russian Church after the Communist Revolution and the serious limitations placed upon religious activities by the Socialist states of the Balkans constitute another major test of Orthodoxy's ability to survive in the modern world. The Church of Russia, greatly reduced in numbers and deprived of means of educational, missionary, and intellectual development, has passed the test successfully thus far: the fact that major literary and intellectual figures like Boris Pasternak and Aleksandr Solzhenitsyn have been professed Christians and that millions of the faithful crowd the few churches which are still open, does give much hope to the Orthodox for the future.

Mere survival, however, is insufficient for a meaningful intellectual development, which requires a system of education and a tradition of scholarship. History has not favoured Eastern Orthodoxy in this respect. The Turkish conquest in 1453 brutally interrupted the potential en-

counter between the tradition of the Greek Fathers and the great Western Renaissance, for which Byzantine theologians seemed to be ready. Similarly, the Communist Revolution of 1917 put an end to the remarkable scholarly and spiritual creativity of Orthodox Russia, whose major Christian intellectual figures were either persecuted or emigrated to the West.

In spite of these events, the areas of potential growth and the signs of revival are not lacking. These can be seen not only in the remarkable capacity for survival under adverse conditions, which Orthodox Christianity has recently shown, but also in the efforts to strengthen the unity between the autocephalous churches—five pan-Orthodox conferences met recently in Rhodes (1961, 1963, 1964), Belgrade, and Geneva—and in the Orthodox expansion to Western countries, especially to the Americas. It remains to be seen what positive contribution this new presence will make in the midst of the contemporary crisis of Western Christianity.

BIBLIOGRAPHY. S.N. BULGAKOV, *L'Orthodoxie* (1932; Eng. trans., *The Orthodox Church*, 1935), the best general introduction to Orthodox doctrine and spirituality; J. MEYENDORFF, *L'Église Orthodoxe hier et aujourd'hui* (1960; Eng. trans., *The Orthodox Church*, 2nd rev. ed., 1969), a history of Eastern Christianity, the schism, and the present situation of Orthodoxy, and *Le Christ dans la théologie byzantine* (1969; Eng. trans., *Christ in Eastern Christian Thought*, 1969), occasionally technical; T. WARE, *The Orthodox Church* (1963), an easily accessible and very complete general introduction; V. LOSSKY, *Essai sur la théologie mystique de l'Église d'Orient* (1944; Eng. trans., *The Mystical Theology of the Eastern Church*, 1957), a classic on God-man relations; A. SCHMEMANN, *Sacraments and Orthodoxy* (1965), the basic approach to liturgy and sacraments; L. OUSPENSKY and V. LOSSKY, *Der Sinn der Ikonen* (1952; Eng. trans., *The Meaning of Icons*, 1956), a fervent introduction to a fascinating aspect of Byzantine and Russian Christianity; P.N. TREMBELAS, *Dogmatique de l'Église orthodoxe catholique*, 3 vol. (1966–68; orig. pub. in Greek, 1959), an all-encompassing textbook for theological students, the only one available in a Western language.

(J.M.)

Eastern Orthodoxy, History of

Eastern Orthodoxy, tracing its origins to the very beginning of Christianity, is one of the three major forces in Christianity, along with Roman Catholicism and Protestantism. Since the Schism of 1054, Eastern Orthodoxy —centred mainly in the Balkans, the Middle East, and Russia—has experienced unfavourable effects from various ethnic, political, social, and antireligious pressures. In spite of such pressures, which the Western churches have not experienced, Eastern Orthodoxy has not only survived but, at times, it also has thrived.

Throughout the medieval period and up to modern times, the Christian Church in the Middle East, in the Balkans, and in eastern Europe continued a historical development distinct from both the Western Latin world and from the ancient Monophysite (emphasizing the divine nature of Christ) or Nestorian (emphasizing the human nature of Christ) churches of the East that did not accept the decrees of the Council of Chalcedon in 451. Holding the faith of the seven ecumenical councils (from 325–787), it always defined itself as the "Orthodox" Church in distinction to Western Roman Catholicism and Eastern non-Chalcedonian churches. After the 16th century, Russian Orthodox missions developed in northern and eastern Asia and finally on the American continent. In the 20th century, numerous Orthodox communities appeared throughout the Western world and, more recently, in Africa.

The Schism in 1054 that separated East and West, the tragic confrontation with Western Crusaders (*e.g.*, the sack of Constantinople in 1204), the Mongolian conquest of Russia (13th century), and the Turkish occupation of the entire Middle East (15th–20th centuries) contributed to the estrangement between East and West, which was initiated in the first centuries of Christianity. With the emergence of Russia in the 16th century and the independence of the Balkan peoples in the 19th century, East and West met again, but their respective histories and religious beliefs were different. Eastern Orthodoxy, which had rejected the medieval papacy of Rome, had not known the Reformation. It had encountered modern secularism only recently. Its structures and organization had developed in the tight framework of the Byzantine, the Ottoman, and the Russian imperial systems. In matters of faith, however, it had preserved the idea that the entire body of the church (the "people of God"), as distinct from any particular institution such as the papacy, was the guardian of the truth. Also, its insistence on tradition, on the liturgy, on sacramental life, and on the mystical experience of God clearly distinguished its religious outlook from the major principles of Protestantism (*e.g.*, justification by grace through faith).

THE CHURCH OF IMPERIAL BYZANTIUM

Byzantine Christianity around AD 1000. At the beginning of the second millennium of Christian history, the church of Constantinople, capital of the East Roman (or Byzantine) Empire, was at the peak of its world influence and power. Neither Rome, which had become a provincial town and its church an instrument in the hands of political interests, nor Europe under the Carolingian and Ottonian dynasties could really compete with Byzantium as centres of Christian civilization. The Byzantine emperors of the Macedonian dynasty had extended the frontiers of the empire from Mesopotamia to Naples (in Italy) and from the Danube River (in central Europe) to Palestine. The church of Constantinople not only enjoyed a parallel expansion but also extended its missionary penetration, much beyond the political frontiers of the empire, to Russia and the Caucasus.

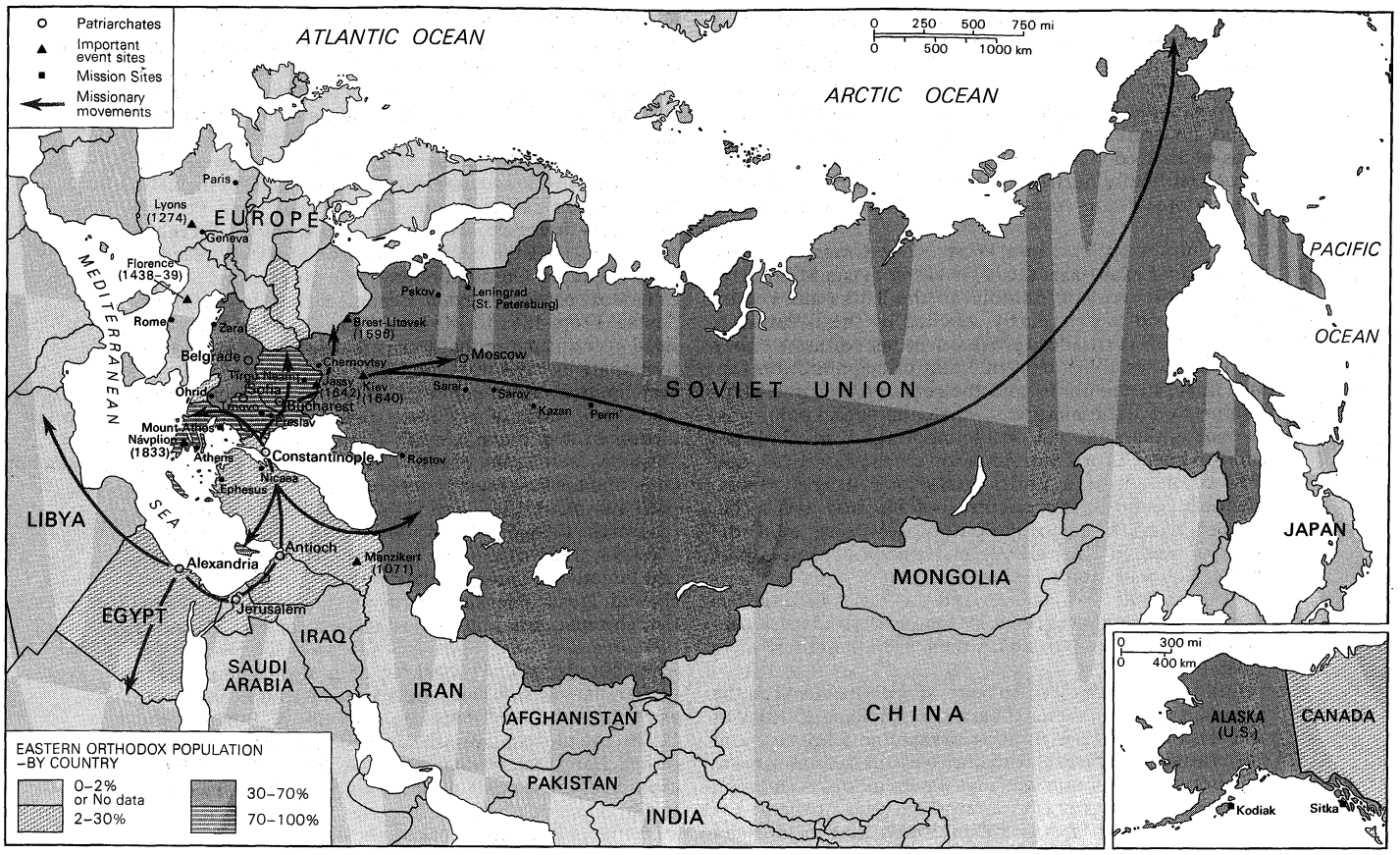
Expansion of Byzantine church and state

Relations between church and state. The ideology that had prevailed since Constantine (4th century) and Justinian I (6th century)—according to which there was to be only one universal Christian society, the *oikoumenē*, led jointly by the empire and the church—was still the ideology of the Byzantine emperors. The authority of the patriarch of Constantinople was motivated in a formal fashion by the fact that he was the bishop of the "new Rome," where the emperor and the senate also resided (canon 28 of the Council of Chalcedon, 451). He held the title of "ecumenical patriarch," which pointed to his political role in the empire. Technically, he occupied the second rank—after the bishop of Rome—in a hierarchy of five major primates, which included also the patriarchs of Alexandria, Antioch, and Jerusalem. In practice, however, the latter three were deprived of all authority by the Arab conquest of the Middle East in the 7th century, and only the emerging Slavic churches attempted to challenge, at times, the position of Constantinople as the unique centre of Eastern Christendom.

The relations between state and church in Byzantium are often described by the term caesaropapism, which implies that the emperor was acting as the head of the church. The official texts, however, describe the emperor and the patriarch as a dyarchy (government with dual authority) and compare their functions to that of the soul and the body in a single organism. In practice, the emperor had the upper hand over much of church administration, though strong patriarchs could occasionally play a decisive role in politics: Patriarch Nicholas Mystikus (patriarch 901–907, 912–925) and Polyeuctus (patriarch 956–970) excommunicated emperors for uncanonical acts. In the area of faith and doctrine, the emperors could never impose their will when it contradicted the conscience of the church: this fact, shown in particular during the numerous attempts at union with Rome during the late medieval period, proves that the notion of caesaropapism is not unreservedly applicable to Byzantium.

Significance of Hagia Sophia

The Church of the Divine Wisdom, or Hagia Sophia, built by Justinian in the 6th century, was the centre of religious life in the Eastern Orthodox world. It was by far the largest and most splendid religious edifice in all of Christendom. According to *The Russian Primary Chronicle*, the envoys of the Kievan prince Vladimir, who visited it in 987, reported: "We knew not whether we were in heaven or on earth, for surely there is no such



The extent of Eastern Orthodoxy in the Balkans, the Middle East, the Soviet Union, and Alaska.
Orthodox Christian population data from *Eastern Orthodox World Directory* by Joe Kuzmission, Copyright 1968 by Branden Press, Inc.

splendor or beauty anywhere upon earth." Hagia Sophia, or the "great church," as it was also called, provided the pattern of the liturgical office, which was adopted throughout the Orthodox world. This adoption was generally spontaneous and based upon the moral and cultural prestige of the imperial capital: the Orthodox Church uses the 9th-century Byzantine Rite.

Monastic and mission movements. Both in the capital and in other centres, the monastic movement continued to flourish as it was shaped during the early centuries of Christianity. The Constantinopolitan monastery of Studius was a community of over 1,000 monks, dedicated to liturgical prayer, obedience, and asceticism. They frequently opposed both government and ecclesiastical officialdom, defending fundamental Christian principles against political compromises. The Studite Rule (guidelines of monastic life) was adopted by daughter monasteries, particularly the famous Monastery of the Caves in Kiev (in Russia). In 963 Emperor Nicephorus II Phocas offered his protection to St. Athanasius, the Athonite, whose laura (large monastery) is still the centre of the monastic republic of Mt. Athos (under the protection of Greece). The writings of St. Symeon the New Theologian (949–1022), abbot of the monastery of St. Mamas in Constantinople, are a most remarkable example of Eastern Christian mysticism, and they exercised a decisive influence on later developments of Orthodox spirituality.

Missionary expansion

Historically, the most significant event was the missionary expansion of Byzantine Christianity throughout eastern Europe. In the 9th century, Bulgaria had become an Orthodox nation and under Tsar Symeon (893–927) had established its own autocephalous (administratively independent) patriarchate in Preslav. Under Tsar Samuel (976–1014) another autocephalous Bulgarian centre appeared in Ohrid. Thus, a Slavic-speaking daughter church of Byzantium dominated the Balkan Peninsula. It lost its political and ecclesiastical independence after the conquests of Byzantine emperor Basil II (976–1025), but the seed of a Slavic Orthodoxy had been solidly planted. In 988 the Kievan prince Vladimir embraced Byzantine Orthodoxy and married a sister of the same emperor Basil. After that time, Russia became an ecclesiastical province of the church of Byzantium, headed by a Greek or, less frequently, a Russian metropolitan appointed from Constantinople. This statute of dependence was not challenged by the Russians until 1448. During the entire period, Russia adopted and developed the spiritual, artistic, and social heritage of Byzantine civilization, which was received through intermediary Bulgarian translators.

Relations with the West. Relations with the Latin West, meanwhile, were becoming more ambiguous. On the one hand, the Byzantines considered the entire Western world as a part of the Roman *oikoumenē* of which the Byzantine emperor was the head and in which the Roman bishop enjoyed honorary primacy. On the other hand, the Frankish and German emperors in Europe were challenging this nominal scheme, and the internal decadence of the Roman papacy was such that the powerful patriarch of Byzantium seldom took the trouble of entertaining any relations with it. From the time of Patriarch Photius (patriarch 858–867, 877–886), the Byzantines had formally condemned the *Filioque* clause, which stated that the Spirit proceeded from the Father and from the Son, as an illegitimate and heretical addition to the Nicene Creed, but in 879–880 Photius and Pope John VIII had apparently settled the matter to Photius' satisfaction. In 1014, however, the *Filoque* was introduced in Rome, and communion was broken again.

The background of the Schism of 1054

The incident of 1054, wrongly considered as the date of the Schism (which had actually been developing over a period of time), was, in fact, an unsuccessful attempt at restoring relations, disintegrating as they were because of political competition in Italy between the Byzantines and the Germans and also because of disciplinary changes (enforced celibacy of the clergy, in particular) imposed by the reform movement that had been initiated by the monks of Cluny (in France). Conciliatory efforts of Emperor Constantine Monomachus (reigned 1042–1055)

were powerless to overcome either the aggressive and uninformed attitudes of the Frankish clergy, who were now governing the Roman Church, or the intransigence of Byzantine patriarch Michael Cerularius (1043–58). When papal legates came to Constantinople in 1054, they found no common language with the patriarch. Both sides exchanged recriminations on points of doctrine and ritual and finally hurled anathemas of excommunication at each other, thus provoking what has been called the Schism.

Invasions from East and West. *The Crusades.* After the Battle of Manzikert (1071) in eastern Asia Minor, Byzantium lost most of Anatolia to the Turks and ceased to be a world power. Partly solicited by the Byzantines, the Western Crusades proved another disaster: they brought the establishment of Latin principalities on former imperial territories and the replacement of Eastern bishops by a Latin hierarchy. The culminating point was, of course, the sack of Constantinople itself in 1204, the enthronement of a Latin emperor on the Bosporus, and the installation of a Latin patriarch in Hagia Sophia. Meanwhile, the Balkan countries of Bulgaria and Serbia secured national emancipation with Western help, the Mongols sacked Kiev (1240), and Russia became a part of the Mongol Empire of Genghis Khan.

The Byzantine heritage survived this series of tragedies mainly because the Orthodox Church showed an astonishing internal strength and a remarkable administrative flexibility.

Until the Crusades, and in spite of such incidents as the exchanges of anathemas between Michael Cerularius and the papal legates in 1054, Byzantine Christians did not consider the break with the West as a final schism. The prevailing opinion was that the break of communion with the West was due to a temporary take-over of the venerable Roman see by misinformed and uneducated German "barbarians," and that eventually the former unity of the Christian world under the one legitimate emperor—that of Constantinople—and the five patriarchates would be restored. This utopian scheme came to an end when the Crusaders replaced the Greek patriarchs of Antioch and Jerusalem with Latin prelates, after they had captured these ancient cities (1098–99). Instead of re-establishing Christian unity in the common struggle against Islām, the Crusades demonstrated how far apart Latins and Greeks really were from each other. When finally, in 1204, after a shameless sacking of the city, the Venetian Thomas Morosini was installed as patriarch of Constantinople and confirmed as such by Pope Innocent III, the Greeks realized the full seriousness of papal claims over the universal church: theological polemics and national hatreds were combined to tear the two churches further apart.

The sack of Constantinople and its aftermath

After the capture of the city, the Orthodox patriarch John Camaterus fled to Bulgaria and died there in 1206. A successor, Michael Autorianus was elected in Nicaea (1208), where he enjoyed the support of a restored Greek empire. Although he lived in exile, this patriarch was recognized as legitimate by the entire Orthodox world. He continued to administer the immense Russian metropolitanate. From him, and not from his Latin competitor, the Bulgarian Church received again its right for ecclesiastical independence with a restored patriarchate in Trnovo (1235). It was also with the Byzantine government at Nicaea that the Orthodox Serbs negotiated the establishment of their own national church; their spiritual leader, St. Sava, was installed as autocephalous (independent) archbishop of Serbia in 1219.

The Mongol invasion. The invasion of Russia by the Mongols had disastrous effects on the future of Russian civilization, but the church survived, both as the only unified social organization and as the main bearer of the Byzantine heritage. The "metropolitan of Kiev and all Russia," who was appointed from Nicaea or from Constantinople, was a major political power, respected by the Mongol Khans. Exempt from taxes paid by the local princes to the Mongols and reporting only to his superior (the ecumenical patriarch), the head of the Russian Church—though he had to abandon his cathedral see of Kiev that had been devastated by the Mongols—acquired an unprecedented moral prestige. He retained ecclesiastical control over immense territories from the Carpathian Mountains to the Volga River, over the newly created episcopal see of Sarai (near the Caspian Sea), which was the capital of the Mongols, as well as over the Western principalities of the former Kievan Empire—even after they succeeded in winning independence (*e.g.*, Galicia) or fell under the political control of Lithuania and Poland.

Attempts at ecclesiastical union and theological renaissance. In 1261 the Nicaean emperor Michael Palaeologus recaptured Constantinople from the Latins, and an Orthodox patriarch again occupied the see in Hagia Sophia. From 1261 to 1453 the Palaeologan dynasty presided over an empire that was embattled from every side, torn apart by civil wars, and gradually shrinking to the very limits of the imperial city itself. The church, meanwhile, kept much of its former prestige, exercising jurisdiction over a much greater territory, which included Russia as well as the distant Caucasus, parts of the Balkans, and the vast regions occupied by the Turks. Several patriarchs of this late period—*e.g.*, Arsenius Autorianus (patriarch 1255–59, 1261–65), Athanasius I (patriarch 1289–93, 1303–10), John Calecas (patriarch 1334–47), and Philotheus Coccinus (patriarch 1353–54, 1364–76)—showed great independence from the imperial power, though remaining faithful to the ideal of the Byzantine *oikoumenē*.

Extension of ecclesiastical jurisdiction

Without the military backing of a strong empire, the patriarchate of Constantinople was, of course, unable to assert its jurisdiction over the churches of Bulgaria and Serbia, which had gained independence during the days of the Latin occupation. In 1346 the Serbian Church even proclaimed itself a patriarchate; a short-lived protest by Constantinople ended with recognition in 1375. In Russia, Byzantine ecclesiastical diplomacy was involved in a violent civil strife; a fierce competition arose between the grand princes of Moscow and Lithuania, who both aspired to become leaders of a Russian state liberated from the Mongol yoke. The "metropolitan of Kiev and all Russia" was by now residing in Moscow, and often, as in the case of the metropolitan Alexis (1354–78), played a directing role in the Muscovite government. The ecclesiastical support of Moscow by the church was decisive in the final victory of the Muscovites and had a pronounced impact on later Russian history. The dissatisfied western Russian principalities (which would later constitute the Ukraine) could only obtain—with the strong support of their Polish and Lithuanian overlords—the temporary appointment of separate metropolitans in Galicia and Belorussia. Eventually, late in the 14th century, the metropolitan residing in Moscow again centralized ecclesiastical power in Russia.

Relations with the Western Church. One of the major reasons behind this power struggle in the northern area of the Byzantine world was the problem of relations with the Western Church. To most Byzantine churchmen, the young Muscovite principality appeared to be a safer bulwark of Orthodoxy than the Western-oriented princes who had submitted to Catholic Poland and Lithuania. Also, an important political party in Byzantium itself favoured union with the West in the hope that a new Western Crusade might be made against the menacing Turks. The problem of ecclesiastical union was, in fact, the most burning issue during the entire Palaeologan period.

The problem of ecclesiastical union

Emperor Michael Palaeologus (1259–82) had to face the aggressive ambition of the Sicilian Norman king Charles of Anjou, who dreamed of restoring the Latin empire in Constantinople. To gain the valuable support of the papacy against Charles, Michael sent a Latin-inspired confession of faith to Pope Gregory X, and his delegates accepted union with Rome at the Council of Lyons (1274). This capitulation before the West, sponsored by the Emperor, won little support in the church. During his lifetime, Michael succeeded in imposing an Eastern Catholic patriarch, John Beccus, upon the Church of Constantinople, but upon Michael's death an Orthodox council condemned the union (1285).

Throughout the 14th century, numerous other attempts at negotiating union were initiated by the emperors of Byzantium. Formal meetings were held in 1333, 1339, 1347, and 1355. In 1369 Emperor John V Palaeologus was personally converted to the Roman faith in Rome. All these attempts were initiated by the government and not by the church, for an obvious political reason; *i.e.*, the hope for Western help against the Turks. But the attempts brought no results either on the ecclesiastical or on the political levels. The majority of Byzantine Orthodox churchmen were not opposed to the idea of union but considered that it could only be brought about through a formal ecumenical council at which East and West would meet on equal footing, as they had done in the early centuries of the church. The project of a council was promoted with particular consistency by John Cantacuzenus, who, after a brief reign as emperor (1347–54), became a monk but continued to exercise great influence on all ecclesiastical and political events. The idea of an ecumenical council was initially rejected by the popes, but it was revived in the 15th century with the temporary triumph of conciliarist ideas (which advocated more power to councils and less to popes) in the West at the councils of Constance and Basel. Challenged with the possibility that the Greeks would unite with the conciliarists and not with Rome, Pope Eugenius IV called an ecumenical council of union in Ferrara, which later moved to Florence.

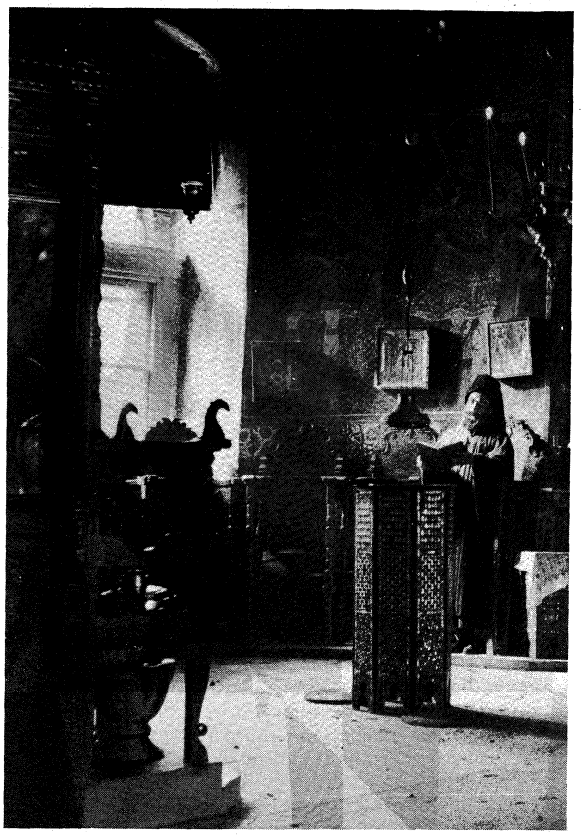
The Council of Ferrara–Florence

The Council of Ferrara–Florence (1438–45) lasted for months and allowed for long theological debates. Emperor John VIII Palaeologus, Patriarch Joseph, and numerous bishops and theologians represented the Eastern Church. They finally accepted most Roman positions—the *Filioque* clause, purgatory (an intermediate stage for the soul's purification between death and heaven), and the Roman primacy. Political desperation and the fear of facing the Turks again, without Western support, was the decisive factor that caused them to place their signatures of approval on the Decree of Union (July 6, 1439). The metropolitan of Ephesus, Mark Eugenicus, alone refused to sign. Upon their return to Constantinople, most other delegates also renounced their acceptance of the council and no significant change occurred in the relations between the churches.

The official proclamation of the union in Hagia Sophia was postponed until December 12, 1452; however, on May 29, 1453, Constantinople fell to the Ottoman Turks. Sultan Mehmed II transformed Hagia Sophia into an Islāmic mosque, and the few partisans of the union fled to Italy.

Theological and monastic renaissance. Paradoxically, the pitiful history of Byzantium under the Palaeologan emperors coincided with an astonishing intellectual, spiritual, and artistic renaissance that influenced the entire Eastern Christian world. The renaissance was not without fierce controversy and polarization. In 1337 Barlaam the Calabrian, one of the representatives of Byzantine Humanism, attacked the spiritual practices of the Hesychast (from the Greek word *hēsychia*, meaning quiet) monks, who claimed that Christian asceticism and spirituality could lead to the vision of the "uncreated light" of God. Barlaam's position was upheld by several other theologians, including Akyndinus and Nicephorus Gregoras. After much debate, the church gave its support to the main spokesman of the monks, Gregory Palamas, who showed himself as one of the foremost theologians of medieval Byzantium (1296–1359). The councils of 1341, 1347, and 1351 adopted the theology of Palamas and, after 1347, the patriarchal throne was consistently occupied by his disciples. John Cantacuzenus, who, as emperor, presided over the council of 1351, gave his full support to the Hesychasts. His close friend, Nicholas Cabasilas, in his spiritual writings on the divine liturgy and the sacraments, defined the universal Christian significance of Palamite theology. The influence of the religious zealots, who triumphed in Constantinople, outlasted the empire itself and contributed to the perpetuation of Orthodox spirituality under the Turkish rule. It also spread to the Slavic countries, especially Bulgaria and Russia. The monastic revival in northern Russia during the last half of the 14th century, associated with the name of St. Sergius of Radonezh, as well as the contemporaneous revival of iconography (*e.g.*, the work of the famous painter Andrey Rublev), would be unthinkable without constant contacts with Mt. Athos, the centre of Hesychasm and with the spiritual and intellectual life of Byzantium.

Pierre Vals—*Paris Match*



Orthodox monk meditating in a chapel on Mount Athos.

Along with the Hesychast revival, a significant "opening to the West" was taking place among some Byzantine ecclesiastics. The brothers Prochorus and Demetrius Cydones, under the sponsorship of Cantacuzenus, for example, were systematically translating the works of Latin theologians into Greek. Thus, major writings of Augustine, Anselm of Canterbury, and Thomas Aquinas were made accessible to the East for the first time. Most of the Latin-minded Greek theologians eventually supported the union policy of the emperors, but there were some—like Gennadius Scholarius, the first patriarch under the Turkish occupation—who reconciled their love for Western thought with total faithfulness to the Orthodox Church.

ORTHODOXY UNDER THE OTTOMANS (1453–1821)

The Christian ghetto. According to Muslim belief, Christians, as well as Jews, were considered as "people of the Book"; *i.e.*, their religion was seen as not entirely false, but incomplete. Accordingly, provided that Christians submitted to the dominion of the caliphate and the Muslim political administration and paid appropriate taxes, they deserved consideration and freedom of worship. Any Christian mission or proselytism among the Muslims, however, was considered a capital crime. In fact, Christians were formally reduced to a ghetto existence: they were the *Rūm millet*, or the "Roman nation" conquered by Islām but enjoying a certain internal autonomy.

Conditions of Christians under Muslim rule

In January 1454 the Sultan allowed the election of a new patriarch, who was to become *millet-bachi*, the head of the entire Christian *millet*, or in Greek the "ethnarch," with the right to administer, to tax, and to exercise justice

over all the Christians of the Turkish empire. Thus, under the new system, the patriarch of Constantinople saw his formal rights and jurisdiction extended both geographically and substantially: on the one hand, through the privileges granted to him by the sultan, he could practically ignore his colleagues, the other Orthodox patriarchs, and, on the other hand, his power ceased to be purely canonical and spiritual but became political as well. To the enslaved Greeks, he appeared not only as the successor of the Byzantine patriarchs but also as the heir of the emperors. For the Ottomans, he was the official and strictly controlled administrator of the *Rūm millet*. In order to symbolize these new powers, the patriarch adopted an external attire reminiscent of that of the emperors: mitre in form of a crown, long hair, eagles as insignia of authority, and other imperial accoutrements.

Consequences of the *millet* system

The new system had many significant consequences. Most important, it permitted the church to survive as an institution; indeed, the prestige of the church was actually increased because, for Christians, the church was now the only source of education and it alone offered possibilities of social promotion. Moreover, through the legal restrictions placed on mission, the new arrangement created the practical identification of church membership with ethnic origin. And finally, since the entire Christian *millet* was ruled by the patriarch of Constantinople and his Greek staff, it guaranteed to the Phanariots, the Greek aristocracy of the Phanar (now called Fener, the area of Istanbul where the patriarchate was, and still is, located), a monopoly in episcopal elections. Thus, Greek bishops progressively came to occupy all the hierarchical positions. The ancient patriarchates of the Middle East were practically governed by the Phanar. The Serbian and Bulgarian churches came to the same fate: the last remnants of their autonomy were formally suppressed in 1766 and 1767, respectively, by the Phanariot patriarch Samuel Hantcherli. This Greek control, exercised through the support of the hated Turks, was resented more and more by the Balkan Slavs and Romanians as the Turkish regime became more despotic, taxes grew heavier, and modern nationalisms began to develop.

It is necessary, however, to credit the Phanariots with a quite genuine devotion to the cause of learning and education, which they alone were able to provide inside the oppressed Christian ghetto. The advantages they obtained from the Porte (the Turkish government) for building schools and for developing Greek letters in the Romanian principalities of Moldavia and Walachia that were entrusted to their rule came to play a substantial role in the rebirth of Greece.

Relations with the West. The Union of Florence became fully inoperative as soon as the Turks occupied Constantinople (1453). In 1484 a council of bishops condemned it officially. Neither the sultan nor the majority of the Orthodox Greeks were favourable to the continuation of political ties with Western Christendom. The Byzantine cultural revival of the Palaeologan period was the first to experience adverse effects from the occupation. Intellectual dialogue with the West became impossible. Through liturgical worship and the traditional spirituality of the monasteries, the Orthodox faith was preserved in the former Byzantine world. Some self-educated men developed a remarkable ability to develop the Orthodox tradition through writings and publications, but they were isolated exceptions. Probably the most remarkable among them was St. Nicodemus of the Holy Mountain, the Hagiorite (1748–1809), who published the famous *Philocalia*, an anthology of spiritual writings, and also translated and adapted Western spiritual writings (*e.g.*, those of the Jesuit founder, Ignatius of Loyola) into modern Greek.

East–West contacts through education and ecumenical correspondence

The only way for Orthodox Greeks, Slavs, or Romanians to acquire an education higher than the elementary level was to go to the West. Several of them were able to do so, but, in the process, became detached from their own theological and spiritual tradition.

The West, in spite of much ignorance and prejudice, had a constant interest in the Eastern Church. At times there was a genuine and respectful curiosity; in other instances, political and proselytistic (conversion) concerns prevailed. Thus, in 1573–81, a lengthy correspondence was initiated by the Lutheran scholars from Tübingen (in Germany), including Philipp Melanchthon. However interesting as a historical event, this correspondence, which includes the *Answers* of Patriarch Jeremias II (patriarch 1572–95), shows how little mutual understanding was possible at that time between the Reformers and traditional Eastern Christianity.

Relations with the West, especially after the 17th century, were often vitiated in the East by the incredible corruption of the Turkish government, which constantly fostered diplomatic intrigues. An outstanding example of such manipulation was the *kharāj*, an important tax required by the Porte at each patriarchal election. Western diplomats were often ready to provide the amount needed in order to secure the election of candidates favourable to their causes. The French and Austrian ambassadors, for example, supported candidates who would favour the establishment of Roman Catholic influence in the Christian ghetto, while the British and Dutch envoys supported patriarchs who were open to Protestant ideas. Thus, a gifted and Western-educated patriarch, Cyril Lucaris, was elected and deposed five times between 1620 and 1638. His stormy reign was marked by the publication in Geneva of a *Confession of Faith* (1629), which was, to the great amazement of all contemporaries, purely Calvinistic (*i.e.*, it contained Reformed Protestant views). The episode ended in tragedy. Cyril was strangled by Turkish soldiers at the instigation of the pro-French and pro-Austrian party. Six successive Orthodox councils condemned the *Confession:* Constantinople, 1638; Kiev, 1640; Jassy, 1642; Constantinople, 1672; Jerusalem, 1672; and Constantinople, 1691. In order to refute its positions, the metropolitan of Kiev, Peter Mogila, published his own *Orthodox Confession of Faith* (1640), which was followed, in 1672, by the *Confession* of the patriarch of Jerusalem, Dosítheos Notaras. Both, especially Peter Mogila, were under strong Latin influence.

These episodes were followed, in the 18th century, by a strong anti-Western reaction. In 1755 the Synod of Constantinople decreed that all Westerners—Latin or Protestant—had invalid sacraments and were only to be admitted into the Orthodox Church through Baptism. This practice of the Greek Church fell into disuse only in the 20th century.

THE CHURCH OF RUSSIA (1448–1800)

The "third Rome." *Origin of the Muscovite patriarchate.* At the Council of Florence, the Greek "metropolitan of Kiev and all Russia." Isidore, was one of the major architects of the Union. Having signed the decree, he returned to Moscow in 1441 as a Roman cardinal but was rejected by both church and state, arrested, and then allowed to escape to Lithuania. In 1448, after much hesitation, the Russians received a new primate, Jonas, elected by their own bishops. Their church became "autocephalous"; *i.e.*, administratively independent under a "metropolitan of all Russia," residing in Moscow. In territories controlled by Poland, Rome (in 1458) appointed another "metropolitan of Kiev and all Russia." The tendencies toward separation from Moscow that had existed in the Ukraine since the Mongol invasion and that were supported by the kings of Poland thus received official sanction. In 1470, however, this metropolitan broke the union with the Latins and re-entered—nominally—the jurisdiction of Constantinople, by then under Turkish control.

Independence of the Moscow metropolitanate

After this, the fate of the two churches "of all Russia" became quite distinct. The metropolitanate of Kiev developed under the control of Roman Catholic Poland. Hard pressed by the Polish kings, the majority of its bishops, against the will of the majority of their flock, eventually accepted union with Rome at Brest-Litovsk (1596). In 1620, however, an Orthodox hierarchy was reestablished, and a Romanian nobleman, Peter Mogila, was elected metropolitan of Kiev (1632). He created the first Orthodox theological school of the modern period.

the famous Academy of Kiev. Modelled after the Latin seminaries of Poland, with instruction given in Latin, this school served as the theological training centre for almost the entire Russian high clergy in the 17th and 18th centuries. In 1686 the Ukraine was finally reunited with Muscovy, and the metropolitanate of Kiev was attached to the patriarchate of Moscow, with approval given by Constantinople.

Muscovite Russia, meanwhile, had acquired the consciousness of being the last bulwark of true Orthodoxy. In 1472 Grand Prince Ivan III (reigned 1462–1505) married Sophia, the niece of the last Byzantine emperor. The Muscovite sovereign began to use more and more of the Byzantine imperial ceremonial, and he assumed the double-headed eagle as his state emblem. In 1510 the monk Philotheus of Pskov addressed Basil III as "tsar," (or emperor), saying: "Two Romes have fallen, but the third stands, and a fourth there will not be." The meaning of the sentence was that the first Rome was heretical, the second—Byzantium—was under Turkish control, and the third was Moscow. Ivan IV, the Terrible, was crowned emperor, according to the Byzantine ceremonial, by the metropolitan of Moscow, Makary, on January 16, 1547. In 1551 he solemnly presided in Moscow over a great council of Russian bishops, the Stoglav ("Council of 100 Chapters"), in which various issues of discipline and liturgy were settled and numerous Russian saints were canonized. These obvious efforts to live up to the title of the "third Rome" lacked one final sanction: the head of the Russian Church was lacking the title of "patriarch." The "tsars" of Bulgaria and Serbia did not hesitate in the past to bestow the title on their own primates, but the Russians wanted an unquestionable authentication and waited for proper opportunity. It occurred in 1589 when the patriarch of Constantinople, Jeremias II, was on a fund-raising tour of Russia. He could not resist the pressure of his hosts and established the metropolitan Job as "patriarch of Moscow and all Russia." Confirmed later by the other Eastern patriarchs, the new patriarchate obtained the fifth place in the honorific order of the Oriental sees, after the patriarchs of Constantinople, Alexandria, Antioch, and Jerusalem.

The "third Rome" ideology

Relations between patriarch and tsar. After the 16th century, the Russian tsars always considered themselves as successors of the Byzantine emperors and the political protectors and financial supporters of Orthodoxy throughout the Balkans and the Middle East. The patriarch of Moscow, however, never pretended to occupy formally the first place among the patriarchs. Within the Muscovite Empire, many traditions of medieval Byzantium were faithfully kept. A flourishing monastic movement spread the practice of Christian asceticism in the northern forests, which were both colonized and Christianized by the monks. St. Sergius of Radonezh (*c.* 1314–92) was the spiritual father of this monastic revival. His contemporary, St. Stephen of Perm, missionary to the Zyryan tribes, continued the tradition of SS. Cyril and Methodius, the "apostles to the Slavs" in the 9th century, in translating the Scripture and the liturgy into the vernacular. He was followed by numerous other missionaries who promoted Orthodox Christianity throughout Asia and even established themselves on Kodiak Island off the coast of Alaska (1794). The development of church architecture, iconography, and literature also added to the prestige of the "third Rome."

Russian monastic and missionary activities

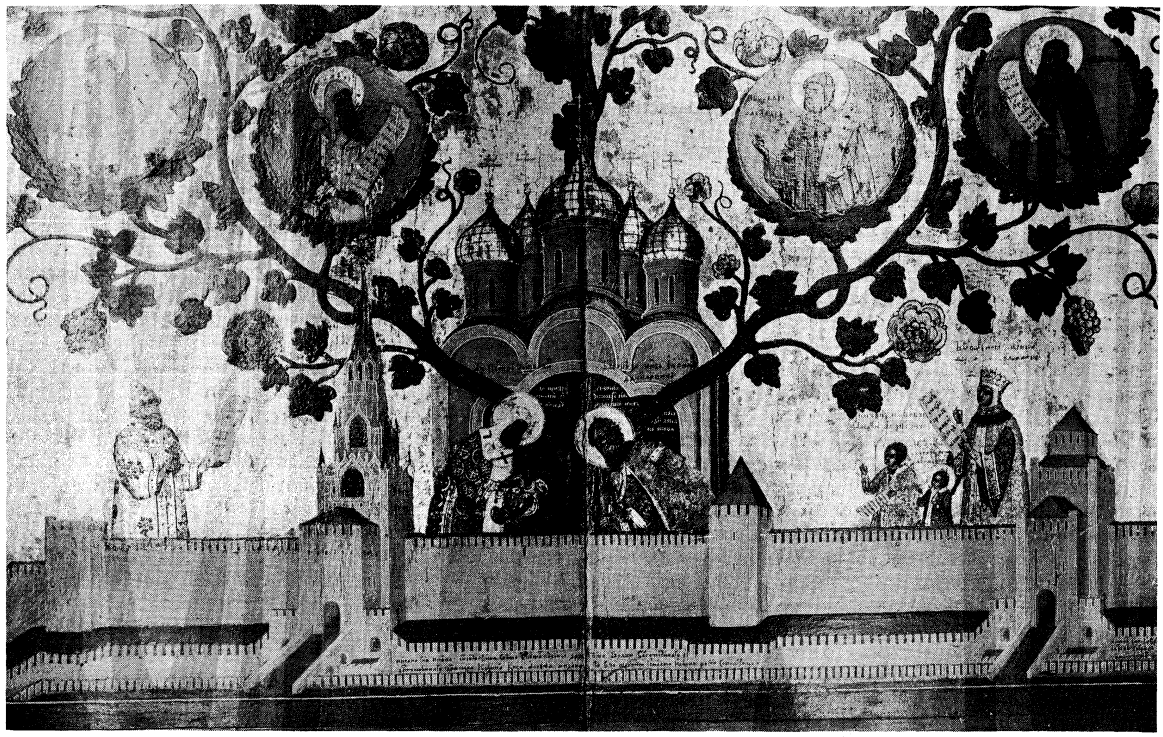
The Muscovite Empire, however, was quite different from Byzantium both in its political system and in its cultural self-understanding. The Byzantine "symphony" (harmonious relationship) between the emperor and the patriarch was never really applied in Russia. The secular goals of the Muscovite state and the will of the monarch always superseded canonical or religious considerations, which were still binding on the medieval emperors of Byzantium. Muscovite political ideology was always influenced more by the beginnings of western European secularism and by Asiatic despotism than by Roman or Byzantine law. Though strong patriarchs of Constantinople were generally able to oppose open violations of dogma and canon law by the emperors, their Russian successors were quite powerless; a single metropolitan of Moscow, St. Philip (metropolitan 1566–68), who dared to condemn the excesses of Ivan IV, was deposed and murdered.

A crisis of the "third Rome" ideology occurred in the middle of the 17th century. Nikon (reigned 1652–58), a strong patriarch, decided to restore the power and prestige of the church by declaring that the patriarchal office was superior to that of the tsar. He forced the tsar Alexis Romanov to repent for the crime of his predecessor against St. Philip and to swear obedience to the church. Simultaneously, Nikon attempted to settle a perennial issue of Russian church life: the problem of the liturgical books. Originally translated from the Greek, the books suffered many corruptions through the centuries and contained numerous mistakes. In addition, the different historical developments in Russia and in the Middle East had led to differences between the liturgical practices of the Russians and the Greeks. Nikon's solution was to order the exact compliance of all the Russian practices with the contemporary Greek equivalents. His liturgical reform led to a major schism in the church. The Russian masses had taken seriously the idea that Moscow was the last refuge of Orthodoxy. They wondered why Russia had to accept the practices of the Greeks, who had betrayed Orthodoxy in Florence and had been justly punished by God, in their view, by becoming captives of the infidel Turks. The reformist decrees of the patriarch were rejected by millions of lower clergy and laity who constituted the Raskol, or schism of the "Old Believers." Nikon was ultimately deposed for his opposition to the tsar, but his liturgical reforms were confirmed by a great council of the church that met in the presence of two Eastern patriarchs (1666–67).

Schism in the Russian Church

The reforms of Peter the Great (reigned 1682–1725). The son of Tsar Alexis, Peter the Great, changed the historical fate of Russia by radically turning away from the Byzantine heritage and reforming the state according to the model of Protestant Europe. Humiliated by his father's temporary submission to Patriarch Nikon, Peter prevented new patriarchal elections after the death of Patriarch Adrian in 1700. After a long vacancy of the see, he abolished the patriarchate altogether (1721) and transformed the central administration of the church into a department of the state, which adopted the title of "Holy Governing Synod." An imperial high commissioner (*Oberprokuror*) was to be present at all meetings and, in fact, to act as the administrator of church affairs. Peter also issued a lengthy *Spiritual Regulation* (*Dukhovny Reglament*) that served as bylaws for all religious activities in Russia. Weakened by the schism of the "Old Believers," the church found no spokesman to defend its rights and passively accepted the reforms.

With the actions of Peter, the Church of Russia entered a new period of its history that lasted until 1917. The immediate consequences were not all negative. Peter's ecclesiastical advisers were Ukrainian prelates, graduates of the Kievan academy, who introduced in Russia a Western system of theological education; the most famous among them was Peter's friend, Feofan Prokopovich, archbishop of Pskov. Throughout the 18th century, the Russian Church also continued its missionary work in Asia and produced several spiritual writers and saints: St. Mitrofan of Voronezh (died 1703), St. Tikhon of Zadonsk (died 1783)—an admirer of the German Lutheran Johann Arndt and of German Pietism—as well as other eminent prelates and scholars such as Platon Levshin, metropolitan of Moscow (died 1803). All attempts at challenging the power of the tsar over the church, however, always met with failure. The metropolitan of Rostov, Arseny Matsiyevich, who opposed the secularization of church property by the empress Catherine the Great, was deposed and died in prison (1772). The atmosphere of secularistic officialdom that prevailed in Russia was not favourable for a revival of monasticism, but such a revival did take place through the efforts of a young Kievan scholar, Paissy Velichkovsky (1722–94), who became the abbot of the monastery of Neamts in Romania. His Slavonic edition of the *Philocalia* con-



"Planting the Tree of the Russian Nation," icon by Simon Ushakov, 1668. Detail shows the Kremlin, Ivan I Kalita, and Tsar Alexis. In the Tretyakov Art Gallery, Moscow.
Novosti Press Agency

tributed to the revival of Hesychast traditions in Russia in the 19th century.

THE ORTHODOX CHURCHES IN THE 19TH CENTURY

Autocephalies in the Balkans. The ideas of the French Revolution, the nationalistic movements, and the ever-living memory of past Christian empires led to the gradual disintegration of Turkish domination in the Balkans. According to a pattern existing since the late Middle Ages, the birth of national states was followed by the establishment of independent, "autocephalous" Orthodox churches. Thus the collapse of the Ottoman rule was accompanied by the rapid shrinking of the actual power exercised by the patriarch of Constantinople. Paradoxically, the Greeks, for whom—more than anyone—the patriarchate represented a hope for the future, were the first to organize an independent church in their new state.

The establishment of independent nationalistic churches

In Greece. In 1821 the Greek revolution against the Turks was officially proclaimed by the metropolitan of Old Patras, Germanos. The patriarchate, being the official Turkish-sponsored organ for the administration of the Christians, issued statements condemning and even anathematizing the revolutionaries. These statements, however, failed to convince anyone, least of all the Turkish government, who on Easter Day in 1821 had the ecumenical (Constantinopolitan) patriarch Gregory V hanged from the main gate of the patriarchal residence as a public example. Numerous other Greek clergy were executed in the provinces. After this tragedy, the official loyalty of the patriarchate was, of course, doubly secured. Unable either to communicate with the patriarchate or to recognize its excommunications, the bishops of liberated Greece gathered in Návplion and established themselves as the synod of an autocephalous church (1833). The ecclesiastical regime adopted in Greece was modelled after that of Russia: a collective state body, the Holy Synod, was to govern the church under strict government control. In 1850 the patriarchate was forced to recognize what was by then a *fait accompli*, and granted a charter of autocephaly (*tómos*) to the new Church of Greece.

In Serbia. The independence of Serbia led, in 1832, to the recognition of Serbian ecclesiastical autonomy. In 1879 the Serbian Church was recognized by Constantinople as autocephalous under the primacy of the metropolitan of Belgrade. This church, however, covered only the territory of what was called "old Serbia." The small state of Montenegro, always independent from the Turks, had its own metropolitan in Cetinje. This prelate, who was also the civil and military leader of the nation, was consecrated either in Austria, or, as in the case of the famous bishop-poet Pyotr II Negosh, in St. Petersburg (1833).

In the Austro-Hungarian empire, two autocephalous churches, with jurisdiction over Serbs, Romanians, and other Slavs, were in existence during the second half of the century. These were the patriarchate of Sremski-Karlovci (Karlowitz), established in 1848, which governed all the Orthodox in the Kingdom of Hungary; and the metropolitanate of Czernowitz (now Chernovtsy) in Bukovina, which, after 1873, also exercised jurisdiction over two Serbian dioceses (Zara and Kotor) in Dalmatia. The Serbian dioceses of Bosnia and Herzegovina, acquired by Austria in 1878, remained autonomous but were never completely independent from Constantinople.

In Romania. The creation of an independent Romania, after centuries of foreign control by Bulgarians, Turks, Greek-Phanariots, and, more recently, Russians, led in 1865 to the self-proclamation of the Romanian Church as autocephalous, even against the violent protests of the Phanar. As in Greece, the new church was under the strict control of the pro-Western government of Prince Alexandru Cuza. Finally, as in the Greek case, Constantinople recognized the Romanian autocephaly under the metropolitan of Bucharest (1885). The Romanians of Transylvania, still in Austria–Hungary, remained under the autocephalous metropolitan of Sibiu and others under the church of Czernowitz.

In Bulgaria. The re-establishment of the Church of Bulgaria eventually was secured, but not without tragedy and even a schism; this happened mainly because the issue of re-establishing the autocephalous church arose at a time when both Greek and Bulgarian populations lived side by side in Macedonia, Thrace, and Constantinople itself, though still within the framework of the Ottoman imperial system. After the Turkish conquest, and especially in the 17th and 18th centuries, the Bulgarians were

governed by Greek bishops and were often prevented from worshipping in Slavonic. This enforced policy of Hellenization was rejected in the 19th century when Bulgarians began to claim not only a native clergy but also equal representation on the higher echelons of the Christian *millet; i.e.*, the offices of the patriarchate. These claims were met with firm resistance by the Greeks. The alternative was a national Bulgarian Church, which was created by a sultan's firman (decree) in 1870. The new church was to be governed by its own Bulgarian exarch, who resided in Constantinople itself and governed all the Bulgarians who recognized him. The new situation was uncanonical, because it sanctioned the existence of two separate ecclesiastical structures on the same territory. Ecumenical Patriarch Anthimus VI convened a synod in Constantinople, which also included the Greek patriarchs of Alexandria and Jerusalem (1872). The council condemned "phyletism"—the national or ethnic principle in church organization—and excommunicated the Bulgarians, who were certainly not alone guilty of "phyletism." This schism lasted until 1945, when a reconciliation took place with full recognition of Bulgarian autocephaly within the limits of the Bulgarian state.

The problem of the national or ethnic principle

After their liberation from the Turkish yoke, the Balkan churches freely developed both their national indentities and their religious life. Theological faculties, generally following German models, were created in Athens, Belgrade (in Yugoslavia), Sofia (in Bulgaria), and Bucharest (in Romania). The Romanian Church introduced the full cycle of the liturgical offices in vernacular Romanian. But these positive developments were often marred by nationalistic rivalries. In condemning "phyletism," the synod of Constantinople (1872) had, in fact, defined a basic problem of modern Orthodoxy.

The church in imperial Russia. *The Spiritual Regulation* of Peter the Great remained in force until the very end of the Russian Empire (1917). Many Russian churchmen consistently complained against the submission of the church to the state, but there was little they could do except to lay plans for future reforms. This they did not fail to do, and in the 20th century the necessary changes were rapidly enacted. Though Peter himself and his first successors tended to deal personally and directly with church affairs, the tsars of the 19th century delegated much authority to the *Oberprokurors*, who received a cabinet rank in the government and were the real heads of the entire administration of the church. One of the most debilitating aspects of the regime was the legal division of Russian society by a rigid caste system. The clergy was one of the castes with its own school system, and there was little possibility for its children to choose another career.

In spite of these obvious defects, the church kept its self-awareness, and among the episcopate such eminent figures as Philaret of Moscow (1782–1867) promoted education, theological research, biblical translations, and missionary work. In each of its 67 dioceses, the Russian Church created a seminary for the training of priests and teachers. In addition, four theological academies, or graduate schools, were established in major cities (Moscow, 1769; St. Petersburg, 1809; Kiev, 1819; Kazan, 1842). They provided a generally excellent theological training for both Russians and foreigners. The rigid caste system and the strictly professional character of these schools, however, were serious obstacles to their seriously influencing society at large. It was, rather, through the monasteries and their spirituality that the church began to reach the intellectual class. More influential than the rigid discipline of the large monastic communities, the prophetic ministry of the "elders" (*startsy*), who acted as living examples of the standards of the spiritual life or as advisers and confessors, attracted large masses of the common people, and also intellectuals. St. Seraphim of Sarov (1759–1833), for example, lived according to the standards of the ancient Hesychast tradition that had been revived in the Russian forests. The *startsy* of Optino —Leonid (1768–1841), Makarius (1788–1860), and Ambrose (1812–91)—were visited not only by thousands of ordinary Christians but also by the writers Nikolay Gogol, Leo Tolstoy, and Fyodor Dostoyevsky. The latter was inspired by the *startsy* when he described in his novels monastic figures such as Zossima in *The Brothers Karamazov*. From the ranks of an emerging group of Orthodox lay intellectuals, the production of a living theology—if less scholarly than in the academies—was taking shape. The great influence of a lay theologian like Aleksey Khomyakov (1804–60), who belonged to the Slavophile (pro-Slavic) circle before it acquired a political flavour, eventually helped in the conversion to Orthodoxy at the end of the century of such leading Marxists as Sergey Bulgakov (1871–1944) and Nikolay Berdyayev (1874–1948).

Significance of Russian monasticism

Missionary expansion also continued, with the theological Academy of Kazan specializing in missions to Islām. By the year 1903, the liturgy was celebrated in more than 20 languages in the region of Kazan. An Orthodox missionary society coordinated these missionary efforts. Nikolay Kasatkin (1836–1913), canonized as a saint in 1970, founded an indigenous church in Japan. A married priest, Ivan Veniaminov, was sent to Alaska in 1822 in order to continue the work of the first mission. His work among the Aleuts, Eskimos, and Tlingit Indians is still continuing today. Under the name of Innocent, Veniaminov was the first Orthodox bishop to reside on American soil (1840).

Disproportionately larger and richer than its sister churches of the Balkans and the Middle East, the Church of Russia included, in 1914, more than 50,000 priests, 21,330 monks, and 73,299 nuns. It supported thousands of schools and missions. It cooperated with the Russian government in exercising great influence in Mid-Eastern affairs. Thus, with Russian help, an Arab (Meletios Doumani) rather than a Greek was elected for the first time as patriarch of Antioch (1899). With the successive partitions of Poland and the reunions with Russia of Belorussian and Ukrainian territories, many Eastern Catholic descendants from those who had joined the Roman communion in Brest-Litovsk (1596) returned to Orthodoxy.

The condition of the Russian Church on the eve of World War I

After 1905, Tsar Nicholas II gave his approval for the establishment of a preconciliar commission charged with the preparation of an all-Russian Church Council. The avowed goal of the planned assembly was to re-establish the church's independence, lost since Peter the Great, and eventually to restore the patriarchate. This assembly, however, was fated to meet only after the fall of the empire.

THE ORTHODOX CHURCH SINCE WORLD WAR I

The almost complete disappearance of Christianity in Asia Minor, the regrouping of the Orthodox churches in the Balkans, the tragedy of the Russian Revolution, and the Orthodox diaspora in the West radically changed the entire structure of the Orthodox world.

The Russian Revolution and the Soviet period. The Church of Russia was less unprepared than generally believed to face the revolutionary turmoil. Projects of necessary reforms had been prepared since 1905, and most clergy did not feel particularly attached to the fallen regime that had deprived the church of its freedom for several centuries. During the rule of the provisional government, in August 1917 a council representing the entire church met in Moscow, including 265 members of the clergy and 299 laymen. The democratic composition and program of the council had been planned by the Pre-Conciliar Commission. It adopted a new constitution of the church that provided for the re-establishment of the patriarchate, the election of bishops by the dioceses, and the representation of laymen on all levels of church administration. It was only in the midst of the new revolutionary turmoil, however, that Tikhon, metropolitan of Moscow, was elected patriarch (October 31, 1917—six days after the Bolshevik take-over). The bloody events into which the country was plunged did not allow all of the reforms to be carried out, but new bishops were elected in several dioceses by the people.

The Bolshevik government, because of its Marxist ideology, considered all religion as the "opium of the

The effects of the Russian Revolution on the church

people." On January 20, 1918, it published a decree depriving the church of all legal rights, including that of owning property. The stipulations of the decree were difficult to enforce immediately, and the church remained a powerful social force for several years. The patriarch replied to the decree by excommunicating the "open or disguised enemies of Christ," without naming the government specifically. He also made pronouncements on political issues that he considered of moral importance: in March 1918 he condemned the peace of Brest-Litovsk that brought an unsatisfactory armistice between Russia and the Central Powers, and in October he addressed an "admonition" to Lenin, calling on him to proclaim an amnesty. Tikhon was careful, however, not to appear as a counter-revolutionary and in September 1919 called the faithful to refrain from supporting the Whites (anti-Communists) and to obey those decrees of the Soviet government that were not contrary to their Christian conscience.

The independence of the church suffered greatly after 1922. In February of that year, the government decreed the confiscation of all valuable objects preserved in the churches. The patriarch would have agreed to that measure if he had had the means to check on the government contention that all confiscated church property would be used to help the starving population on the Volga. The government refused all guarantees but supported a group of clergy who were ready to cooperate with it and to overthrow the patriarch. While Tikhon was under house arrest, this group took over his office and soon claimed the allegiance of a sizable proportion of bishops and clergy. This became known as the schism of the "Renovated" or "Living" Church, and it broke the internal unity and resistance of the church. Numerous bishops and clergy faithful to the patriarch were tried and executed, including the young and progressive metropolitan Benjamin of Petrograd. The "Renovated" Church soon broke the universal discipline of Orthodoxy by admitting married priests to the episcopate and by permitting widowed priests to remarry.

Upon his release, Tikhon condemned the schismatics and many clergy returned to his obedience. But he also published a declaration affirming that he "was not the enemy of the Soviet government" and dropped any opposition to the authorities, setting a pattern that is still continued. Tikhon's attitude of conformism did not bring immediate results. His designated successors (after he died in 1925) were all arrested. In 1927 the "substitute *locum tenens*" (holder of the position) of the patriarchate, Metropolitan Sergius, pledged loyalty to the Soviet government. Nevertheless, under the rule of Joseph Stalin in the late 1920s and 1930s, the church suffered a bloody persecution that claimed thousands of victims. By 1939 only three or four Orthodox bishops and 100 churches could officially function: the church was practically suppressed.

Post-World War II developments in the Soviet Union

A spectacular reversal of Stalin's policies occurred, however, during World War II. Sergius was elected patriarch in 1943 and the "Renovated" schism was ended. Under Sergius' successor, Patriarch Alexis (1945–70), the church was able to open 25,000 churches and the number of priests reached 33,000. But a new anti-religious move was initiated by Prime Minister Nikita Khrushchev in 1959–64, reducing the number of open churches to less than 10,000.

With the election of Patriarch Pimen (1971), which followed the death of Alexis, the future of the church in the Soviet Union is still uncertain. In spite of 50 years of antireligious propaganda, the church still commands the loyalty of millions, yet, in order to survive, its leaders must show total cooperation with the government.

The Balkans and eastern Europe. In bringing about the fall of the Turkish, Austrian, and Russian empires, World War I provoked significant changes in the structures of the Orthodox Church. On the western borders of the Soviet Union, in the newly born republics of Finland, Estonia, Latvia, and Lithuania, the Orthodox minorities established themselves as autonomous churches. The first three joined the jurisdiction of Constantinople, and the Lithuanian diocese remained nominally under Moscow. In Poland, which now included several million Belorussians and Ukrainians, the ecumenical patriarch established an autocephalous church (1924) over the protests of Patriarch Tikhon. After World War II, the Estonian, Latvian, and Lithuanian autonomies were again suppressed, and in Poland the Orthodox Church was first reintegrated to the jurisdiction of Moscow and later was declared autocephalous again (1948).

In the Balkans, changes were even more significant. The five groups of Serbian dioceses (Montenegro, patriarchate of Karlovci, Dalmatia, Bosnia-Herzegovina, Old Serbia) were united (1920–22) under one Serbian patriarch, residing in Belgrade, the capital of the new Yugoslavia. Similarly, the Romanian dioceses of Moldavia-Walachia, Transylvania, Bukovina, and Bessarabia formed the new patriarchate of Romania (1925), the largest autocephalous church in the Balkans. Finally, in 1937, after some tension and a temporary schism, the patriarchate of Constantinople recognized the autocephaly of the Church of Albania.

Relations between church and state in the Balkans

After World War II, Communist regimes were established in the Balkan states. There were no attempts, however, at liquidating the churches entirely, similar to the persecutions that took place in Russia in the 1920s and 1930s. In both Yugoslavia and Bulgaria, church and state were legally separated. In Romania, paradoxically, the Orthodox Church remained legally linked to the Communist state. With its solid record of resistance to the Germans, the Serbian Church was able to preserve more independence from the government than its sister churches of Bulgaria and Romania. Generally speaking, however, all the Balkan churches adopted an attitude of loyalty to the new regime, according to the pattern given by the patriarchate of Moscow. At that price, they could keep some theological schools, some publications, and the possibility to worship. This is also the case of the Orthodox minority in Czechoslovakia, which was united and organized into an autocephalous church by the patriarchate of Moscow in 1951. Only in Albania did a Communist government announce the total liquidation of organized religion, following the Cultural Revolution of 1966–68.

Among the national Orthodox churches, the Church of Greece is the only one that preserved the legal status it acquired in the 19th century as the national state church. As such, it was supported by the successive political regimes of Greece. It could also develop an impressive internal mission. The Brotherhood Zoe ("Life"), organized according to the pattern of Western religious orders, was quite successful in creating a large system of church schools, in sponsoring various Christian organizations of professional men and women, and in publishing religious literature, yet did not always succeed in closing the gap between a conservative church and the Western-oriented intellectuals.

Conditions of the Eastern patriarchates

The Orthodox Church in the Middle East. As a result of the Greco–Turkish War, the entire Greek population of Asia Minor was transferred to Greece (1922); the Orthodox under the immediate jurisdiction of the ecumenical patriarchate of Constantinople were thus reduced to the Greek population of Istanbul and its vicinity. This population, rapidly shrinking in recent years, is now reduced to a few thousand. Still recognized as holding an honorary primacy among the Orthodox churches, the ecumenical patriarchate also exercises jurisdiction over several dioceses of the "diaspora" and, by consent of the Greek government, over the Greek islands. The impressive personality of Patriarch Athenagoras I (1948–72), who was succeeded by Demetrios, contributed to its prestige on the pan-Orthodox and ecumenical levels. The patriarchate convened pan-Orthodox conferences in Rhodes, Belgrade, and Geneva and has been coordinating preparations for a "Great Council" of the Orthodox Church scheduled for the 1970s.

Together with the ecumenical patriarchate, the ancient sees of Alexandria, Antioch, and Jerusalem are remnants of the Byzantine imperial past, but under the present conditions they still possess many opportunities

of development: Alexandria, as the centre of emerging African communities (see below *The Orthodox diaspora and missions*); Antioch, as the largest Arab Christian group, with dioceses in Syria, Lebanon, and Iraq; and Jerusalem, as the main custodian of the Christian holy places in that city.

The two ancient churches of Cyprus and Georgia, with their quite peculiar history, continue to play important roles among the Orthodox sister churches. Autocephalous since 431, the Church of Cyprus survived the successive occupations, and often oppressions, by the Arabs, the Crusaders, the Venetians, the Turks, and the English. Following the pattern of all areas where Islām was predominant, the archbishop is traditionally seen as the ethnarch of the Greek Christian Cypriots. Archbishop Makarios also became the first president of the independent Republic of Cyprus in 1960. The Church of Georgia, isolated in the Caucasus in a country that became part of the Russian Empire in 1801, is the witness of one of the most ancient Christian traditions. It received autocephaly from its mother Church of Antioch as early as the 6th century and developed a literary and artistic civilization in its own language. Its head bears the traditional title of "Catholicos-Patriarch." When the Russians annexed the country in 1801, they suppressed Georgia's autocephaly and the church was governed by a Russian "exarch" until 1917 when the Georgians re-established their ecclesiastical independence. Fiercely persecuted during the 1920s, the Georgian Church survives to the present day as an autocephalous patriarchate.

Orthodoxy in the United States. The first Orthodox communities in what is today the continental United States were established in Alaska and on the West Coast, as the extreme end of the Russian missionary expansion through Siberia (see above *The church in imperial Russia*). Russian monks settled on Kodiak Island in 1794. Among them was St. Herman (died 1837, canonized 1970), an ascetic and a defender of the natives' rights against their exploitation by ruthless Russian traders. After the sale of Alaska to the United States, a separate diocese "of the Aleutian Islands and Alaska" was created by the Holy Synod (1870). After the transfer of the diocesan centre to San Francisco and its renaming as the diocese "of the Aleutian Islands and North America" (1900), the original church establishment exercised its jurisdiction on the entire North American continent. In the 1880s, it accepted back into Orthodoxy hundreds of "Uniate" parishes of immigrants from Galicia and Carpatho-Russia, particularly numerous in the northern industrial states and in Canada. It also served the needs of immigrants from Serbia, Greece, Syria, Albania, and other countries. Some Greek and Romanian communities, however, invited priests directly from the mother country without official contact with the American bishop. In 1905 the American archbishop Tikhon (future patriarch of Moscow) presented to the Russian synod the project of an autonomous, or autocephalous, church of America, whose structure would reflect the ethnic pluralism of its membership. He also foresaw the inevitable Americanization of his flock and encouraged the translation of the liturgy into English.

Organization and canonical status of American Orthodox churches

These projects, however, were hampered by the tragedies that befell the Russian Church following the Russian Revolution. The administrative system of the Russian Church collapsed. The non-Russian groups of immigrants sought and obtained their affiliation with mother churches abroad. In 1921 a "Greek Archdiocese of North and South America" was established by the ecumenical patriarch Meletios IV Metaxakis. Further divisions within each national group occurred repeatedly, and several independent jurisdictions added to the confusion.

A reaction against this chaotic pluralism manifested itself in the 1950s. More cooperation between the jurisdictions and a more systematic theological education contributed to an increased desire for unity. A standing conference of bishops was established in 1960. In 1970, the patriarch of Moscow, reviving Tikhon's project of 1905, formally proclaimed its diocese in America (which had been in conflict with Moscow since 1931 on the issue of "loyalty" to the Soviets) as the autocephalous Orthodox Church in America, totally independent from administrative connections abroad. The ecumenical patriarchate of Constantinople, however, protested this move, turned down a request for autonomy presented by the Greek archdiocese (the largest single Orthodox body in the U.S.), and reiterated its opposition to the use of English in the liturgy (1970). This latest crisis of American Orthodoxy involves the very understanding of the Orthodox presence in the Western world, centring on the question of the utility of preserving the ethnic ties of the past.

The Orthodox diaspora and missions. Since World War I, millions of east Europeans were dispersed in various areas where Orthodox communities had never existed before. The Russian Revolution provoked a massive political emigration, predominantly to western Europe and particularly France. It included eminent churchmen, theologians, and Christian intellectuals, such as Bulgakov, Berdyayev, and V.V. Zenkovsky, who were able not only to establish in Paris a theological school of great repute but also to contribute significantly to the ecumenical movement. In 1922 Patriarch Tikhon appointed Metropolitan Evlogy to head the émigré churches, with residence in Paris. The authority of the metropolitan was challenged, however, by a group of bishops who had left their sees in Russia, retreating with the White armies, and who had found refuge in Sremski-Karlovci as guests of the Serbian Church. Despite several attempts at reconciliation, the "Synod" of Karlovci, proclaiming its firm attachment to the principle of tsarist monarchy, refused to recognize any measure taken by the re-established patriarchate of Moscow. This group transferred its headquarters to New York and is also known as the "Russian Orthodox Church outside of Russia." It has no canonical relation with the official Orthodox patriarchates and churches. A "Ukrainian Orthodox Church in exile" finds itself in a similarly irregular canonical situation. Other émigré groups found refuge under the canonical auspices of the ecumenical patriarchate.

Theological and missionary activities after World War I

After World War II, a very numerous Greek emigration took place to western Europe, Australia, New Zealand, and Africa. In East Africa, without much initial effort on their part, these Greek-speaking emigrants have attracted a sizable number of black Christians, who have discovered in the Orthodox liturgy and sacramental worship a form of Christianity more acceptable to them than the more dogmatic institutions of Western Christianity. Also, in their eyes, Orthodoxy has the advantage of having no connection with the colonial regimes of the past. Orthodox communities, with an ever increasing number of native clergy, are spreading in Uganda, Kenya, and Tanzania. Less professionally planned than the former Russian missions in Alaska and Japan, these young churches constitute an interesting development in African Christianity.

Ecumenical involvement. Between the two world wars, many Orthodox churchmen of the ecumenical patriarchate of Constantinople, of Greece, of the Balkan churches, and of the Russian emigration took part in the ecumenical movement. After World War II, however, the churches of the Communist-dominated countries failed to join the newly created World Council of Churches (1948): only Constantinople and Greece did so. The situation changed drastically in 1961, when the patriarchate of Moscow applied for membership and was soon followed by other autocephalous churches. Before and after 1961, the Orthodox consistently declared that their membership did not imply any relativistic understanding of the Christian truth, but that they were ready to discuss with all Christians the best way of restoring the lost unity of Christendom, as well as problems of common Christian action and witness in the modern world.

During the reign of Pope John XXIII, when Roman Catholicism became actively involved in ecumenism, the Orthodox—after some hesitation—contributed to the new atmosphere. The spectacular meetings in the 1960s

between Patriarch Athenagoras and the Pope in Jerusalem, Istanbul, and Rome, the symbolic lifting of ancient anathemas, and other gestures were signs of rapprochement, although they are sometimes mistakenly interpreted as if they were ending the Schism itself. In the Orthodox view, full unity can be restored only in the fullness of truth witnessed by the entire church and sanctioned in sacramental communion.

BIBLIOGRAPHY. The best special account of the church in Byzantium until 1261 is found in G. EVERY, *The Byzantine Patriarchate*, 2nd ed. rev. (1962); general information about the entire history of the Eastern Church is given in A. SCHMEMANN, *The Historical Road of Eastern Orthodoxy* (1963; orig. pub. in Russian, 1954). Of special interest for 19th and 20th century history of Orthodoxy are J. MEYENDORFF, *L'Église orthodoxe: hier et aujourd'hui* (1960; Eng. trans., *The Orthodox Church: Its Past and Its Role in the World Today*, 1962); and T. WARE, *The Orthodox Church* (1963). S.A. RUNCIMAN, *The Great Church in Captivity: A Study of the Patriarchate of Constantinople from the Eve of the Turkish Conquest to the Greek War of Independence* (1968), is an excitingly written critical evaluation of the Greek struggle for survival in the Ottoman Empire. CHARLES A. FRAZEE, *The Orthodox Church and Independent Greece, 1821–1852* (1969), uses contemporary diplomatic archives to describe the role of the Greek Church in an event that also involved all major European powers. The best general account of Russian medieval Christianity is by G. FEDOTOV, *The Russian Religious Mind*, 2 vol. (1946–66); on contemporary problems and events, see S. BOLSHAKOFF, *The Foreign Missions of the Russian Orthodox Church* (1943); and particularly, N. STRUVE, *Les Chrétiens en U.R.S.S.*, 2nd rev. ed. (1964; Eng. trans., *Christians in Contemporary Russia*, 1966).

(J.M.)

Eastern Rite Churches

Eastern rite churches, consisting of the Byzantine, Alexandrian, Antiochene, Chaldean, and Armenian rites, are Eastern Christian churches that trace their origins to various ancient national or ethnic Christian groups in the East but are in union with the Western Catholic Church under the jurisdiction of the papacy at Rome.

Nature and significance. The Eastern rite churches include the heterogeneous groupings of Eastern Catholics in communion with the Roman Catholic Church. Eastern Catholics, in contrast to Western, or Latin, Catholics, who have a history continuous from the 1st century AD, trace their origins largely to the failure of the ecclesiastical authorities at the Council of Ferrara–Florence in 1439 to unite Christians of the East and West. Stimulated by this unsuccessful beginning, however, and encouraged also by the later missionary activities of monastic orders, such as the Jesuits, Dominicans, Franciscans, and Capuchins, the proponents of the goal of the eventual reunion of Eastern and Western Christians began to achieve some elements of success. The Brest-Litovsk Union of 1596—under which all but two Ukrainian Orthodox bishops submitted to the demand of their Polish Catholic king that they accept the primacy of the pope—in a substantial way signalled the effective advent of Eastern rite churches. Other smaller groups had united with Rome in previous centuries, but the Ukrainians who were united with Rome at this time were the largest branch of Eastern Catholics to move in that direction. Prior to this event, Eastern Catholics were limited to Italo-Albanians in southern Italy and Sicily, a large number of Maronites (Lebanese Christians of the Syro-Antiochene rite) who became associated with Rome in the 12th century, and some Armenians in the Syria–Lebanon region who also trace their relationship with Rome to the 12th century. A number of Nestorians (followers of the 5th-century patriarch of Constantinople, Nestorius, who was declared a heretic) were united with Rome in 1551, Ruthenians (an east central European people) in 1595, Romanians of Transylvania in 1698, and Melkites (Syrian Christians of the Byzantine rite) in 1724. Political factors also played a role during the reunion process; Eastern Christians were, and are, much influenced by nationalistic loyalities in their respective regions. As these various groups of Eastern Catholics grew in number, Rome encouraged and established ecclesiastical hierarchies.

Significance of the Brest-Litovsk Union

To understand more fully the hierarchies of Eastern Catholics, the larger context of Eastern Christianity must be reviewed. Eastern Catholic churches correspond in kind to the more numerous Byzantine Orthodox churches and the Eastern independent, or Oriental, churches; *i.e.*, those that do not accept the decrees of the ecumenical Council of Chalcedon (451). Within this fuller context, Eastern Catholics as a group are the smallest segment within Eastern Christianity. Furthermore, from the viewpoint of the Byzantine and non-Chalcedonian traditions, Eastern Catholics may be looked upon with suspicion, primarily because of the latinizing influence found in their ranks. Hence the majority of Byzantine Orthodox and Eastern independent churches characterize Eastern Catholics as "uniate" churches. The expression uniate is taken from the Slavic *uniya*, a term coined by the opponents of the Brest-Litovsk Union. "Uniatism" implies hybridism, or the tendency for latinization, and hence a betrayal of one's ancient and nationalistic tradition. Eastern rite churches would prefer to be considered as united churches rather than uniate, with its negative implications.

Eastern rite churches make manifest the pluralistic composition of the Roman Catholic tradition. Permission for the various Eastern rites is a concession from Roman Catholic canon law, for a rite indicates more than liturgy; it points to an entire life style and discipline of a church tradition. Eastern Catholic rites—in contradistinction to Western Catholic rites—permit a married clergy and the immediate admission of baptized infants to the sacraments of communion (the Lord's Supper) and confirmation. In the "Decrees on the Catholic Churches of the Eastern Rite" adopted by the second Vatican Council (1962–65), the Roman pontiff reaffirmed the pledge of his predecessors to preserve the rites of the Eastern churches. "All of the Eastern Rite members should know and be convinced," states the decree, "that they can and should always preserve their legitimate liturgical rite and their established way of life, and that these may not be altered except to obtain for themselves an organic improvement." At present, there exist five distinct Eastern rite traditions: the Alexandrian, Antiochene, Armenian, Chaldean, and Byzantine.

The Byzantine rite. By far the most significant Eastern rite is the Byzantine. It affects the most persons, is the most widely distributed, and has had an impact upon the four other rites. From the original Byzantine rite, 13 canonical rites were derived: Albanian (dispersed; *i.e.*, no longer indigenous to the land of its origin), Belorussian (dispersed, but with a bishop in London), Bulgarian, Greek, Hungarian, Italo-Albanian, Melkite, Romanian (suppressed by the government in 1948), Russian (dispersed), Ruthenian, Slovak (suppressed by the government in 1950, re-established in 1968), Ukrainian, and Yugoslav. Several of the dispersed or suppressed rites have memberships in other countries, especially in the United States. Originally the Byzantine rite represented the Church of Constantinople and was Greek, but as Constantinople extended its political and ecclesiastical influence, the rite was translated into the vernacular of the diverse peoples it embraced. Liturgically, the Byzantine rite is based upon the liturgies of St. Basil, a 4th-century theologian and monastic leader, and St. John Chrysostom, a 4th-century scriptural interpreter and patriarch of Constantinople, as well as on the Presanctified Liturgy (in which bread and wine consecrated on the previous Sunday is used in communion), which legend attributes to Pope St. Gregory the Great (reigned 590–604). The Liturgy of St. Basil is celebrated ten times a year, the Liturgy of the Presanctified during Lent, and the Liturgy of St. Chrysostom during the greater part of the year. The authorship of these liturgies extends back to the temple ritual tradition of the Jews and the accumulative contributions of the early Church Fathers. In effect, the liturgies of Basil and Chrysostom represent a culmination of a process of accumulation and accretion rather than exact authorship. The liturgies generally are divided into three parts: the prayers of preparation, liturgy of the catechumens, and the liturgy of the faithful.

Distribution of the Byzantine rite

The Byzantine rite embraces a wide group of peoples. The Syrian Christians are known as Melkites (imperialists; Syrian *malka*, "king" or "emperor"). Melkites today are found mainly in the patriarchates of Alexandria, Antioch, and Jerusalem. (A patriarchate comprises all dioceses of a particular rite in a certain territory and is presided over by a patriarch.) Since the time of the Catholic Melkite patriarch Cyril VI (1724–59), there has been an uninterrupted line of Catholic Melkite patriarchs. Residing at Damascus (Syria) or Cairo (Egypt), the patriarch assumes the title "Patriarch of Antioch and of all the East, of Alexandria and of Jerusalem." The Melkites have been divided between Catholic and Orthodox Melkites, but union was briefly achieved under Patriarch Dorotheos I, who accepted the decrees of the Council of Ferrara–Florence in the 16th century. Under the rule of Maximos III Mazlūm (1833–55), progress and reform began and then continued in the 20th century under the ecumenical leadership and foresight of Maximos IV Sa'igh, elected patriarch in 1947 and made a cardinal in 1965.

Among the Catholics of the Byzantine rite, the Ukrainians and Ruthenians are among the most numerous migrants to North America. The Rev. Ivan Valansky in 1886 opened the doors of the first Catholic church of Byzantine rite in the United States at Shenandoah, Pennsylvania. Originally from the region west and southwest of Russia, many of these Eastern Catholics left their homeland to avoid being pressed into Orthodox churches. In 1965 their archbishop Josef Slipyj was named a cardinal. Romanian Catholics of the Byzantine rite, also once numerous, experienced similar pressures in 1948 when the government encouraged incorporation into the Orthodox Church of Romania. Though the number of Catholics of Byzantine rite found today among Greeks, Bulgarians, Russians, Italo-Albanians, Georgians, Finns, Estonians, Latvians, Hungarians, Yugoslavians, and Belorussians has dwindled, their widespread existence reveals the extensive influence of the Byzantine rite. Over 2,000,000 Catholics of the Byzantine rite are scattered throughout the non-Communist world. In the 1940s over 5,600,000 Byzantine-rite Catholics lived in countries with Communist governments; the statistics are unavailable for their present numbers.

The Alexandrian rite. Eastern Catholics of the Alexandrian rite are found in North Africa among Egyptians and Ethiopians. Numerically a minority among Coptic and Ethiopian Orthodox Christians, these Catholics identify with the same heritage of apostolic beginnings associated with the Evangelist St. Mark, who, according to legend, travelled in the 1st century AD to Alexandria, then second in importance only to Rome.

Egyptian Catholics. The Catholic Coptic patriarch in Egypt is called significantly the "Patriarch of Alexandria and of All the Preaching of St. Mark." The history of Eastern Catholics in this region, however, has been turbulent. In 6th-century Egypt, two lines of patriarchs were established, one representing the majority of Coptic Christians who were non-Chalcedonian and the other representing the minority in communion with Rome. The

Reunion of Coptic Christians with Rome

Council of Ferrara–Florence led to premature expectations of reunion under the Coptic patriarch John XI, but it never materialized. The 17th century brought renewed gestures for reunion through the efforts of representative Capuchins and Franciscans, but with little results until the 18th century, when a Coptic Orthodox bishop, Amba Athanasius, became a Catholic. In the meantime, Raphael Tukhi, another convert studying in Rome, prepared liturgical books in Coptic and enhanced the beginnings of a Coptic Catholic Church of the Alexandrian rite that began to take shape in the 19th century. In 1895 Rome elevated a Coptic Catholic priest, George MaCaire, who took the name Cyril Makarios, as vicar apostolic to the whole diocese of Egypt, thus giving further recognition to the small number of Eastern Catholics in the land. Further growth was seen in the 20th century; Amba Mark II Khuzam was patriarch until 1958; and his successor, Amba Stephen Sidaruss, was made cardinal in 1965. The Coptic liturgy is derived from the Greek liturgy of Alexandria. Today, Catholic Copts enjoy civic recognition; their strength is found among the least privileged class of Egyptians. They now number about 95,000, with a patriarchate at Alexandria and three other dioceses.

Ethiopian Catholics. The Ethiopian Catholics have never had a strong foothold among the Ethiopian populace, whose Christian roots date from the middle of the 4th century. Early attempts to establish a Catholic Church in Ethiopia followed from the activities of Portuguese explorers in the 16th century, primarily those of a physician named Bermudes. Later, Pope Julius III in 1555 appointed a Jesuit, João Nunes Barreto, as "patriarch of Ethiopia" and sent him and his clerical court to Ethiopia. He died (1561) before reaching Ethiopia. In the 17th century, Pedro (or Pero) Páez, another Jesuit, had more success through his influence upon the autocratic ruler Malak Sagad III (Susenyos), who summarily ordered obedience to Rome and in 1622 installed Afonso Mendes as patriarch. This resulted in a period of bloodshed and chaos among the Christians in Ethiopia. When Sagad was succeeded on the throne by his son Fasilidas, the latter was faithful to the Ethiopian Orthodox Church, and thus the process of forced obedience was halted; in 1636 Mendes was expelled from the country. Against this background, the Ethiopian government banned Catholic priests from the country for some 200 years. Renewed and more discrete measures were begun by Lazarists and Capuchins in the 19th century. In the early 1970s there were administrative centres of Ethiopian Catholics at Addis Ababa (the metropolitan see), Asmara, and Adigrat; seminaries at Asmara and Adigrat; and preseminaries at Keren (in Eritrea), Hebo, and Atat. The greatest response to the Catholic Ethiopians (now numbering over 76,000) is found among the pagan Galla tribe. In 1950 an Ethiopian was appointed by Rome to be titular bishop of Suzusa and given the title "Apostolic Administrator for Catholics of the Ethiopian Rite." The Egyptians and Ethiopians employ their own Coptic and Ge'ez languages, respectively, in their Alexandrian liturgies.

The Antiochene rite. *Maronites.* The largest single group of Eastern Catholics in the Middle East and throughout the world in the early 1970s was the Maronite Church, which adheres to the Antiochene rite. With 600,000 in Lebanon and about 350,000 or more abroad, the Maronites, like their Phoenician ancestral fathers, maintained a spirit of trade and adventure that has established communities around the world. They have about 50 parishes in the United States and sizable groups in Uruguay, Argentina, Brazil, and even Australia and South Africa. There are no non-Catholic Maronites; their patriarch is called "Maronite Patriarch of Antioch and All the Orient." While maintaining local autonomy over their church, Maronites have been in association with Rome since the 12th century. The Maronite patriarch Jeremias II al-Amshiti participated at the Lateran Council in 1215. Maronites have retained their ancient Syriac liturgy and have maintained a deep sense of their religious and social identity in spite of Muslim domination, pressures for latinization, Turkish rule, and influences of a French protectorate in the first half of the 20th century. No doubt this testifies to the spirit of independence nurtured by their isolation in mountain dwellings. Maronites also derive great strength from an uninterrupted history of monastic rule. With a patriarchate at Bkerke (Lebanon), the Maronite rite also has two archdioceses (Aleppo, Syria; and Beirut, Lebanon), eight dioceses, an apostolic administration (governed by Rome), and an apostolic exarchy (an ecclesiastical territory under authority delegated by the pope) in the United States.

Phoenician origins

Syrian Catholics. Unlike the Maronites, the Syrian Catholics emerged from no very unified background. The Syrians were divided after the Council of Chalcedon (451) between Melkites and Jacobites (those not accepting the decrees of Chalcedon). Early efforts in the 13th and 14th centuries by Dominicans and Franciscans to bridge the division in favour of Rome were not successful. Communication between the Jacobite patriarch Naamat-Allah and Pope Gregory XIII in 1583 opened the

Ancient divisions among Syrian Christians

way for future contacts with Rome. By 1656 a Jacobite, Andrew Akhijan, was converted to Catholicism and in 1667 was acknowledged by Pope Clement IX to be the successor to the Jacobite throne at Mardin. Thus, the Catholic patriarchate was initiated among the Syrians. The Catholic succession has continued, and in more recent times Ignatius Gabriel Tappouni (died 1968) was raised to the cardinalate in 1936, the first Syrian to be appointed to this position. Seeking to maintain continuity with the past and encouraged by Rome, Syrian Catholics have traced the Antiochene liturgical rite to the ancient Liturgy of St. James, using the same original Edessene dialect of Syriac. The purity of the language, however, has not healed the checkered history that the Syrian Catholics have known in their relations with the Jacobite Syrians. Within the patriarchate there are about 79,000 members, and approximately 10,000 live elsewhere.

The Malankars. The Jacobites living in the region of Malankar, or Malabar, in southwestern India, like the Syrian Jacobites, reacted to the latinization forced on them in the 16th century by Portuguese traders who did not appreciate fully the Chaldean (or Syro-Oriental) rite practiced by Malabar Christians. In the mid-17th century, under the leadership of an archdeacon, Thomas Parampil, these Indian Catholics turned from the Chaldean rite and followed the Jacobite Syrians of Antioch. The complicated history of the Jacobites of Malabar has been one of the particular tragedies of Indian Christian history. In the 20th century, Ghiverghis Panikerveetil, a priest and educator, created a new monastic community called "Of the Imitation of Christ." In 1925 he was consecrated bishop and took the name of Mar (Lord) Ivanios. He was commissioned by his synod to explore with Rome whether the past history of schism initiated by Parampil could be corrected. Rome subsequently welcomed the bishops along with their adopted Antiochene rite into communion with Rome. At first, only Mar Ivanios and Mar Teothilos responded favourably to the invitation and were received in 1930. These Malankara Catholics, as they are called, distinguish themselves from Chaldean- and Latin-rite Catholics in Malabar, and their growing number (about 162,000) is causing concern among the larger number of Jacobites in South India.

The Chaldean rite. *Chaldeans.* The Chaldean rite is East Syrian, in contrast to the Antiochene rite, which is West Syrian. Both share a common Syriac origin, however, and liturgically are related to the early Jerusalem-Antioch liturgy. Originally, the Christians of this rite were from Mesopotamia and Chaldea, descendants of the ancient Babylonians, later spreading throughout Asia, including India. The name Chaldean was first used at the Council of Ferrara–Florence by Pope Eugenius IV in 1445 when former Nestorians of Cyprus were reconciled with Rome, after having been reconverted to Roman Catholicism en masse in the previous century. In the 16th century, following his profession of faith to Rome, the superior of the Rabbān Hormizd monastery near al-Qūsh (Turkey), John Sulaka, was appointed (1553) by Pope Julius III to be the patriarch to Catholic Nestorians. He was assassinated two years later, and contacts with Rome became weaker. During this period the term Chaldean came into common use to distinguish the followers of Rome from the other Nestorians of the East. Both groups have their roots in the Syrian Church of the East, which was in existence before the birth of Nestorius. The successors of Sulaka (later reverting to Nestorianism) took on the name "Simeon"—Sulaka had called himself Simeon VIII—and the title "Patriarch-Catholicos of Babylon of the Chaldeans." Other Chaldean Catholics continued their relationship with Rome, and under the leadership of the metropolitan of Diyarbakir (Turkey), Joseph, who became a Catholic in 1672, the line of Catholic Chaldean patriarchs was re-established. The Edessene dialect of the Syriac language is used in the liturgy with the Scripture lessons in the vernacular. With its patriarchal archdiocese in Baghdad (Iraq), the Chaldean Rite Church has eight other archdioceses and six dioceses. Approximately 235,000 Chaldeans are under the jurisdiction of the patriarchate.

Origins of the Chaldean rite

Syro-Malabar Catholics. The Chaldean rite was carried by Syrian Christians to the Malabar coast of India, where the Malabar Christians have long considered themselves Christians of St. Thomas, direct descendants of that Apostle's missionary efforts. The Chaldean rite can be traced back to 6th-century contacts with missionaries from Mesopotamia. Their liturgical language was Syriac and their bishops were Chaldeans ("Babylonians"). The Malabarese (Syro-Indians) came to the attention of Rome as a result of conflicts that erupted under Portuguese political control from the 15th to the 17th centuries. The Portuguese sought latinization and control of the Malabarese; the ensuing spirit of discontent culminated in the schism (1653) led by the Malankar leader Thomas Parampil. Through the efforts of Carmelite friars under the apostolic vicar Giuseppe Sebastiani, sent by Pope Alexander VII in 1656 to quell the discontent, the majority of the Malabarese resumed their allegiance to Rome, apparently because Sebastiani had consecrated Alexander Parampil, a cousin of Thomas, as bishop. Upon Alexander's death, however, a Carmelite of the Latin rite was chosen to succeed him. Only recently has indigenous Indian leadership been developed among the Malabarese. Pope Leo XIII appointed three Indian bishops in 1896, and in 1923 Pope Pius XI re-established a Syro-Indian hierarchy, with the archdiocesan seat at Ekrakulam. In 1956 another archdiocese was established at Changanācher. There are, in addition, five suffragan sees and four apostolic exarchies (ecclesiastical jurisdictions with bishops directly under the pope). In the 1970s the vernacular, Malayalam, was taking on increased importance, replacing Syriac, within the liturgical life of the Malabarese. Approximately 1,800,000 persons adhere to the Syro-Malabar Church.

The Armenian rite. Armenians trace their Christian history to the missionary activities of the apostles St. Bartholomew and St. Thaddaeus, but their actual history as a "Christian nation" dates to the beginning of the 4th century when Christianity became the prevailing religion in Armenia under the leadership of Gregory the Parthian, surnamed by the Armenians Lowsavorich' (the Illuminator), who converted their king Tiridates (Trdat) III, and with him the nation, to the Christian faith. Armenians assert that they are the first Christian nation—a distinction that has cost them much hardship and suffering due to persecutions even into the 20th century.

The first Christian nation

The Armenian Catholic Church—under the jurisdiction of Rome rather than under the Armenian Apostolic Orthodox patriarchs of Echmiadzin (Armenian Soviet Socialist Republic) or Cilicia (Lebanon)—has the smallest number of followers. As early as the 12th century there were Armenian Catholics among the Armenians who fled from Muslim oppressors and established the Kingdom of Cilicia in Turkey. This came to an end in 1375, but Armenian monks known as the Friars of Unity laid the groundwork for the future Catholic Armenian Church under Dominican influence. Their goal was realized with the election of the Armenian bishop of Aleppo (Syria) Abraham Ardzivian—who took the name Peter—as catholicos of Sis (Kozan) in 1740 under Pope Benedict XIV. Because of rivalries and numerous difficulties under the Turks, the Catholic see was moved from Sis to Kuraym to Bzummār and, finally, to Beirut (the last three in Lebanon), where it is located today. The head of Armenian Catholics is called "Patriarch of the Catholic Armenians and Catholicos of Cilicia" and has always taken the name of Peter. The jurisdiction of the Armenian Catholic patriarchs includes Syria, Lebanon, Egypt, and Iraq, as well as Turkey. Other Armenian Catholics under the jurisdiction of Rome are in France, Greece, and Romania. The most illustrious patriarch has been Gregory Peter XV Agagianian (died 1971), elected in 1937 and made a cardinal in 1945. Due to the press of duties as the head of the Congregation for the Propagation of the Faith, for the Roman Catholic Church, he resigned as patriarch in 1962, and his successor became Ignatius Peter XVI Batanian. The Mekhitarists, an order of Armenian Catholic monks, have centres in Venice and in Vienna. The monks have performed noteworthy service

Persecution of Armenian Christians

in the preservation of ancient Armenian literature and in the field of publications. The liturgy continues to be practiced in the classical Armenian, a testimony to the endurance of Armenian Christians, who in 1965 marked the 50th anniversary of their most recent persecution under the Turks (approximately 1,500,000 Armenians, including Armenian Apostolic Orthodox, Armenian Catholics, and Armenian Protestants, were killed because they would not forfeit their faith). Approximately 60,000 Armenian Catholics are under the jurisdiction of the patriarchate, and 40,000 outside.

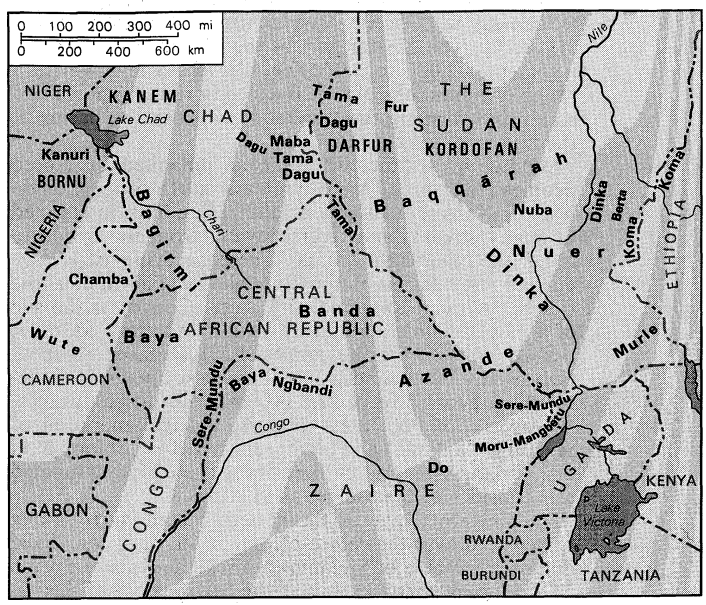
The Eastern rite churches today. The majority of Eastern rite churches find themselves today as did their forefathers, from the 7th century onward, within an Islāmic environment, though there also exist Marxist and Hindu contexts for Indian Christians and the complex forces of secularism that confront all the rites. Furthermore, Eastern Catholics, except for the Maronites, as well as Protestants, are a "Christian minority" among Christians of the independent churches not in union with Constantinople. It is this minority status that presents the greatest challenge to Eastern Catholics as they seek to find a creative religious witness in their world. Eastern Catholics generally have an outstanding reputation among other Christians and Muslims, despite their small numbers, for establishing an orderly community of followers. Accused by many non-Eastern Catholics of being too organized and latinized by Rome, they are also held in suspicion by Muslims as foreign agents of the West. Thus, the conditions for Eastern Catholics today is not easy. The most distinctive contribution of Eastern Catholics, especially within Islāmic areas, is made through their educational institutions, which look toward a new era of dialogue and coexistence with their Muslim neighbours.

BIBLIOGRAPHY. AZIZ S. ATIYA, *History of Eastern Christianity* (1968), a study limited to Eastern, non-Byzantine churches; DONALD ATTWATER, *The Christian Churches of the East*, 2nd rev. ed., 2 vol. (1961–62), an overall survey of all Eastern Churches, however, statistical material is dated; *Eastern Catholic Worship* (1945), contains the text of Eastern Church rites, with commentary; F.E. BRIGHTMAN (ed.), *Liturgies, Eastern and Western* (1896, reprinted 1965), a classic study of the liturgies; CARNEGIE SAMUEL CALIAN, *Icon and Pulpit* (1968), an ecumenical study of the scope and diversity of Eastern Christians for a Western audience; IRENEE-HENRI DALMAIS, *Les Liturgies d'Orient* (1959; Eng trans., *Eastern Liturgies*, 1960), a commentary and discussion of Eastern rites from a Roman Catholic perspective; FRANCIS J. MCGARRIGLE *et al.*, *The Eastern Branches of the Catholic Church* (1938), includes six studies on the Oriental rites compiled by the Liturgical Arts Society; the essays on "The Code of Canon Law and the Catholics of Eastern Rites" and "Liturgy and Asceticism in the Eastern Church" are particularly helpful; ARCHDALE A. KING, *The Rites of Eastern Christendom*, 2 vol. (1947–48), a complete and detailed study of the respective Eastern rite churches; SEVERIEN SALAVILLE, *An Introduction to the Study of Eastern Liturgies* (1938), a detailed study of Eastern liturgies indicating types, relations, and history; MAXIMOS IV (ed.), *Voix de l'église en Orient* (1962; Eng. trans., *The Eastern Churches and Catholic Unity*, 1963), highlights the search for identity on the part of Eastern Catholics within a Western church; TIMOTHY WARE, *The Orthodox Church* (1963), a clearly written history of the Orthodox Church by a convert to Orthodoxy from Anglicanism.

(C.S.Ca.)

Eastern Sudan, Cultures of the

The name Sudan (Country of the Blacks) was given by medieval Arab writers to the savanna belt that extends across the continent of Africa from Cape Verde on the Atlantic to the rise of the highlands of Ethiopia. Topographically, this Sudanic belt is a plateau averaging somewhat less than 1,500 feet (450 metres) in altitude, with a length of over 3,000 miles (4,800 kilometres) and an average width of about 650 miles (1,000 kilometres). This belt of open country, with its clearly marked dry season of intense heat and its wet season of growth, is bounded along most of its southern fringe by thick forests and along its northern edge by desert. The area from around Lake Chad eastward, traditionally designated as the eastern Sudan, is now shared between the republics of The Sudan, Zaire (Congo [Kinshasa]), Chad, Cameroon, Nigeria, and the Central African Republic. The societies within eastern Sudan portray diverse linguistic, ethnic, social, cultural, and religious characteristics.



Distribution of the peoples of the eastern Sudan.

PEOPLES AND LANGUAGES

Because of the great cultural diversity, it is difficult to classify the types of cultures found among the various peoples in the eastern Sudan. The selection of cultures that follows merely suggests some rather prominent cultural patterns, illustrative of the wide range present.

The cattle-herding Baqqārah

Bedouin Arabs invading southward from Arabia and North Africa reached the area of Lake Chad as early as the 16th century; gradually they intermarried extensively with the Negroid inhabitants, so that many tribes came to exhibit Negroid characteristics. Toward the end of the 18th century, the Baqqārah (nomadic cattlemen), who claim Arab descent, moved eastward to the central regions of what is now the Democratic Republic of The Sudan. This part of the savanna belt is flat and tree covered and suitable to an economy of nomadic cattle husbandry. The majority of the peoples living there have taken not only to raising cattle but also to cultivating millet, sorghum, cotton, and other traditional crops. All are Muslim. Some of the major Baqqārah tribes are the Humr, who number about 55,000, and the Messiria, Hawazma, and Rizeigat.

In the Nuba Hills, a rugged, almost mountainous area in a region traditionally known as Kordofan and now a province of the Democratic Republic of The Sudan, there live a Negro people called the Nuba, who, because of the terrain, have tended to be isolated from the peoples migrating about them in the Nile Valley and other areas of the eastern Sudan. Numbering about 500,000, the Nuba are subdivided into over 50 different ethnic groups with a multiplicity of languages. Despite their ethnic and linguistic diversity, there exits what has been called a "Nuba culture" based upon some affinities such as clan structure and the regulation of intertribal conflicts. Very little is known about the ancient history of the Nuba tribes, but there are oral traditions that they migrated from the west. The country of the Nuba is well watered, densely wooded, and rich in vegetation. The Otoro is the largest tribe (about 40,000 people), inhabiting eight hills in the eastern part of Nuba country.

Between Kordofan and Lake Chad, along the northern fringes of the eastern Sudan, there are a number of Negro peoples who were among the first sub-Saharan inhabitants to be strongly influenced by the Arabian commerce and culture penetrating from the north (from about the 11th century onward). Situated on the African frontier of the

trans-Saharan caravan routes, these peoples were in a position to develop more diversified cultures and elaborate economies and social structures. Representative of these populations are the groups residing in Darfur, the Republic of The Sudan's westernmost province—the largest resident group being the Fur, after whom the province is named, who occupy the centre and number about 1,300,000. The greater part of the country of the Fur is a plateau—2,000 to 3,000 feet above sea level—with a range of mountains of volcanic origin running north and south and forming the watershed between the basins of the Nile and Lake Chad.

Another important cultural zone inhabited by Negro peoples is the broad band lying just north of the Congolese rain forest, in an area of undulating hills, intersected by many streams and rivers flowing from the divide to either the Nile or the Congo. The zone, lying astride the current boundaries of the Republic of The Sudan, the Central African Republic, and Zaire (Congo), is the site of typical Sudanic agriculture—with sorghum and millet usually constituting the staples. Representative of these peoples are the Azande, one of the best known Sudanic-speaking peoples of central Africa, who probably number about 750,000; they comprise an agglomeration of invading Sudanic and indigenous tribes differing in origin, language, and culture but thoroughly assimilated and considering themselves all as Azande.

The western fringe of the eastern Sudan, surrounding Lake Chad, was the centre of one of the most long-lived and most notable of Sudanic empires, the Kanem-Bornu. The ruling dynasty, known as the Sef, was established as early as the 8th century in Kanem (now a province of Chad) and removed later to Bornu (now a province of Nigeria) and lasted until 1846—making it perhaps the longest reigning dynasty in world history. The empire flourished from the 11th century on, when it developed as the Sudanese terminus of the important trans-Saharan trade route running to Libya; most of the peoples of the empire became Muslims at that time. Today the dominant ethnic group in the area is that of the Kanuri of Bornu, a tall Negro people who number about 1,250,000 and inhabit open savanna country. Their current economy is based on millet agriculture, but they remain a commercial people with a well-developed internal and external trade in various crafts, cowhides, and other goods.

Major language groups

The major language groups in the eastern Sudan are Congo-Kordofanian, to which the Azande and Nuba languages belong, and Nilo-Saharan, to which the languages of the Kanuri and Fur belong; the term Sudanic languages is applied to these and other languages spoken by peoples from Ethiopia to Nigeria. The Baqqārah tribes are entirely Arabized, and most of the Fur, too, now speak Arabic. Even among the Nuba, Arabic is becoming the lingua franca.

SOCIAL ORGANIZATION

Political and territorial organization. Broadly speaking, the societies of the eastern Sudan exhibit two types of political organization: the hierarchical systems of the Azande, Kanuri, and Fur; and the segmentary systems of the Humr Baqqārah and Otoro. Zandeland, for example, is divided into a number of autonomous chiefdoms. The structure of authority within each chiefdom is pyramidical, with chiefs at the apex of the hierarchy (previously kings), followed by sub-chiefs, deputies, and finally homestead heads. The homestead is the basic political, social, and economic unit. Absolute authority lies with the aristocratic Avungara clan, and commoners can only reach the positions of deputy or homestead head. Local groupings, in which descent is reckoned in the male line, are not associated with clans or sections of clans; they are political and administrative units. Clans have little corporate life, and genealogical links between clansmen are seldom known.

The Kanuri also have a hierarchical administrative and political structure. The household or compound is the basic political unit; a number of households join to form wards, hamlets, villages, and districts. At the head of this hierarchy is the *shehu*, who is the formal head of the government and the religious leader of the state. With the exceptions of the *shehu* and heads of households, heads of wards, hamlets, and districts are appointed. As among the Azande, members of agnatic descent groups (descent groups traced through the male line) are dispersed, and political units do not coincide with kinship groupings.

A similar political and administrative structure of the centralized type also existed among the Fur. There was a sultan at the head of the state, which was divided into four regions in turn divided into districts, subdistricts, and villages. Each village had a council of elders who decided minor cultivation disputes and enforced their decisions by advice and warning. The rights of the village were vested in its inhabitants jointly. Ultimate authority lay with the sultan, who could depose officials beneath him when their power became a threat to his dominance.

In contrast, the Humr Baqqārah have a political system based on a segmentary lineage organization. The tribe is divided into two sections, each of which is divided into a further five sections, each comprising major lineages, camps, and extended families. All these groups have potential leaders. The significant residential and herding unit is the camp, the composition of which changes with the seasons. Within each tribe, and at every level, there is a process of splitting, migration, and resettling that results in a continuous change of alliance among groups and individuals. Blood feuds occur between segments and are settled by payment of blood money. Power among the Humr Baqqārah stems from wealth and strength of personality.

Political organization of the Otoro

The Otoro political system consists of a number of territorial segments that do not coincide with kinship groupings. Clan members are scattered in different localities; the basic political unit is the hill community, whose members share a tract of land and a common code of morality. Feuding between hill communities has been constant, but members of the same hill community must not kill one another. The Otoro have recognized tribal boundaries defined by periodically renewed intertribal treaties. Because of the continual raids and wars between segments, a chieftainship, with the power to use force to maintain peace, has been established.

Family and kinship patterns. In all five societies descent is reckoned in the male line, but the significance of agnatic ties among kin groups differs from one society to another. Among the Azande there are 188 exogamous clans that function primarily as political and administrative units. Clan members are expected to join together to celebrate births, marriages, and funerals. The Azande recognize obligations to the kin of both parents, but greater importance attaches to the paternal connection. Until he is married an Azande has not achieved a man's full status. Marriage is a series of events, each of which is marked by the transfer of goods from one group of kin to the other. Wives of a dead man are inherited by his sons or brothers. Polygyny is also practiced and is regarded as a means of extending affinal relationships and acquiring support. Although divorce is now common, a broken marriage was considered a shameful thing because it destroyed the network of relationships.

As among the Azande, Kanuri descent groups are based upon the individual's relations to the kin of his parents, with emphasis on the patrilineal line. There are two types of relationship that are significant in kinship, economic, and political spheres: discipline–respect relationships (modelled on the father–son relationships) and shame–avoidance relationships demanding subdued and self-controlled behaviour (*e.g.*, a junior must not speak until he is spoken to by his senior, nor can they eat together from the same bowl). Kanuri practice cousin marriage with virilocal residence (*i.e.*, residence with or near the husband's parents), and one in two marriages is polygynous. A person depends on his kinsmen, friends, and political and economic patrons for help over his marriage. Divorce is easy, and the rate is high: 80–90 percent.

The Fur also reckon descent patrilineally, but residence is customarily uxorilocal (with or near the wife's parents). If a husband disagrees with his affines, however, he may take his wife to live with his own group. Cousin

marriage, sororate (customary marriage with a deceased wife's sister), levirate (customary marriage to an elder brother's widow), and polygyny are practiced.

Lineage organization of the Humr Baqqārah

Among the Humr Baqqārah, members of the smallest lineage (*surra*), together with their dependents, form a single camp. The organization of a *surra* depends on the number of cattle and the distribution of their ownership among its members. Each *surra* has a leader who is wealthy but who has no administrative functions unless he is already a member of the local government. Whereas members of a camp form the basic unit of cooperation over herding, the household is the main unit of cooperation in agriculture, although some activities connected with agriculture involve a wider group. Preferred marriage within the *surra* is with a parallel cousin. Many first marriages, contracted to conform to the expectation of elders, end in divorce, but in subsequent marriages people are free to choose their marriage partners.

Among the Otoro there are 35 patrilineal clans of various size. With the exception of the Chungur clan, which is the traditional holder of the hereditary chieftainship, all clans are socially of equal order. Clan members intermarry with each other, although clan exogamy was the rule in the past. In the economic and religious spheres the clan does not exercise influence, but it does impose upon its members the collective duty of blood feud in the case of homicide between clans. Patrilineal descent is important in determining the social identity of a person, inheritance, and rights and duties concerning marriage and bridewealth (gifts from the groom and his kinsmen to the father of the bride and his kinsmen). Though the Otoro are patrilineal, matrilineal ties are also important. Polygyny and leviratic marriages are practiced, and bridewealth payments establish wider contact between social groups. Divorce is negotiated and settled by the families concerned. The marriage is dissolved by the total refund of the bridewealth to the former husband.

Social stratification. All the groups have some form of class distinction. The Azande's Avongara clan recognizes rigid distinctions between chiefs and commoners, and generally among the Azande there are distinctions between conquering and conquered groups and slaves. Political, economic, and social status depend on birth, but the politically superior group, the Avongara, must maintain their position by organized political and military means.

Kanuri society is also hierarchical, the Shehu being the highest political and religious authority and being most prominent in trade and other economic activities. There are also slaves. The introduction of western-type education and native governments produced new classes with non-traditional premises and aspirations.

The highest political office among the Fur has been that of the sultan, who has been surrounded by a body of councillors dependent on royal patronage. Locally, there has been a descending hierarchy of hereditary fief stewards, village heads, and, lastly, heads of households. Ironworkers have been a despised class not allowed to intermarry with ordinary Fur. Among the Humr Baqqārah, the main distinction is between persons of Arab descent (who hold the positions of power) and non-Arabs. In other respects, a person's wealth and personality determine his successful bid for leadership.

Traditionally, among the Otoro there were no political offices, but only a special hereditary ritual office—that of "chief of the Path," who acted as intermediary in peace negotiations between conflicting parties. Today, however, there are chiefs selected by the local government from among persons of local wealth and importance. In addition there are age-grades; membership in each grade (there are five grades) lasts three years, after which all the members are promoted together to the next higher grade. Promotion is marked with festivities, and members of each age-grade live separately. This system marks the development of a person from an ordinary youth to that of a "big man." Girls also form age-grade groups.

Socialization and education. In all five societies under comparison, there are ritual and ceremonial practices marking the stages in the life cycle of the individual—birth, circumcision, puberty, marriage, and death. There is little formal education among the Azande, Otoro, Humr Baqqārah, and Fur.

Circumcision distinguishes boyhood from manhood. It is a recent introduction among the Azande, having spread from the south despite being opposed by the Avongara. Among the Kanuri, no mention of the newly born baby is made, and even its sex is unknown outside the household except for very close friends of the family. Among the Otoro, an individual achieves higher status as he is promoted from one age-grade to another. There is also an association of cicatrization (scarring) with a test of manhood: killing an enemy entitles a person to have a small pattern of scars on his back. Similar patterns are also made on the upper arms of successful hunters. Men take part in wrestling and fighting as part of their training for manhood. Among the Azande and Nuba, there are very few schools at elementary level established by missions. Education has made little headway among the Humr Baqqārah because it comes into conflict with their way of life—people are reluctant to send their children to schools because livestock and cattle herding will suffer. There are a few Qur'ānic schools among the Fur, in which elementary Arabic, arithmetic, and the Qur'ān are taught. The Kanuri also have Qur'ānic schools and are the most advanced in terms of formal education. They also have an advanced system of government and administration, which came into operation during the colonial period at the beginning of the 20th century.

Qur'ānic schools of the Fur and Kanuri

ECONOMIC ORGANIZATION

Settlement patterns. The five societies selected for comparison exhibit three different patterns of settlement. The Otoro and Azanda live in scattered groups. Otoro homesteads are scattered irregularly over hilltops and valleys, a number of homesteads constituting a village, villages combining to form hill communities, a number of which make up the tribe. Similarly, the Azande live in family groups in scattered homesteads separated by strips of bush. Among the Otoro and the Azande, the homesteads are the basic social and economic unit. In contrast, the Kanuri and Fur have more compact settlements. The Kanuri settlements, composed of walled compounds, vary in size from Maiduguri, the capital, with over 100,000 people, down to small hamlets of three to four households. The Fur live in homesteads, a number of which constitute a village. The third type of settlement pattern is that of the Humr Baqqārah, who live in tent camps. Their nomadic existence does not permit the formation of permanent settlements.

Production and technology. The economy of most of the peoples of eastern Sudan is dependent on cultivation, with animal husbandry, and sometimes hunting and fishing, providing an important supplement to agriculture. A wide variety of crops, including grain, sesame, vegetables, sorghum, maize, groundnuts, and cotton, are grown, the latter two sometimes as cash crops. Among the Humr Baqqārah cultivation is a subsidiary occupation. They move in a regular seasonal cycle according to the availability of water and grass. In winter, or at the end of the rains, they move south, and in spring, or during the rains, they move north. Some cultivation of millet and, in recent years, of cotton as a cash crop, is undertaken. Throughout their territory the Humr Baqqārah have communal grazing rights, whereas cultivable plots of land are owned individually and handed down from father to son.

The technology of these societies is simple: among the Otoro the making of bedsteads, mat weaving, and pottery are undertaken. The Azande, on the other hand, are prominent as craftsmen and artists. Their superior material culture, particularly their knives, spears, and shields, was one of the factors by which they dominated their neighbours and brought about the spread of their culture. Basketmaking, net weaving, pottery, smelting, metalworking, and ivory and wood carving are also undertaken. Among the Kanuri weaving, metalworking, dyeing, tanning, smithing, and other occupations are organized under headmen in each settlement. Each occupation

Prominence of Azande arts and crafts

comes under an overall headman who lives in the capital city. There are ironworkers in every Fur village; other Fur crafts include tanning and weaving of cloth and basketry. The Humr Baqqārah produce leather goods and basketry to meet their own needs.

Property and exchange systems. Land among the Otoro, Azande, Kanuri, and Fur is the principal economic asset. Among the Otoro land can be acquired through inheritance, purchase, lease, and by clearing and cultivating new areas. The Otoro and the rest of the Nuba tribes convert nearly their whole agricultural surplus into livestock. Although livestock are used for clan sacrifices, they are not sold for money, and neither are they slaughtered for meat. Among the Nuba tribes, however, people use livestock for certain standardized payments such as bridewealth or gifts to kin relations. Iron, because of its rarity and its use in weapons and tools, became the standard medium of exchange. Handicraft products are paid for in grain, sesame, hoes, spears, goats, axheads, and, nowadays, in money.

Whereas livestock is a secondary source of wealth to the Otoro, cattle are the primary source of wealth, prestige, and political position for the Humr Baqqārah. Cattle were once the medium of exchange, but during the 20th century cash has become significant.

Before the introduction of cotton and money, wealth among the Azande was primarily in the form of perishable agricultural produce, and it was customary to destroy the property of a man after his death. Only chiefs were able to accumulate wealth, since they received tribute and prisoners of war as perquisites. There was little exchange between households, although iron tools and spears were used in bridewealth payments.

Among the Fur, property consists of houses, domestic articles, rolls of cloth, and cattle kept mainly for resale. Rights over land are held jointly by descent groups and are vested in a titleholder. There are organized marketplaces in which the medium of exchange has been, and still is, money. Previously, cloth may have served as a medium of exchange. Agricultural products are exchanged for tools and utensils, cloth, and other commodities. Labour is exchanged for beer.

The Kanuri have organized marketplaces in which agricultural produce and craft products are sold. Money is thus the medium of exchange and constitutes the main form of wealth among the Kanuri. Land is an important economic asset, but nowadays wealth and prestige lie in success in trade, in politics, in the civil service, and in the acquisition of Western educaton.

RELIGIOUS PRACTICES

The Humr Baqqārah, Fur, and Kanuri peoples adhere to Islāmic beliefs and practices, Islām having come to them through Arab influence. Traditional local practices, however, coexist with Islāmic beliefs. Among the Fur, for example, the splashing of sanctuaries with a flour and water paste is carried out to ensure fertility. There are also rain cults thought to have been introduced from farther west. Sacrifices are made at shrines and at ancestral tombs when the rains are likely to fall. The office of rainmaker is hereditary.

Witchcraft of the Otoro and Azande

The Otoro and Azande have their own local beliefs and practices, which are significant as a means of social control, and Islām and Christianity have very little influence. Among the Otoro there is a widespread belief in oracles and witchcraft as a means of punishing offenders and establishing justice. Charms bought from Arab or West African charm sellers, diviners, grain priests, and rainmakers are used to find and punish evildoers. Leprosy, imagined to be caused by supernatural powers, is believed to be connected with offenses such as sexual intercourse in forbidden kinship degrees and homicide. Witchcraft, a magic at the disposal of any individual, is effective only if directed against a person guilty of a crime.

Among the Azande the power of witchcraft is inherited. Diviners, oracles, vengeance magic, and the use of leeches are means of counteracting it. The Avongara clan may not be accused of witchcraft because the chief and his oracle are considered infallible.

DEVELOPMENTS IN THE 20TH CENTURY

The influence of European contact and of self-government on the societies of the eastern Sudan has varied. Innovations introduced among the Humr Baqqārah have included cotton cultivation, use of cash, wage labour, and improvements in breeding. With the introduction of local government, the traditional power structure was altered, but the traditional agnatic groupings remain as units of administration. The economic and social importance of cattle has not changed.

Although the traditional economic organization of the Kanuri persists, more people are seeking individual advancement in trade, administration, and politics and by means of Western education. The household, however, remains the economic and social unit, and the superior–subordinate relations within it have continued as a basis for maintaining the traditional society and its values. The traditional power structure of the Kanuri overlaps with their modern political organization based on the emergence of political parties.

The Otoro have experienced a number of changes introduced by Europeans and Arabs. Larger tracts of land are being worked, new crops have been introduced, modern chieftainship has been instituted, and a new system of law and order has been introduced. The younger generation of Nuba who work outside their homeland bring back new habits and interests, and the authority of the older generation is thus weakened. Nevertheless, the Nuba still retain their social structure and identity.

Since the incorporation of Darfur into The Sudan at the turn of this century, a new administration has been introduced. The independence of the Fur and other tribes has been modified by the formation of regional local governments, each with its own judicial and financial powers. The basic tribal local government, however, is still founded on the rule of hereditary or traditional headmanship of a tribe or on sultanates.

Attempted resettlement of the Azande

The most noticeable social and economic changes have occurred among the Azande, who since the 1920s have been constantly settled and resettled. In 1945 a resettlement scheme was introduced based upon the manufacture of cloth, cooking oil, and soap from cotton grown by the Azande. It was at first successful, but eventually proved unacceptable to the Azande themselves since it was not sufficiently in accord with their value system. Sporadic revolt in the area has also delayed development.

BIBLIOGRAPHY. E.E. EVANS-PRITCHARD, *Witchcraft, Oracles and Magic Among the Azande* (1937, reprinted 1958) and *The Azande: History and Political Institutions* (1971), are the best authoritative accounts; while CONRAD C. REINING, *The Zande Scheme* (1966), deals with recent changes among the Azande due to resettlement. P.T.W. BAXTER and AUDREY BUTT, *The Azande, and Related Peoples of the Anglo-Egyptian Sudan and Belgian Congo* (1953), offers a general ethnographic account. S.F. NADEL, *The Nuba* (1947), remains the only anthropological study of the Nuba tribes. A.J. ARKELL, "The History of Darfur, 1200–1700 A.D.," *Sudan Notes and Records*, 32:37–70, 207–238 (1951) and 33:129–155, 244–275 (1952); A.C. BEATON, "The Fur," *ibid.*, 29:1–39 (1948); and G.D. LAMPEN, "History of Darfur," *ibid.*, 31:177–209 (1950), are useful historical accounts. H.G. BALFOUR-PAUL, "A Prehistoric Cult Still Practised in Muslim Darfur," *Jl. R. Anthrop. Inst.*, 86:77–86 (1956); and A.C. BEATON and R.C. COOKE, "Bari and Fur Rain Cults and Ceremonies," *Sudan Notes and Records*, 22:181–203 (1939), are of ethnographic interest. FREDRIK BARTH, "Economic Spheres in Darfur," in RAYMOND FIRTH (ed.), *Themes in Economic Anthropology* (1967), is a recent account that gives an interesting analysis of aspects of economic organization among the Fur. IAN G. CUNNISON, *Baggara Arabs* (1966), is the best anthropological study of lineage organization and politics among the Baqqārah. C.G. and B.Z. SELIGMAN, *Pagan Tribes of the Nilotic Sudan* (1932); and H.A. MACMICHAEL, *A History of the Arabs in The Sudan and Some Account of the People Who Preceded Them and of the Tribes Inhabiting Dárfūr*, 2 vol. (1922, reprinted 1967), and *The Tribes of Northern and Central Kordofán* (1912, reprinted 1967), are classics that provide further ethnographic and historical data. J.D. TOTHILL (ed.), *Agriculture in The Sudan* (1948), remains the only detailed authority on agriculture in The Sudan. RONALD COHEN, *The Kanuri of Bornu* (1967), is a recent anthropological study of the Kanuri; and S.J. HOGBEN and A.H.M. KIRK-GREENE, *The Emirates of North-*

ern Nigeria (1966), offers a well-documented historical study. C.K. MEEK, *The Northern Tribes of Nigeria*, 2 vol. (1925, reprinted 1969), provides comprehensive data on the cultures of northern Nigeria. A.H.M. KIRK-GREENE, "The Linguistic Statistics of Northern Nigeria: A Tentative Presentation," *African Language Review*, 6:75–101 (1967), provides recent statistics on the languages and population of northern Nigeria. JOSEPH H. GREENBERG, *The Languages of Africa* (1963); and A.N. TUCKER, *The Eastern Sudanic Languages* (1940), provide detailed linguistic classifications and data on the eastern Sudan. Important travel books are HEINRICH BARTH, *Travels and Discoveries in North and Central Africa, . . . 1849–55*, 5 vol. (1857–58; reprinted in 3 vol., 1965); M.O. AL-TUNUSI, *Voyage au Darfour* (1845); IGNAZ PALLME, *Reitebeschreibung von Kordofan* (1843; Eng. trans., *Travels in Kordofan*, 1844); DIXON DENHAM, HUGH CLAPPERTON, and WALTER OUDNEY, *Narrative of Travels and Discoveries in Northern and Central Africa, in the Years 1822, 1823*, and *1824* (1826); H.R. PALMER, *Sudanese Memoirs*, 3 vol. (1928); WILHELM JUNKER, *Reisen in Afrika 1875–1886*, 3 vol. (1889–91; Eng. trans., *Travels in Africa During the Years 1875–1886*, 3 vol., 1890–92); JOHN and KATE H. PETHERICK, *Travels in Central Africa, and Explorations of the Western Nile Tributaries*, 2 vol. (1869); and GUSTAV NACHTIGAL, *Sahārā und Sûdân*, 3 vol. (1879–89).

(A.Al-Sh.)

Eastern Woodlands Indians

At the time of the European discovery of North America, the area stretching from Lake Superior to the Atlantic coast was occupied by many different Indian groups speaking languages that belong to three distinct language families: Iroquoian, Algonkian, and Siouan. The major speakers of northern Iroquoian languages include the Iroquois confederacy (Mohawk, Oneida, Onondaga, Cayuga, Seneca, and, later, Tuscarora), the Huron confederacy, Tionontati (Petun or Tobacco Nation), Wyandot (a group of Hurons and Tionontati), Neutral, Wenrohronon, Erie, Susquehanna (Conestoga), and Laurentian Iroquois. The major speakers of Algonkian languages include the Passamaquoddy, Malecite, Abnaki, Penobscot, Pennacook, Massachuset, Nauset, Wampanoag, Narraganset, Niantic, Pequot, Mohegan, Nipmuc, Pocomtuc, Mahican, Wappinger, Montauk, Delaware (Lenni Lenape), Powhatan, Ojibwa (Chippewa), Menominee, Sauk (Sac), Kickapoo, Miami, and Illinois. The Winnebago, who also lived in the eastern woodlands, spoke a Siouan language.

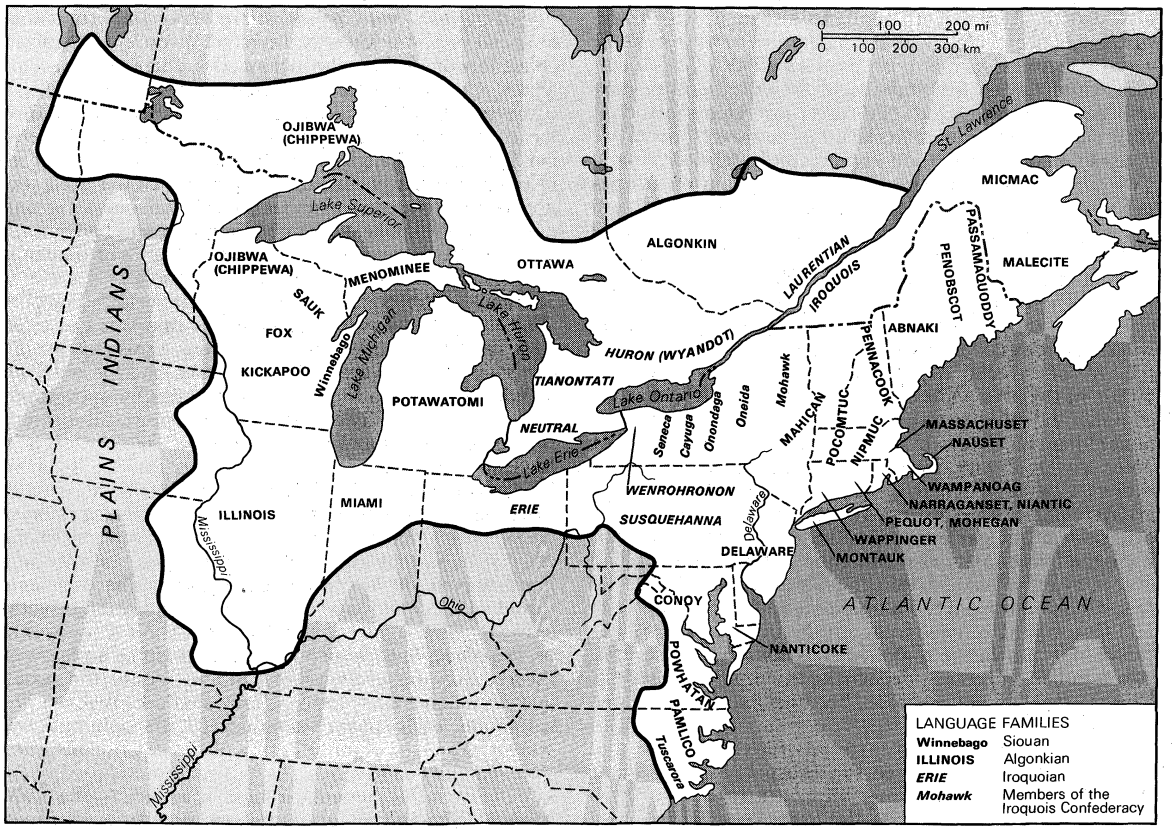
Forest resources

Although the area generally was a wooded one, it was in the open spaces in the forests that many of the best natural resources were to be found. In consequence, the heaviest population concentrations were near or along the seacoast, lakes, ponds, marshes, creeks, and rivers. There animals could be hunted, fish caught, birds taken, leaves, seeds, and roots of wild plants gathered, shellfish collected, and crops grown. Certain areas were favoured with resources not found elsewhere in the region. In certain parts of the upper Great Lakes area, wild rice (*Zizania aquatica*) grew in abundance, and the Menominee especially depended on it. Buffalo (bison) roamed the plains–prairie area, and such groups as the Sauk, Fox, Illinois, and Miami, who lived near the prairie, hunted them. On the Atlantic Coast, shellfish were plentiful and played an important part in the diet.

TRADITIONAL CULTURE PATTERNS

Economic systems. *Food production and consumption.* Except in the northern part of the upper Great Lakes area and the northern part of the Eastern Seaboard—areas that do not have a growing season long enough for corn (maize) to mature—some agriculture was practiced. The three principal plants cultivated were corn, beans, and squash, plants the Iroquois termed the three sisters because they were to be found like sisters together in the fields. Generally, it was the women who cultivated the fields, planting the seed, weeding, and harvesting the crops. Men, however, usually aided in clearing the land preparatory to planting. The type of agriculture practiced is often termed horticulture, as domesticated draft animals were not used. The implements used in planting were the hoe and the digging stick (dibble stick), a wooden stick pointed at one end that was used to make a hole in the ground into which the seeds could be dropped and then covered.

Adapted from *International Journal of American Linguistics* (1953)



Distribution of Eastern Woodlands Indians.

Of all the groups in the area, the Iroquoian tribes were probably the most dependent on agriculture, but even they did not rely wholly on it. The population of the area was never great enough to have exhausted the supply of wild plants and animals, and as a result the Indians living there did not have to expend the time and effort necessary for intensive agriculture. They also could avoid a dull and monotonous, though balanced, diet confined to corn, beans, and squash and could enjoy a rich and varied diet merely by travelling to the proper places at the proper times of the year to collect food. Leaves of wild plants were gathered in the spring and cooked as greens. Later in the year, berries and other seeds were gathered. Fish were caught in the spring, summer, and fall—different species being plentiful at different times of the year. Birds were taken, and in certain areas the spring roosting places of pigeon provided virtually unlimited amounts of squab. In early spring, the maple trees were tapped for their sap. Fall was the best time for hunting, and numbers of families went into the forests at this season of the year to hunt. The seasonal round of activities was, then, a varied one and one that, contrary to common opinion, afforded a great deal of leisure time, for the Indians did not wander aimlessly around, barely avoiding starvation by collecting what few eatable foods were available. They relied, instead, on their knowledge of the country to provide them with a diverse diet and diverse experiences in obtaining it.

Some cooking was accomplished by roasting over the fire or baking in the ashes, but the most popular method was boiling. The typical meal included soup cooked in a pottery or bark vessel. Dishes, however, were made not of pottery but of bark or of carved wood, as were the spoons used in eating. Each individual had his own bowl and spoon, and so an invitation to a feast was likely to include the invitation to "Come, and bring your bowl and spoon." Corn was generally prepared by pounding it with a wooden pestle in a wooden mortar hollowed out of a tree trunk. Before pounding, the corn was usually soaked in ashes to make removal of the hull easier. Occasionally, however, the corn was ground between two flat stones.

Housing, transportation, and weapons. As the land provided the Indian with food, so also it provided him with other necessities of life. The culture of 19th-century white American farmers has been labelled a "wood culture," and the earlier Indian cultures of the area may also be so characterized: the forest provided the Indian with many of the raw materials that he used to make his houses (as well as the food implements described above).

Wigwams and longhouses

Two types of houses were built: the wigwam by the upper Great Lakes and Eastern Seaboard Algonkians and the longhouse by the Iroquoians. The wigwam was made by driving a number of pointed poles into the ground to make a circular or oval floor plan. These poles were tied together with strips of bark and reinforced with other poles tied horizontally to make a dome-shaped framework that was covered with bark, reeds, or woven mats, the type of covering depending on the availability of materials in the area. A fire in the centre provided heat for cooking and for warmth. Although the Iroquoian longhouse, like the wigwam, was made of a framework of poles covered with bark sheets, it was arbor-shaped rather than dome-shaped, having straight sides and an arched roof. After European contact, the roof shape often was of the gable type. Rather than being circular or oval, the longhouse was rectangular in floor plan and had a door at either end. A longhouse was approximately 20 feet wide and ranged from less than 50 feet to more than 200 feet in length, depending on the number of families living in it. Down the middle of the house were the fires, two families on either side sharing the fire in the middle. The average longhouse probably had five fires and ten families.

The forest also provided materials for the dugouts and bark canoes commonly used for transport in this area of many lakes and streams, the birchbark of the northern portions of the area making the best canoes. The forest also provided materials for the frames of the snowshoes, which made travel in the winter easier and which were essential in the north (a man on snowshoes could outrun a moose or other large animal in the snow and take him with an ease that was not possible in more clement seasons). The bows and arrows and the spears used in hunting and warring were also made of wood provided by the forest. Points for the arrows and spears were chipped from flint or other suitable stone, as were the knives and other sharp-edged implements.

Clothing. Animals of the forest provided the skins from which the women made clothing. The basic item of men's dress was the breechclout, a strip of skin drawn between the legs and held in place by looping it over a belt at the waist. For protection from the cold or while travelling in the forest, leggings (basically, two tubes of skin also attached to the waist belt) were added. A cape or robe of skin or fur was also worn in cold weather. The basic item of women's dress was a skirt, to which might be added leggings tied at the knee and a cape or robe. Both men and women wore moccasins, the soft-soled and heelless shoe adapted, among other things, for use with the snowshoe. The clothing of both sexes might be decorated with painting, porcupine-quill embroidery, shells, or shell beads. For special occasions, such as feasts and war expeditions, the body might also be decorated with paint and jewelry. (The face might also be tattooed.) Long hair was admired and might be greased to give added lustre. A number of men plucked out some hair and cut the remainder to form roaches or other distinctive styles.

Social organization. The relatively small numbers of Indians who lived in the area and their particular manner of obtaining the essentials of life—food, clothing, and shelter—did not require an elaborate social organization to coordinate their activities. Rather, it was a way of life that tended to foster individual self-reliance. Division of labour was simple: women took care of the children, did the cooking and other household tasks, made the clothing and basketry containers, and did much of the agricultural work; and the men hunted, fished, traded, warred, and made the houses and the implements that they used. Beyond this division of labour by sex, there was little specialization.

Bands, tribes, and confederations. An important unit of social organization was the band, a loosely organized collection of people who occupied a particular tract of land and who recognized a common identity, including a name for themselves. A number of names now known locally as Indian tribal names were band names. Several bands comprised a tribe, which, like the band, was loosely organized and which in many parts of the area was not so much a political unit as a cultural one—a group of people who spoke a common language and had similar customs. (It is the names of tribes rather than of bands that are listed at the beginning of this article.)

Leaders and councils

Although chieftainships often were inherited, personal ability was the basis for the influence that was exercised by a chief (sometimes termed sachem, especially in the eastern part of this area). The man who had the requisite abilities was chosen to succeed. Particularly important to a chief was his ability to persuade. As one result, oratory was highly valued and developed into a fine art; even in English, the power of Indian oratory is evident. Typically, the councils of the Indians involved the making of speeches, but the intent of this oratory was not to impress others with mere rhetoric but to find a solution to the issue at hand that all could agree to. If unanimity was not achieved, no action could be taken. The dissidents would either continue to express their opposition or withdraw; in either case, the effectiveness of the group would be weakened. Speech making, then, served as a means of ascertaining the diversity of opinion within the group and the manner in which consensus could be reached, for commonly each speaker summarized the opinions previously expressed before offering his own. Councils often were lengthy affairs.

Councils (including some religious councils—the Indians typically did not draw a sharp distinction between sacred and secular affairs) normally opened with the

smoking of a pipe. In the Midwest, this practice was elaborated into the calumet ceremony used to greet foreigners and on other occasions. It is from this custom that the phrase "smoke the peace pipe" is derived.

League of the Iroquois

The most elaborate and powerful political organization developed in this area was that of the League of the Iroquois, but even this league rested on the political principles to be found elsewhere. Unanimity was required (the failure to achieve this unanimity being one of the reasons for the Iroquois defeat at the hands of the Revolutionists during the Revolutionary War), and oratory was highly regarded. Tradition has it that the league was formed as a result of the efforts of Dekanawida and Hiawatha probably sometime in the 15th or 16th century. The intent of the league was to establish peace among the five tribes of the Iroquois (the sixth member of the league, the Tuscarora, was to join it in the early part of the 18th century); and thus one of the most important parts of the constitution of the league established the price for murder: ten strings of wampum (shell beads) for the life of a man and 20 strings of wampum for the life of a woman, to be paid to the kinsmen of the murdered person. (If a man murdered another man, the total paid by the relatives of the murderer was 20 strings of wampum—ten for the murdered man's life and ten for the murderer's life, which had been forfeited by the act of murder; the total for murder of a woman by a man was 30 strings of wampum, and so on.) Before the establishment of the league, tribes might go to war if the compensation was not considered just, a common custom in the area. (Witchcraft, not murder, was the most serious crime among Indians and evoked capital punishment.) The fixing of the price prevented war from occurring between members of the league, although not between members of the league and other tribes.

If the League of the Iroquois was a league of peace to its members, it was not to other Indian groups. In the 17th century, beaver had become scarce in Iroquois territory. They began a series of wars with their neighbours, defeating and dispersing the Hurons and Tionontati in 1649–50 (a number of Hurons and Tionontati were to join forces to become the Wyandot), the Neutrals in 1651, the Erie in 1656, and the Susquehanna in 1675. Earlier, about 1638, the Wenrohronons had joined the Hurons in consequence of Iroquois attacks on them. The Iroquois then took on still more distant Indian peoples.

The founding of the League of the Iroquois has been dated variously from about 1460 to about 1600, but, regardless of whether the date is pre- or post-Columbian, it seems likely that its strength and effectiveness increased as a consequence of European influence on the continent. Somewhat similarly, loose confederacies of various Indian bands were established along the Eastern Seaboard after European contact, in part to counter the force of the whites. Later, diverse groups of Indians living in the upper Great Lakes area banded together in a like manner, under the leadership of particularly influential chiefs, in order to affect events in that region.

Warfare

It seems likely that warfare among Indians increased in the years following European settlement and the establishment of trading posts. War parties were, typically, groups of men gotten together for the purpose of raiding for loot or for captives. For example, if compensation for murder had not been agreed on, a war party might be raised to avenge the murder and might include the taking of a captive to replace the murdered person. Before setting out, there might be a ceremony in which the men would boast of previous exploits, both to gain recruits and to rouse feelings to the proper pitch.

Among the Iroquoian, captives taken during war might be killed or brought back to the village, where they were either tortured to death or adopted into the tribe. If the captive was taken to compensate for a previous murder, he was often given to the family of the murdered person, and that family decided whether the captive was to be tortured to death or adopted as a member of the family. If the decision was for torture, members of the community tried to excel each other in the invention of tortures. The captive tried to accept these tortures without crying out—a practice that has contributed to the white stereotype of the stoic Indian. A captive who was killed might be scalped. Scalping may have been introduced by the Europeans, but the evidence is not conclusive. It may have been merely an elaboration of the old Indian custom to take a piece of hair or clothing of the killed person as a trophy of the event.

Clans. An important kind of social group in the area, particularly in the more southerly parts, where agriculture was relatively more important and population density higher, was the clan (sometimes now termed sib). The division of a tribe into clans served both to divide the community into smaller cooperating units and to provide a means of uniting it with others. Members of the same clan had certain obligations toward one another, regardless of tribal or community affiliations. They were expected to provide hospitality to visitors of the same clan, for example. Loyalty to clan often overshadowed that to the nuclear family, but the marriage tie among Indians often was a relatively unstable one and frequently was marked by little or no ceremony. In contrast, clan affiliation continued for life. On occasion, clan membership was as important as tribal, band, or village membership.

Names and relationships

Clans often have names that refer to an animal. The Seneca clans, for example, are Turtle, Bear, Beaver, Wolf, Snipe, Hawk, Deer, and Heron. The animal, or totem (a word adopted into English from the Algonkian word meaning "kinsmen"), had a special relationship to the clan. Members of a clan considered themselves to be related, whether an actual genealogical relationship could be traced or not. Partly because of this view, the members of a clan were not permitted to marry within the clan (that is, clans were exogamous).

Among the Iroquoians and the Delawares, a child automatically belonged to the clan of his mother (that is, the clans were matrilineal). Among the Winnebago and many other upper Great Lakes (Central) Algonkians, a child belonged to the clan of his father (that is, the clans were patrilineal—such patrilineal clans sometimes being termed gentes; singular, gens). Thus, for example, an Iroquois child of a father who belonged to the Wolf clan and a mother who belonged to the Turtle clan was a member of the Turtle clan, the clan of his mother. Further, he could not marry (without being accused of committing incest) a girl who also belonged to the Turtle clan. And if he married a girl, say, of the Bear clan, she did not become a member of the Turtle clan but remained a member of her own clan, the Bear clan, because clan membership was for life. (In a few instances, an individual might be adopted into another clan, but this had nothing to do with marriage.) In tribes that had clans, everyone in the tribe belonged to a clan; and, if a captive or other person was adopted into the tribe, he was also adopted into a clan of that tribe.

Commonly in this area, each clan owned a number of names, and a newborn child was given one of the clan's names that was not currently in use (that is, a name belonging to someone then deceased or a name that had been given up for a new name by someone still living). Certain names carried special responsibilities, such as those belonging to the 50 chiefs of the Iroquois League. When a chief died, the women of his clan decided on a successor, who was a member of the same clan. If the successor was approved by the other chiefs, he was given the name of the deceased chief in the condolence ceremony, in which the chief was "raised up" and resuscitated by giving his name to his successor.

Among many upper Great Lakes tribes, the clan owned a bundle of sacred objects, with which there was associated a ritual that was performed periodically by members of the clan. At least in recent times, the Iroquois had no comparable clan ceremonies; rather, a significant part of their ritual life centred on ceremonies in recognition of foods as they matured. These rituals included the maple, strawberry, bean, green corn, and harvest festivals, as well as the midwinter ceremony.

In some groups, such as the Winnebago and the Iroquois, the clans were grouped into two sides or moieties

("halves"), which had reciprocal obligations. Among the Iroquois, for example, members of a moiety buried the dead of the opposite moiety. Traditionally, the moieties were exogamous, and thus, in a sense, they provided wives for each other.

Medicine societies. Also often important in Indian religion are the medicine societies, so termed because one of their important functions was curing and because their membership consisted of individuals who had undergone such cures. The most famous medicine society among the upper Great Lakes Indians was the Mide Society (Midewiwin, or Grand Medicine Society), a society that held elaborate annual or semi-annual meetings during which various magical feats were performed. Of the various Iroquois medicine societies, the False Face Society is most familiar to whites because the carved wooden masks worn by members of this society during their rituals often are to be found exhibited in museums and pictured in books on Indian art.

Belief systems. Not all curing was performed by members of medicine societies. Certain individuals (often termed medicine men or shamans in anthropological literatures or powwows in the literature on New England Indians) had the power to cure, a power that was often indicated in a vision or dream. Dreams were especially important, because, generally among the people of this area, dreams indicated not only the causes of illness and an individual's power to cure but also the means of maintaining good fortune in various aspects of life. So much attention was paid to dreams that, among some Indians, a mother customarily asked her children in the morning if they had dreamed, for dreams were to be cultivated and attended to. Dreams could also influence the decisions of councils; as some early writers said, "The dream is the God of the country." Although the vision had some importance and although boys in the various tribes might seek a vision experience (particularly about the time of puberty), the vision quest (and the concomitant acquisition of a guardian spirit) was not as important in the area as a whole as it was among the Plains Indians.

Dreams and visions

The reliance on dreams, on the other hand, should not be interpreted as meaning that these Indians lived in a fantasy world. In these societies in which great emphasis was placed on self-reliance and on individual competence, attention to the content of dreams provided a means of understanding oneself and of bringing to consciousness knowledge stored in the unconscious, including knowledge as to where one's greatest abilities lay; for dreams and visions might indicate whether one had special ability in warfare, hunting, and the like.

Indians, moreover, did not lack real knowledge of herbal medicine. Indeed, it was the reputed success of their medicine that added credibility to the value of bottles of "Indian oil" and other medicines sold by the touring "Indian shows" put on by whites. Certain other practices, such as the sweat baths taken in small structures built for the purpose, also were beneficial in the treatment of others. The Indian practice of soliciting the aid of medicine societies and shamans to cure illnesses not cured by "natural" means is not unlike the modern-day treatment of psychosomatic illnesses.

More difficult for a member of Western civilization to comprehend is the "spirit world" of the Indian, for it appears to the outsider as if the Indian world was populated by a number of spirits, including spirits of such natural objects as the sun, the moon, thunder, and the trees. Some anthropologists have interpreted these beliefs as meaning that the Indians believed in pure supernatural power—that is, a force in the universe that, like electricity, could attach itself to certain persons and things. The Algonkian word *manito*, the Siouan word *wakan*, and the Iroquoian word *orenda* have been interpreted as referring to such supernatural power. There is some recent evidence, however, suggesting that this difference between Western and Indian thought rests on a different classification of the animate and inanimate worlds. The Indians classed as animate such things as the sun and the moon. And even those things that they regarded as inanimate they treated as "beings"—to be feasted as human beings are, to be called by kinship terms as human beings are, and in general to be paid attention to. Their exploits, like those of human beings, are reported in tales and myths.

COLONIAL HERITAGE

When the whites arrived on the North American continent, they brought manufactured goods that the Indians wanted and new diseases that they did not. Certain of the new diseases proved particularly virulent among Indians because they did not have the immunity that the whites had developed through centuries of exposure. At one time or another, all the Indian tribes were decimated by epidemics. In one notable instance, an epidemic swept over New England, killing many Indians in 1616–17, just a few years before the Pilgrims landed at Plymouth, a catastrophe that the Pilgrims saw as a sign of God's favour. (One reason that the Indian Squanto was so friendly to the Pilgrims was that he had been taken as a captive to Europe, spent some time in England, and had returned after the epidemic.)

Trade

The extensive trade that developed between the Indians and the Europeans (French, English, and Dutch) rested on the European desire for furs, particularly for beaver fur (beaver fur makes a fine felt, and beaver hats were the rage in Europe), and on the Indian desire for European-manufactured objects, particularly guns, kettles, needles, fishhooks, other metallic objects, and beads.

The Europeans soon discovered the Indians' desire for wampum (tubular beads made of shell) and thus established "wampum factories" in New Jersey and on Long Island. The Indians used wampum for jewelry and for gifts or signs to mark significant occasions. Important messages, for instance, were accompanied by strings of wampum; the making of treaties involved the exchange of wampum belts to confirm the sincerity of the parties and to serve as reminders of the agreement. They also were used on other political and religious occasions and kept as reminders of the occasions. Because it was valuable, wampum became a medium of exchange not only between Indians and white traders but also among the colonists. Because the coinage in common use in the colonies was already various, including Spanish, Portuguese, French, and Dutch coins as well as English ones, the adoption of wampum as another medium of exchange was an easy matter. Wampum, however, was not used as money before European contact.

Indian resistance to the settlers

The initial European settlement clung to the Atlantic coast—the sea providing the lifeline to the European homeland that the colonists needed—and thus the Eastern Seaboard Indians were first affected by the newcomers' desire for land. They were ill equipped to counter the invasion. Not only were their numbers relatively small and made even smaller by the epidemics, but their political organization was not of the kind that easily led to unified action of numbers of men. Friction between Indians and colonists did occasionally erupt, however, as in the Pequot War (1637) and King Philip's War (1675–76). Such resistance could not be maintained long, and these eastern Indians, as other Indians after them, began to adopt European ways as a means of survival. Such adoption often involved the acceptance and practice of Christianity as well as other features of European life. Some missionaries were especially influential. John Eliot, for example, accomplished the monumental task of translating the Bible into Algonkian and publishing it. There remain in New England and on Long Island some few reservations on which the Indians still are to be found. Further south, however, the Delawares gradually sold their land in the 18th century and began to drift west, becoming involved in the affairs of the country there.

The Iroquois fared somewhat better than the coastal Algonkians. In the 17th century and in the 18th century, they lived too far inland for European settlement, and only parts of their eastern territory were settled by whites. Also, the Europeans wished to maintain their trading relations with them and, through them, with Indians living farther west. The Iroquois themselves took advantage of this situation, engaging in war and diplo-

macy to maintain their position of power in the Northeast. Their power was finally broken in the American Revolution when George Washington, aware of the alliance of a number of Iroquois with the British, sent a punitive expedition into what is now upstate New York. After the Revolution, a number of Iroquois moved to Canada, though many remained in New York state. With the exception of many Oneida who moved to Wisconsin, those who remained in New York successfully resisted 19th-century efforts to move them (along with other eastern Indians) west of the Mississippi, and a number of them still live on the six Iroquois reservations in the state.

Like the Indians farther east, those of the upper Great Lakes area were greatly affected by the fur trade. The French established a series of trading posts, and the English challenged them for control of the area. Indians from the east, such as the Delawares, Wyandots, Ottawas, and Shawnee, drifted into the area seeking furs and land. The result was a series of wars and skirmishes between the English and the French and the Indians. Among the Indians, there emerged a series of prophets who attempted to revitalize Indian society and a series of chiefs who attempted to unite the Indians for the purposes of war. Notable among these were the Delaware Prophet, the Ottawa Pontiac, the Miami Little Turtle, the Shawnee Tecumseh and his half brother, the Shawnee Prophet (Tenskwatawa), and the Sauk Keokuk and Black Hawk. The upper Great Lakes Indians responded to various pressures put on them by whites, and many of them moved west. They are now to be found on reservations and other settlements of the more western states, though some still live in the upper Great Lakes area.

If there is any single thread that runs through the complex history of Indian–white relations, it is probably that of ambivalence. The Indians had certain resources that the whites wanted, and the whites had certain items that the Indians wanted. At times, these interests could be pursued to the benefit of all involved, but too often the interests of the various groups were in conflict. Emotionally, too, the attitudes have often been those of ambivalence: the Indians both wanting acceptance by whites and wanting their own separate identity and the whites both wanting to be rid of the Indians and admiring them. This ambivalence is likely to continue, for their histories and cultures are now so entwined that it is impossible to separate them.

BIBLIOGRAPHY. Names, locations, and historical events are summarized in FREDERICK WEBB HODGE (ed.), *Handbook of American Indians, North of Mexico*, 2 vol. (1907–10); and JOHN R. SWANTON, *The Indian Tribes of North America* (1952). Indian cultures of the area are summarized in ROBERT E. and PAT RITZENTHALER, *The Woodland Indians of the Western Great Lakes* (1970); W. VERNON KINIETZ, *The Indians of the Western Great Lakes, 1615–1760* (1940); and in general surveys of North American Indians as CLARK WISSLER, *Indians of the United States*, rev. ed. (1966); RUTH M. UNDERHILL, *Red Man's America* (1953); and OLIVER LA FARGE, *A Pictorial History of the American Indian* (1956). Descriptions of particular cultures include: FRANK G. SPECK, *Penobscot Man: The Life History of a Forest Tribe in Maine* (1940); W.D. and R.S. WALLIS, *The Micmac Indians of Eastern Canada* (1955); ROGER WILLIAMS, *A Key into the Language of America* (1643, reprinted 1971), despite its formidable title a delightful book and the best description of a southern New England Indian people; ALDEN T. VAUGHAN, *New England Frontier: Puritans and Indians, 1620–1675* (1965); WILLIAM W. NEWCOMB, JR., *The Culture and Acculturation of the Delaware Indians* (1956); LEWIS H. MORGAN, *League of the Ho-de-no-sau-nee, or Iroquois* (1851, reprinted 1966); GEORGE T. HUNT, *The Wars of the Iroquois* (1940); ANTHONY F.C. WALLACE, *The Death and Rebirth of the Seneca* (1970); ELISABETH TOOKER, *An Ethnography of the Huron Indians, 1615–1649* (1964); BRUCE G. TRIGGER, *The Huron Farmers of the North* (1969); PAUL RADIN, *The Winnebago Tribe* (1923); WALTER JAMES HOFFMAN, *The Menomini Indians* (1896); RUTH LANDES, *Ojibwa Sociology* (1937, reprinted 1969); WILLIAM T. HAGAN, *The Sac and Fox Indians* (1958); and BERT ANSON, *The Miami Indians* (1970). Modern reservation life is described in EDMUND WILSON, *Apologies to the Iroquois* (1960); and FREDERICK O. GEARING, *The Face of the Fox* (1970). An extensive listing of books and articles on particular tribes is given in GEORGE PETER MURDOCK, *Ethnographic Bibliography of North America*, 3rd ed. (1960).

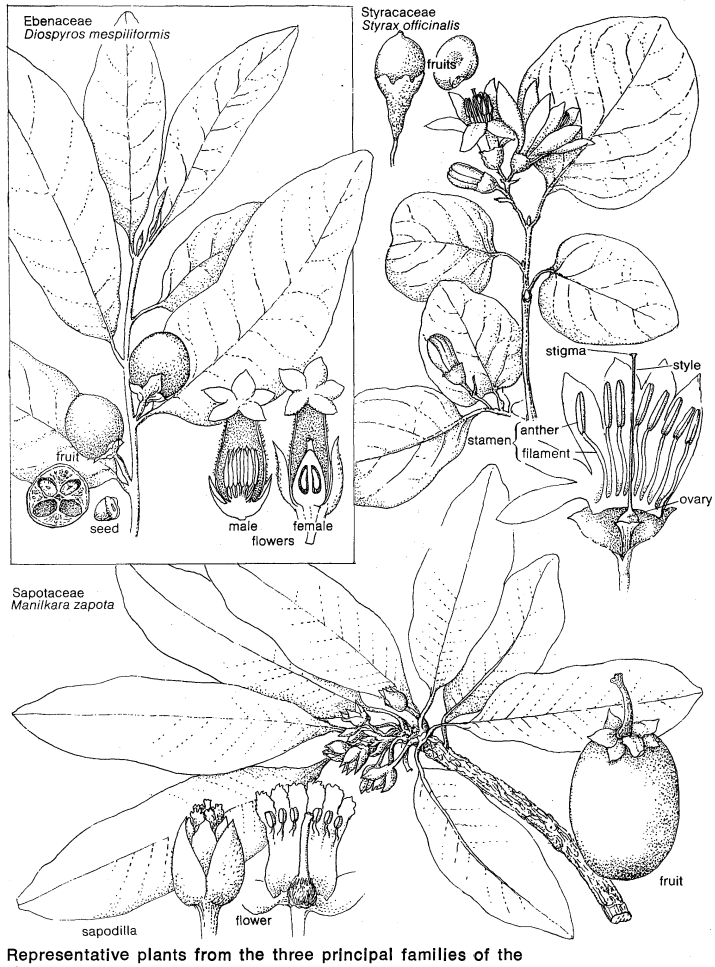
(E.To.)

Ebenales

The ebony order of flowering plants (Ebenales) consists essentially of woody trees or shrubs growing in the tropics. They have simple, smooth-margined leaves arranged alternately along the stems and bear radially symmetrical flowers that display diverse floral plans. The sepals and petals usually number four or five, and they are always fused. The male reproductive structures, or stamens, are variable in number and are arranged in one or more whorls. The female structure, or gynoecium, is superior —*i.e.*, positioned above the point of attachment of the sepals, petals, and stamens—and it is usually composed of many sections called carpels. The ovules, or seed precursors, are attached to a central axis in the ovary. Glandular, or secretory, cells are often present in the flowers.

Five families are included in the ebony order. Only three of these are of numerical or economic importance, however. They are the Ebenaceae, the Sapotaceae, and the Styracaceae.

Drawing by M. Pahl based on (*Diospyros* flowers) J. Hutchinson, *Families of Flowering Plants;* the Clarendon Press, Oxford



Representative plants from the three principal families of the ebony order.

GENERAL FEATURES

Distribution and abundance. The family Styracaceae contains about 150 species, which are grouped into ten genera; these plants grow mostly in hot climates. The most important genus, *Styrax*, contains about 100 species, of which one is found in the Mediterranean area (*Styrax officinalis*).

The family Ebenaceae contains only four to six genera, although it has a total of some 450 species. The genus *Diospyros* (persimmon and ebony) is the most important

one in the family both numerically (350 species) and economically. The present-day distribution of the family Ebenaceae is largely pantropical. One-third of the species are found in Asia (India, Indonesia, Indochina, and the Philippines); another third are in Africa and Madagascar; and, of the remaining third, one-half are in Central and South America and the other in Oceania (including Australia). There are, of course, a few species in subtropical climates, and four are found in warm temperate zones (North America and Asia). Ecologically, 85 percent of the tropical species grow in forests, with 60 percent in the evergreen forests. The zone of distribution of different species of the family Ebenaceae varies enormously; for example, *Diospyros virginiana* (persimmon) extends from Texas to Florida and northern Connecticut, but *D. mafiensis* is known only from a very small area in East Africa.

The four-continent distribution of Ebenaceae

The family Sapotaceae contains 63 to 126 genera, depending on the authority consulted, and more than 600 species, whose features contrast sharply depending on whether the species is found in the New World or the Old World. For example, no genus with a two-whorled calyx or with more stamens than petals is to be found in the Americas. Genera of plants with ovaries of three or fewer chambers occur very frequently in the Americas, but are exceptional in the Old World. The same contrast applies to the relative distribution of unisexual—separate male or female—flowers. In Africa the area from Guinea to the Congo Basin is richly endowed with species of the family Sapotaceae, particularly in Cameroon and Gabon. Madagascar has both pantropical and endemic (strictly native) genera. The island of Mauritius has an endemic genus (*Labourdonnaisia*). The genus *Northea* is found only in the Seychelles. The dry forests, savannas, and steppes that range from The Sudan to the Zambezi River are poor in Sapotaceae species, apart from a few endemics, of which *Butyrospermum parkii* is of economic importance—its oily seeds are pressed to obtain shea butter. New Caledonia is outstanding in the Australo-Papuan zone for the number of endemics of this family. Several regions of tropical America also have particular genera, in addition to genera with more extensive distributions.

Economic importance. *Family Styracaceae.* Economically, the best known representative of the family Styracaceae is the Malayan *Styrax benzoin*, a tall tree that is the source of the resin, benzoin. The resin forms in cavities caused by the destruction of tissues in the wood of stems and branches. *Styrax tonkinensis* is also a source of resin; deep incisions are made in the trunks, and the drops of resin that form in the cuts are collected about two months later. The resins contain vanillin, benzoic acid, and cinnamic acid. The Arab benzoin trade dates back before the Christian Era and provided incense to the Middle East and China. In the Western world it is used for perfumes and as a balsam and antiseptic. Several plants of the family Styracaceae are also used for ornamental purposes (*e.g., Halesia, Styrax*).

Ebony family. The family Ebenaceae is of economic importance especially for its timber, which is used in cabinets. The wood, which is denser than water and hence sinks, has a hard, fine grain and is difficult to work, but it takes a magnificent polish. The heartwood of a number of species of *Diospyros* is used to provide a variety of colours: green in *D. chloroxylon* from India; red in *D. rubra* from the Mascarene Islands; white in *D. chryophyllos*, also from the Mascarene Islands; and variegated from *D. hirsuta.* Finally, the black woods, the true ebonies, are obtained from *D. dendo* and *D. mespiliformis*, which grow in equatorial Africa; *D. crassiflora* in Gabon and Cameroon; *D. ebenum* in Ceylon; *D. melanoxylon* in India; *D. perrieri* and *D. haplostylis* in Madagascar; and *D. celebica* in the Celebes. There are also other species that produce true ebonies. The quality of the wood is related to chemical changes in the cell walls that occur with respect to the formation of gums. As a result of abusive and indiscriminate exploitation, natural stands of these trees have been considerably diminished.

Edible fruits

The fruits of some species of *Diospyros* are edible. The name itself signifies "celestial fruit" or "fruit of the gods." The best known of these is the kaki, or Japanese persimmon, *Diospyros kaki,* cultivated in Japan. About 800 varieties of this species have been developed, some of which have been acclimatized to the Mediterranean area. The fruits, large orange-red berries, are very astringent until fully ripe, at which time the tannin content is completely transformed into insoluble crystals. The juice is then sweet and palatable.

Although *D. kaki* has been cultivated in Japan for several centuries, it originated in China where it is also highly valued. The fruits of *D. virginiana*, the common persimmon of the United States, are used in jam making or dried like prunes. *D. ebenaster*, the black sapote, is a popular fruit, but little known outside Mexico. The dark colour of the flesh may make it unattractive to the uninitiated, yet it is considered to be equal in quality to the Japanese persimmon. *D. lotus*, a native of Central Asia and the Middle East, is grown in the Mediterranean area of western Europe as an ornamental tree. *D. mespiliformis* from Africa not only has edible fruit but its leaves, roots, and bark possess bactericidal properties. The bark of *D. multiflora* is used in the preparation of arrow poisons.

Family Sapotaceae. There are many species of the family Sapotaceae that are important as sources of timber, fruit, and gamboge (a resin). The heavy, hard, and long-lasting woods of some species of *Pouteria* (white balata) and *Manilkara* (red balata) are harvested in the Guianas and in Brazil, as is *Chrysophyllum prieurei*, also from the Guianas, and *Mastichodendron* species (ironwoods) from the Antilles. In Africa among the commercially important species are *Tieghemella heckelii* from the Ivory Coast and Gabon and *Autranella congolensis* of Central Africa. In tropical Asia several species of *Palaquium* are cut for timber, as is *Sideroxylon zwagei* in Indo-Malaya.

The fruits of some Sapotaceae species are edible. The sapodilla (fruit of *Manilkara zapota*, sometimes called *Nispero achras*) has been described as luscious, cool, agreeable, and fragrant. This plant, a native of southern Mexico, not only furnishes an excellent fruit but also a gum, known even to the early Aztecs, who chewed it. Now the milky latex is extracted from the trunk and from unripe fruit, boiled with aromatic ingredients, and made into chewing gum. This useful tree has been cultivated in many hot regions: West Bengal in India, the Hawaiian Islands, and parts of Africa. The sapote (*Calocarpum mammosum*) is another useful fruit from Central America. It is said that Cortés, the Spanish conqueror of Mexico, and his army were able to stay alive by eating this fruit during their march from Mexico to Honduras. It is also cultivated in the West Indies, South America, and the Philippines. In the West Indies the sapote is called marmalade plum; in Cuba, mamey colorado. The star apple (from *Chrysophyllum cainito*) is another fruit of Central America and the West Indies.

Other fruits of the family Sapotaceae with local popularity are the canistel (*Lucuma nervosa*), grown in Cuba and Florida; the abia (*Pouteria caimito*) in Peru; and the gambeya (*Synsepalum dulcificum*) in Africa.

Economically valuable oils from seeds

The seeds of many Sapotaceae species contain useful oils, and those of several species of *Madhuca* from India are used both locally in the production of mohwra and illipe butter and exported for the manufacture of edible fats. The oil cakes made of these seeds are rich in saponins (glucose-containing substances that produce soapy lathers) and toxic substances, which have to be eliminated before the cakes can be used as animal fodder. The *beurre de Karité* ("butter of Karité") of The Sudan and the Nile River area across to Senegal is extracted from the seeds of *Butyrospermum paradoxum* (now *Vitellaria paradoxa*), a variously shaped tree native to the savannas of the area. This oil is an important element in the nutrition of the native populations of those areas. In southern Morocco the seeds of *Argania spinosa* yield an oil totalling about 50 percent of the weight of the kernel. The oil, of strictly local use, is an important item in the diet of the tribes of the area.

Other species of the family Sapotaceae furnish latex for

commercial use. Gutta-percha, related to rubber but not elastic, is used to insulate submarine electric cables and in the manufacture of golf balls; the substance is found in several species of *Palaquium* from Malaysia (*e.g.*, *P. gutta* and *P. leiocarpum*). In tropical America *Manilkara bidentata* provides a latex known as balata gum, which goes into the manufacture of diving belts.

NATURAL HISTORY

Little is known about the biology of the family Ebenaceae. The aromatic flowers of many species open at night, and the petals fall the next morning. Parthenocarpy, the production of fruits without fertilization, has been observed in *Diospyros kaki*, from which seedless fruits have been obtained. The dormancy of the seeds in some species has been shown to result from the mechanical resistance of the seed coats.

Sapotaceae as keys to the ages of forests

Ecologically, the family Sapotaceae is represented most abundantly in damp forests, where they give an indication of the age of the forest by their quantity relative to other species. They are usually large-sized trees. Members of this family are also fairly numerous in the coastal thickets of tropical seas and along river banks. Although they are practically lacking in the arid zones, spiny representatives grow in semi-arid areas, and small trees less than ten feet high occur in thickets in New Caledonia.

Only the cultivated species have really been studied. It has been observed, however, that certain species (*e.g.*, *Butyrospermum paradoxum*) are a particular attraction to bees and that others are pollinated at night by bats.

FORM AND FUNCTION

Family Styracaceae. The family Styracaceae is composed of small trees or shrubs with alternately arranged leaves that lack stipules, small leaflike appendages at the bases of leafstalks. The leaves are characterized by a hairy covering of stellate (branched in starlike patterns) plant hairs or scales. They have five-parted flowers with a calyx that is often fused.

The corolla (collection of petals) is almost many petalled. The male reproductive complex is usually composed of two whorls of stamens (diplostemony) although single whorls sometimes occur. The gynoecium is located above the attachment point of the other flower parts or is partly enclosed by the bases of the sepals, petals, and stamens—its position is thus said to be superior or semi-inferior. It is composed of three to five segments called carpels, which form the same number of interior chambers, or locules. The partitions between the locules, however, are incompletely formed and exist only in the lower half of the ovary. One or more erect or pendulous ovules are present in each locule. The fruits are drupes (stony-seeded, nonsplitting fruits) or capsules, containing one or more seeds.

The ebony family. The family Ebenaceae consists of trees and shrubs with alternate (rarely opposite) simple, smooth-margined leaves that lack stipules. The plant hairs are generally formed of single cells but sometimes occur in tufts, or at times are branched into two parts. Glandular hairs also occur.

The solitary flowers occur in leaf axils (upper angles between leafstalks and stems) or form restricted, few-flowered inflorescences (clusters). They occasionally occur even on the trunk or older branches (cauliflory). They are radially symmetrical, and their parts occur in basic numbers of three to seven. The plants are dioecious—the male and female flowers occur on separate plants—but both flower sexes often have vestigial or sterile organs of the opposite sex in their structure. The stamens are generally attached either to the inner surface of the tubular structure formed by the joined lower portions of the petals (corolla tube) or at its base. They are either equal in number to the corolla lobes (petals) or two to four times as many (up to 120). In the female flowers the superior ovary is composed of two to eight carpels with as many locules or more, because of the formation of false septa, or walls. Each locule contains one or two ovules suspended from its summit. The fruit is usually a berry containing one to 16 seeds.

Family Sapotaceae. The family Sapotaceae consists mainly of large trees that are easily recognizable by the sticky latex throughout the plant. Their branched plant hairs also are distinctive, but the arms, or branches, of the hairs may not always be equal. The leathery leaves are simple and lack stipules; foliation is dense, particularly at ends of branches where the leaves form tufts.

Sapotaceous flowers may be solitary, grouped in axillary inflorescences, or cauliflorous. They are usually small, regular in shape, and bisexual. Each flower has a calyx of four to eight overlapping sepals in a single whorl. The corolla has as many parts as has the calyx, but the overlapping petals are always more or less fused. The stamens occur in two, three, or even more whorls; their arrangement is characteristic: attached, singly or in groups, opposite the petals. The superior ovary is composed of fused carpels with one to 30 locules, each containing one ovule attached along the central axis of the ovary.

The fruit is usually a hardened berry containing one to five seeds (rarely more), which are liberated after the fallen fruit rots on the ground. The seeds characteristically have a thick, smooth, and shiny seed coat and a large, dull, basal scar (hilum). The position, size, and shape of the hilum is often used to delimit groups within the family. The seeds contain a well-developed endosperm (nutrient tissue for the developing embryo) and are frequently rich in oil.

EVOLUTION AND CLASSIFICATION

Evolution. The order Ebenales is thought to have had its ancestry in the order Theales, especially the family Theaceae. The family Styracaceae is considered to be the most primitive within the order. *Styrax officinalis*, which grows today in the Mediterranean region, is known as a fossil from the Eocene Epoch (38,000,000 to 54,000,000 years ago) from areas where it is now lacking (*e.g.*, central Germany).

Fossil specimens

Fossil members of the genera *Diospyros* and *Royena* have been uncovered in strata of the Upper Cretaceous (about 65,000,000 years ago and earlier), and wood samples of the family Ebenaceae have been taken from Oligocene rock (26,000,000 to 38,000,000 years ago) in Hungary.

Members of the family Sapotaceae lived during the Upper Cretaceous, according to the fossil record, with representatives of the genus *Sideroxylon* in Europe as late as the Oligocene Epoch. The family is now pantropical, with one-third of its genera in Central and South America, one-third in Africa, and one-third in India and the Pacific.

Classification. *Annotated classification.*

ORDER EBENALES

Trees or shrubs. Leaves simple, entire (smooth-margined), alternate. Flowers radially symmetrical; sepals and petals of 4 or 5 fused segments each. Stamens variable in number, arranged in one or more whorls. Ovary superior (sepals, petals, and stamens arise at its base), composed of many carpels (ovule-bearing segments), and with axile placentation (ovules attach along the central axis).

Family Styracaceae

Small trees or shrubs. Leaves without stipules; stellate (star-shaped) hairs or scales present on leaves. Latex (milky fluid) absent. Flowers bisexual, occasionally polygamous (with bisexual and unisexual flowers present on the same plant). Stamens usually in two whorls, alternate with the petals. Ovary superior to inferior, composed of 3 to 5 carpels with 1 or more erect or pendulous ovules per locule (chamber) Fruit a drupe (stone-seeded fleshy fruit) or capsule. Ten genera and about 150 species distributed mostly in hot climates.

Family Symplocaceae

Trees and shrubs. Leaves simple, alternate, often leathery. Flowers bisexual, radially symmetrical, in racemes (clusters with a central axis and lower flowers maturing first). Sepals 5, fused, overlapping. Petals 5, fused, or in 2 series of 5 each, fused. Stamens 5, or in 2 or 3 whorls or more of 5 each, epipetalous (attached to inside of floral tube) or free of corolla (petals). Ovary inferior (surrounded by the basal portions of the sepals, petals, and stamens) or semi-inferior, of 2 to 5 carpels and locules. Ovules 2 to 4 per locule, anatropous (inverted and straight), pendulous, on axile placenta.

Fruit drupaceous, 1 seed in each locule of the stone. Two genera and about 500 species distributed in the tropics and subtropics except Africa.

Family Lissocarpaceae

With the characteristics of the family Ebenaceae except the ovary is inferior. One genus (*Lissocarpa*) and 2 species distributed in tropical South America.

Family Ebenaceae (ebony family)

Trees or shrubs. Leaves simple, entire, alternate, and without stipules. Latex absent. Flowers solitary or in few-flowered clusters, radially symmetrical, 3- to 7-parted, usually unisexual with separate sexes on different plants (dioecious), occasionally polygamous. Stamens inserted at the base of the corolla tube (hypogynous) or on the corolla tube (epipetalous), either equal in number to the corolla lobes or 2 to 4 times as many (up to 120). Ovary superior, of 2 to 8 carpels with the same number of locules as carpels or more through the formation of false septa. Ovules 1 or 2 per locule, anatropous, suspended from summit of locule. Fruit usually a berry containing 1 to 16 seeds. Four to 6 genera and about 450 species distributed largely within the tropics of both hemispheres with a few species in warm temperate zones.

Family Sapotaceae

Mostly large trees with sticky latex throughout and characteristic plant hairs bearing 2 branches (2-armed hairs). Leaves simple, usually entire and leathery, alternate, without stipules, usually in dense foliation particularly at branch ends where tufts of leaves occur. Flowers usually small, radially symmetrical, bisexual, rarely unisexual or polygamous, solitary or grouped in axillary clusters (in the upper angles between leaves and stem) or borne directly on the trunk or older branches (cauliflorous). Calyx of 4 to 8 overlapping sepals in a single whorl, fused at the base or completely free. Corolla parts as many as the sepals, the petals always more or less fused. Stamens in 2, 3, or more whorls, epipetalous, inserted singly or in groups opposite petals. Ovary superior, of fused carpels, with 1 to 30 locules and axile placentation. Ovules 1 per locule, anatropous, erect. Fruit a hardened berry with 1 to 5 seeds, rarely more. Genera number 63 to 126 depending on authority consulted, with more than 600 species mostly of damp tropical forests.

Critical appraisal. A considerable difference of opinion continues on the subdivision of the family Sapotaceae into tribes or subfamilies, as well as on the delimitation and position of its genera. One authority, for example, in his treatment of the order included 63 genera; another, 126. The lack of complete herbarium material complicates the task of placing the various taxonomic groups (genera and species) within such a homogeneous family. The characteristics available for separating genera and species are, according to some workers, a continuous network of smoothly intergrading features that do not permit easy separation into smaller groups. As a result, it is almost impossible to assess the evolutionary trends of the family.

Some authorities remove the family Styracaceae from the order Ebenales. The pollen is similar to that of the family Ebenaceae, however, but is also rather like that of the Cornaceae and Nyssaceae, both of which are families of the order Cornales.

BIBLIOGRAPHY. A. AUBREVILLE, *Sapotacées* (1964), one of the most important monographic studies of recent years; C. BAEHNI, *Mémoires sur les sapotacées*, 3 vol. (1938–65); "Sapotacées du Surinam et de la guyane anglaise," *Candollea*, 14:61–77 (1952); and "Pouteriae novae," *Candollea*, 18:161–176 (1962–63), a series of publications of great interest by one of the most active specialists of his time; R. LETOUZEY and F. WHITE, *Ebénacées*, vol. 11 of A. AUBREVILLE and J.F. LEROY, *Flore du Cameroun* (1970), a regional flora but the most recent taxonomic treatment of the African (Cameroonian) species of the family Ebenaceae; E. PERROT, *Matières premières usuelles du règne végétal*, vol. 2, pp. 1707–1737 (1944), a good reference work on plant products, being particularly devoted to products of the Ebenales; W. POPENOE, *Manual of Tropical and Subtropical Fruits* (1927), much useful data included on the fruit of the Ebenales; L. EMBERGER, *Traité de botanique (systématique)*, vol. 2, pp. 849–859 (1960), a recognized treatise on modern systematic botany containing information on the Ebenales; F. WHITE, "Notes on the Ebenaceae," *Bull. Jard. Bot. État Brux.*, 26:237–246 and 277–307 (1956), 27:515–531 (1957), and 33:345–367 (1963); and *Bolm. Soc. Broteriana*, Series 2, 36:97–100 (1962), taxonomic treatments of African species of the family Ebenaceae; C.E. WOOD, JR. and R.B. CHANNELL, "The Genera of the Ebenales in the Southeastern United States," *J. Arnold Arbor.*, 40:1–35 (1960), an excellent reference work with a particularly rich bibliography.

(J.Mi./C.E.B.B.)

Ebert, Friedrich

Friedrich Ebert, a leader of the German Social Democratic movement, became the first Social Democratic president of the German Republic. At heart a constitutionalist who wanted to establish parliamentary democracy but abhorred revolution, he was throughout his life attacked by conservatives of every shade as well as by radical leftists whom he did not scruple to suppress in bloody clashes that occurred during the early years of the republic.

Archiv fur Kunst und Geschichte, Berlin

Ebert, *c.* 1924.

Ebert was born on February 4, 1871, in Heidelberg, the son of a master tailor. He learned the saddler's trade and travelled through Germany as a journeyman saddler. He soon became a Social Democrat and trade unionist, representing so-called revisionist—gradualist, liberal—"trade-union" Socialism, without, however, displaying a deep interest in the ideological struggles of Marxism. His attention was always directed toward practical improvement in the living conditions of the German working class and, above all, its social and moral betterment.

Work with the Social Democratic party

In 1905 Ebert became secretary general of the German Social Democratic Party (SPD). The party had steadily increased in membership and electoral support and had accumulated physical assets and property. He updated party administration, introducing typewriters and filing systems that the party had lacked until that time due to its fear of house searches.

Ebert succeeded August Bebel as party chairman in 1913. Under his leadership, the Social Democratic movement gained increasing influence in German national politics. It was Ebert, in particular, who on August 3, 1914, prevailed upon German Social Democrats to support the war appropriations. The action of the German Social Democratic Party did not differ from that of the other Socialist parties of Europe, in which nationalist feelings remained stronger than internationalist convictions. To its own detriment, Ebert's party gave the "Fatherland" its unconditional support without requiring that Germany adopt a genuine peace policy. In consequence, it lacked the power to force the government to adopt a policy through which Germany might have escaped the crushing defeat that was to destroy the empire and also eventually Ebert's postwar policy.

Ebert could not hold the entire party to his course for long. In March 1917, a left-wing faction left the party to become the Independent Social Democratic Party of Germany (USPD), strenuously rejecting war appropriations and Germany's war policy. Another group split from the SPD to form the Communist Party of Germany

(KPD). The leftists who had withdrawn from the SPD sought a social revolution, while Friedrich Ebert and his party wanted to establish a German parliamentary democracy. Even in the midst of the war, the Catholic Centre Party, the Democratic Party (previously the Progressive Party), and the Social Democrats had formed the so-called Black–Red–Gold (Weimar) coalition, named after the colours of the flag of the liberal revolution of 1848.

With Ebert's active cooperation, a new government, headed by Prince Max of Baden and consisting of the three parties of the Black–Red–Gold coalition, was organized in October 1918 through a sweeping constitutional reform that in essential respects foreshadowed the Weimar Constitution. Thus, Germany would not have needed a revolution to achieve parliamentary democratic reform, and Ebert did everything he could to prevent such a revolution from occurring. "I hate the revolution like sin," he later said to Chancellor Max of Baden. But the revolution of November 1918 was not made by Germans to bring about the advent of the republic, democracy, or even Socialism. For nearly all Germans, the revolution had only one aim: peace. Rightly or wrongly, the German people believed that Emperor William II would not secure peace for Germany.

The revolution, winning its race with peace, came three days before the armistice. It triumphed in Berlin on November 9, and on the same day Prince Max of Baden, acting on his own authority, asked Friedrich Ebert to replace him as chancellor. Ebert, who still hoped to establish a regency for the Emperor, actually held office as chancellor for one day. On November 10 he yielded to the *fait accompli* of the revolution and set up an entirely Socialist government, with representatives from the SPD and USPD. Calling itself the Council of People's Representatives, the government derived its authority from the Workers and Soldiers Council, which claimed to speak for Germany and the German Republic but in truth had been elected rather arbitrarily by the factories and regiments of Berlin alone. Ebert was determined to place the power of the Council of People's Representatives and the Workers and Soldiers Council in the hands of a freely elected German parliament as soon as possible. He wished to see a liberal coalition government rather than a Socialist regime in power.

The elections of January 1919 gave the Black–Red–Gold coalition a majority of 85 percent. The republic's first government, under Ebert's fellow party member Philipp Scheidemann, was based on this tripartite coalition, and the new German constitution, the Weimar Constitution, so-called after the town in which it was drawn up, was the work of the coalition. By the votes of the three parties forming the coalition, Ebert was elected the first president of the republic.

Election to the presidency

Ebert and Hugo Preuss, a professor of constitutional law whom he had charged with the task of drafting the constitution, wished to alter the organic structure of the *Reich*. But the old German states (the *Länder*, or territories) successfully resisted the "unitary state" (*Einheitsstaat*) of Ebert and Preuss. Prussia in particular continued to exist as a state. The groups and forces that had until then been the pillars of the old Germany also remained intact, for the first years of the Weimar Republic were taken up by the bloody civil war which the government, under Ebert's presidency, waged against the leftist Socialists and Communists, who had been Ebert's former comrades. The republic exhausted itself in the civil war against Communism and lacked the strength to carry out the basic changes in the *Reich* that might have placed the republic on a lasting foundation. The workers did not want to make an armed defense of the democratic republic. So Ebert and his friend Gustav Noske, the defense minister, had recourse to volunteer groups, the *Freikorps*, which were principally composed of officers of the old army, and suppressed the Communist uprising out of hatred of Communism rather than love of the republic. The old officer corps formed the backbone of the Reichswehr, the republic's army. Together with the officer class and the old officialdom, the Junkers—the landed gentry east of the Elbe River—with their great estates and influence in social and political life, also survived the revolution.

With the elections to the republic's first parliament on June 6, 1920, the Black–Red–Gold coalition lost its majority and was never to regain it. The Social Democratic Party thereby lost its commanding position in the *Reich*, and the political constellation on which Ebert's leadership had been based dissolved. The electoral defeat was a direct result of the Treaty of Versailles. At that time many Germans, including Ebert, were convinced that the peace of Versailles aimed at the destruction of Germany. The resulting loss of confidence in the Black–Red–Gold coalition was the death blow of the Weimar Republic, although in fact the country's strength and stability had remained untouched.

Nevertheless, the first consequence of the Treaty of Versailles was a coup d'etat against the republic, the Kapp Putsch, by radical nationalists, a part of the Reichswehr, and the *Freikorps* which were to be disbanded under the provisions of the peace treaty. The coup of March 13, 1920, led by Wolfgang Kapp, a provincial bureaucrat who planned a restoration of the monarchy, collapsed after a few days, but Ebert's dream of a reconciliation between the army and the Social Democrats was shattered.

Soon thereafter, the government was confronted by a near-fatal crisis. When, in January 1923, Germany was declared in default of coal deliveries under the reparations provisions of the Treaty of Versailles, France, wishing to settle the reparations question decisively, occupied the Ruhr territory. Ebert, like nearly all Germans, supported national resistance and the general strike in the Ruhr, which was directed toward ending foreign military control. But Germany suffered as a result of the strike, in which eventually millions became idle. Inflation assumed staggering proportions, and the country experienced its most severe social and political crisis. Adolf Hitler nearly succeeded in seizing power in Bavaria. Chancellor Wilhelm Cuno, an independent, appointed on the eve of the Ruhr struggle as a man whom Ebert particularly trusted, was helpless in the face of the crisis. Gustav Stresemann, of the right of center People's Party, who succeeded Cuno, brought the crisis under control. Ebert initially appointed him only with hesitation and treated him with reserve but finally gave him his full support. He bitterly rebuked his own party when, protesting Stresemann's move to a more rightest position, it dropped out of the governing coalition and thus brought about the chancellor's resignation in November 1923. In point of fact, Ebert's party had thereby eliminated itself from active participation in German national politics for many years to come.

Last years

The unity of the *Reich* was preserved. Inflation was ended through monetary reform, and a means for resolving the reparations question was partially resolved in an American proposal providing for their reduction. The evacuation of the Ruhr district was in sight. Yet much of the German right persisted in its defamation of Friedrich Ebert. The judgment of a German court, which ruled that Ebert had committed high treason, at least in the legal sense, during the war by his support of a munition worker's strike, contributed to his early death. He died February 28, 1925, at the age of 54.

BIBLIOGRAPHY. Ebert's writings, speeches, and notes may be found in *Friedrich Ebert: Schriften, Aufzeichnungen, Reden*, with previously unpublished material from his estate, compiled by FRIEDRICH EBERT, JR., with a brief biography by PAUL KAMPFFMEYER, 2 vol. (1926); see also *Friedrich Ebert: Kämpfe und Ziele* (1927); PAUL LOEBE *et al.*, *Friedrich Ebert und seine Zeit* (1928); HERMANN MUELLER-FRANKEN, *Die November-Revolution Erinnerungen* (1928); and PHILIPP SCHEIDEMANN, *Memoiren eines Sozialdemokraten*, 2 vol. (1928; Eng. trans., *Memoirs of a Social Democrat*, 2 vol., 1929). There is no definitive biography of Ebert, but useful information may be found in MICHAEL FREUND, "Friedrich Ebert," *Die grossen Deutschen*, vol. 4 (1957); MAX PETERS, *Friedrich Ebert*, 2nd ed. (1954), in German; and CARL E. SCHORSKE, *German Social Democracy, 1905–1917* (1955).

(M.Fre.)

Echinodermata

The Echinodermata comprise a phylum of marine invertebrate animals whose extant representatives include sea lilies and feather stars (Crinoidea), starfishes or sea stars (Asteroidea), brittle stars or serpent stars (Ophiuroidea), sea urchins (Echinoidea), and sea cucumbers (Holothuroidea). Beginning with Lower Cambrian strata almost 570,000,000 years old, echinoderms have a rich fossil history and are well represented by many bizarre groups, most of which are now extinct.

Echinoderms have been known to man since ancient times; echinoids, for example, were used extensively by Greeks and Romans for medicinal purposes and as food. During the Middle Ages, fossil echinoids and parts of fossil crinoids were objects of superstition. In the early part of the 19th century, the Echinodermata (meaning spiny-skinned), recognized as a distinct group of animals, occasionally were associated with the coelenterates in a group known as the Radiata. Echinoderms now may be separated into 20 classes, based mainly on differences in skeletal structures. The number of extant species exceeds 6,000, and approximately 13,000 fossil species have been described.

GENERAL FEATURES

Structure. Echinoderms have a skeleton composed of numerous plates of mineral calcium carbonate (calcite). Part of the body cavity, or coelom, is a water-vascular system, consisting of fluid-filled vessels that are pushed out from the body surface as tentacles and are used in locomotion, feeding, respiration, and sensory perception. The conspicuous five-rayed, or pentamerous, symmetry of living echinoderms tends to obliterate their fundamental bilateral symmetry; *i.e.*, body of mirror-image halves.

Colour and size range

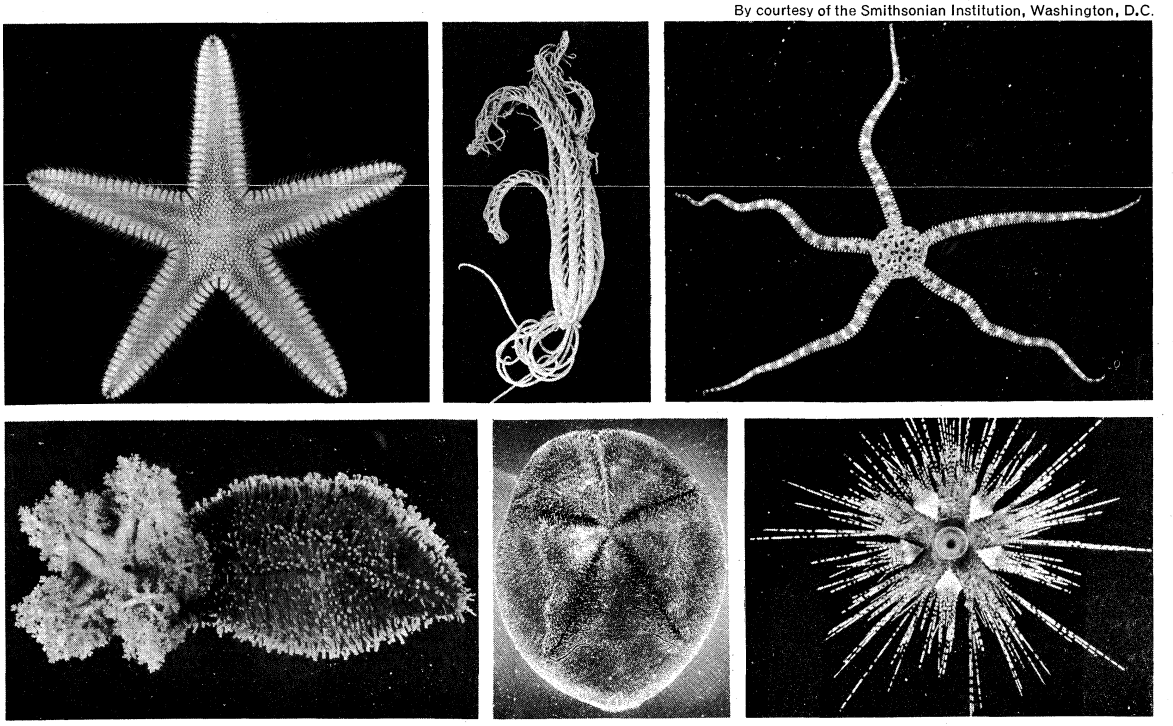
Diversity. Echinoderms exhibit various forms, sizes, and colours and especially brilliant colours such as reds, oranges, greens, and purples. Many tropical species are dark brown to black, but lighter colours are common among species not normally exposed to strong sunlight. Although most echinoderms are small in size, ranging to several centimetres (one inch is 2.54 cm) in diameter, some attain relatively large sizes; *e.g.*, some sea cucumbers (Holothuroidea) are as long as two metres, and some starfishes (Asteroidea) have a diameter of one metre (about 3.3 feet). Among the largest echinoderms were fossil crinoids (sea lilies), which attached to surfaces by holdfasts on stems exceeding 20 metres (66 feet) in length.

Distribution and abundance. Diverse echinoderm faunas consisting of many individuals and many species are found in all marine waters of the world except the Arctic, where few species occur. Asteroids (starfishes) and echinoids (sea urchins), including globular spiny urchins and flattened sand dollars, are commonly found along the seashore. Although many species are restricted to specific temperate regions, Arctic, Antarctic, and tropical forms often are widely distributed; numerous species associated with coral reefs, for example, range across the entire Indian and Pacific oceans. Many of the echinoderms of Antarctica are distributed around the continent; those with a floating (planktonic) immature, or larval, stage may be widely distributed, carried great distances by ocean currents. Some species, particularly in Antarctic and deep-sea regions, which have achieved a wide distribution without benefit of a floating larval stage, may have done so by migration of adults across the sea floor or, in the case of shallow-water species, by passive transport across oceans in seaweed.

IMPORTANCE

Role in nature. Echinoderms play an important role in the general ecology of the sea; they are, for example, efficient scavengers of decaying matter on the sea floor and they prey upon various organisms, thereby helping to regulate their numbers. On the other hand, they produce vast numbers of larvae that provide food for other planktonic organisms. By burrowing into rocks along a shore, echinoderms help to alter shorelines; in addition, they affect the structure of sea floor sediments. Echinoderm skeletons, which formed a large part of massive limestone beds, now are used by paleontologists to date them.

By courtesy of the Smithsonian Institution, Washington, D.C.



Representative extant echinoderms.
(Top left) An asteroid, *Astropecten*, showing typical five-rayed symmetry. (Top centre) A crinoid, the feather star *Notocrinus*. The long arms carrying pinnules and slender cirri are used for attachment to a substrate. (Top right) An ophiuroid, *Ophiomyxa flaccida*, has long slender arms and disc with mouth. (Bottom left) A holothurian, *Cucumaria*, has a soft body with numerous tube feet and ten tentacles around the mouth. (Bottom centre) An irregular echinoid, the sand dollar, *Meoma ventricosa*. (Bottom right) A regular echinoid, *Astropyga*, has long spines, test with five-rayed symmetry, and a soft anal cone central structure.

Relation to man. *Advantages.* Some of the larger species of tropical holothurians (sea cucumbers), known commercially as *trepang* or *bêche-de-mer*, are dried and used in soups, particularly in the Orient. Raw or cooked mature sex organs, or gonads, of sea urchins (Echinoidea) are regarded as a delicacy in parts of the world, including Europe, the Mediterranean region, and Chile. Some tropical holothurians produce a toxin, known as holothurin, which is lethal to many kinds of animals; Pacific island natives kill fish by poisoning waters with holothurian body tissues that release the toxin. Holothurin does not appear to affect man; in fact, the toxin, which reduces the rate of growth of certain types of tumours, may have medical significance. The easily obtained eggs and spermatozoa of echinoderms, particularly those of echinoids (sea urchins), have been used to conduct much research in developmental biology; indeed, echinoids have been collected in such large numbers that they have become rare or have disappeared altogether from the vicinity of several marine biological laboratories.

Holothurin toxin

Disadvantages. Some starfishes, which prey upon commercially usable mollusks such as oysters, have caused extensive destruction of oyster beds. Sea urchins along the California coast have prevented the regrowth of commercial species of seaweed by eating the young plants before they could become firmly established. The "crown-of-thorns" starfish, which feeds on living polyps of reef corals, has caused extensive short-term damage to coral reefs in some parts of the Pacific and Indian oceans.

NATURAL HISTORY

Reproduction and life cycle. *Types of reproduction.* In most species the sexes are separate; *i.e.*, there are males and females. Reproduction is usually sexual, involving fertilization of eggs by spermatozoa, though several species of holothurians (sea cucumbers), asteroids (starfishes), and ophiuroids (brittle stars) can also reproduce asexually by splitting in half and growing new parts.

In sexual reproduction, eggs and spermatozoa are shed into the water, where the eggs are fertilized. Spawning consists of the shedding of numerous eggs (up to several million) by females and spermatozoa by males. Most echinoderms spawn on an annual cycle, with the spawning period normally lasting one or two months during spring or summer; several species, however, are capable of spawning all year. Spawn-inducing factors are complex and may include external influences such as temperature, light, and salinity of the water; in the case of one Japanese feather star (Crinoidea), spawning is correlated with phases of the moon and takes place during early October, when the moon is in the first or last quarter. Many echinoderms aggregate before spawning, thus ensuring fertilization of eggs. Some also display a characteristic unusual behaviour during the spawning process; some asteroids and ophiuroids raise the centre of the body off the sea floor; holothurians may raise the front end of the body and wave it about.

Types of development. After an egg is fertilized, the development of the resulting embryo into a juvenile echinoderm may proceed in a variety of ways. Small eggs with little yolk material develop into free-swimming larvae that become part of the plankton, actively feed on small organisms for a period of time, then transform, or metamorphose, into juvenile echinoderms and settle to the bottom. Large eggs with greater amounts of yolk material may develop into a larval form that is planktonic but subsists upon its own yolk material, rather than feeding upon small organisms, before eventually transforming into a juvenile echinoderm. Methods of development involving an egg, planktonic larval stages, and a juvenile form are termed indirect. Large eggs with abundant quantities of yolk material may develop directly into juvenile echinoderms, without passing through a larval stage; this type of development is termed direct.

Direct and indirect development

In direct development, the young usually are reared by the female parent. Parental care or brood protection ranges from actual retention of young inside the body of the female until they are born as juveniles to care of the young on the outer surface of the body. Brood protection is best developed among Antarctic, Arctic, and deep-sea echinoderms, in which young may be held around the mouth or on the underside of the parent's body, as in some starfishes (Asteroidea) and sea cucumbers (Holothuroidea), or in special pouches on the upper surface of the body, as in some sea urchins (Echinoidea), sea cucumbers, and asteroids.

During indirect development, the fertilized egg divides many times to produce a hollow ball of cells (gastrula) pushed in at one end to form a primitive gut. Echinoderms resemble vertebrates and some other invertebrate groups in that the hole through which the gut opens to the outside (blastopore) marks the position of the future anus; the mouth is at the opposite end of the body from the blastopore. A pair of subdivided hollow pouches arise from the gut and develop into the body cavity (coelom) and water-vascular system.

The gastrula develops into a basic larval type called a dipleurula larva, characterized by bilateral symmetry; hence the name, which means little two sides. A single band of hairlike projections, or cilia, is found on each side of the body and in front of the mouth and anus. The characteristic larvae found among the living classes of echinoderms (*i.e.*, auricularia, bipinnaria, and pluteus larvae) are modifications of the basic dipleurula pattern.

Larval types

Because the ciliated band of the dipleurula larva of holothurians becomes sinuous and lobed, thus resembling a human ear, the larva is known as an auricularia larva. The dipleurula larva of asteroids develops into a bipinnaria larva, with two ciliated bands, which also may become sinuous and form lobes or arms; one band lies in front of the mouth, the other behind it and around the edge of the body. Echinoids and ophiuroids have complex advanced larvae closely similar in type. The larva, named pluteus, resembles an artist's easel turned upside down and has fragile arms formed by lobes of ciliated bands and supported by fragile rods of calcite, the skeletal material. The echinoid larva (echinopluteus) and the ophiuroid larva (ophiopluteus) usually have four pairs of arms but may have fewer or more. An extra unpaired arm on the plutei of sand dollars and cake urchins extends downward, presumably to help keep the larva up-

From (top) R.R. Shrock and W.H. Twenhofel, *Principles of Invertebrate Paleontology* (1953), reproduced by permission of McGraw-Hill Book Co., Inc.; (bottom) *McGraw-Hill Encyclopedia of Science and Technology* (1960), reproduced by permission of McGraw-Hill Book Co., Inc.

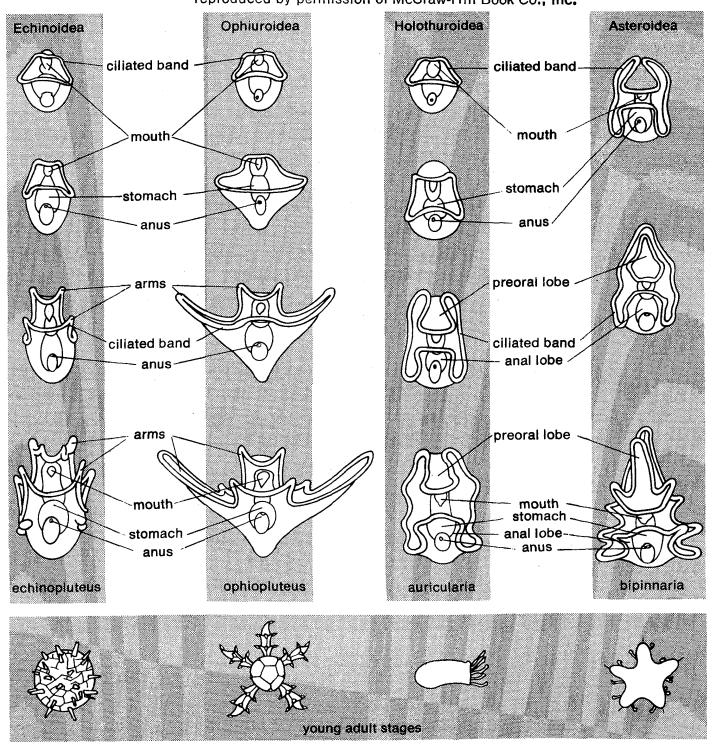


Figure 1: Development of echinoderms from similar early larval stages to diverse forms of young adults.

right. The crinoids (sea lilies), which apparently lack a dipleurula larval stage, have a barrel-shaped larva called a doliolaria larva. The doliolaria larva also occurs in other groups; in holothurians, for example, it is the developmental stage after the auricularia larva, which may not occur in some species. A doliolaria larva usually contains large quantities of yolk material and moves with the aid of several ciliated bands arranged in hoops around the body.

Although most larval stages are small, often less than one millimetre (0.04 inches) in length, some holothurian larvae 15 millimetres (0.6 inches) long have been described, and the length of bipinnaria larvae of some starfishes may exceed 25 millimetres (0.9 inches).

Metamorphosis

After a few days to several weeks as plankton, echinoderm larvae undergo a complex transformation, or metamorphosis, that results in the form known as the juvenile echinoderm. During metamorphosis, the fundamental bilateral symmetry is overshadowed by a radial symmetry dominated by formation of five water-vascular canals. Among holothurians, echinoids, and ophiuroids, the larvae metamorphose as they float, and the young sink to the sea floor; among crinoids and asteroids, however, the larvae attach to the sea floor prior to metamorphosis.

Longevity, growth, and regeneration. The average life span of echinoderms is about four years, and some species may live as long as eight or ten. The rate of growth may depend to a large extent upon factors such as type and abundance of available food and temperature of the surrounding water.

The ability to regenerate, or regrow, lost or destroyed parts is well developed in echinoderms, especially sea lilies (Crinoidea), starfishes (Asteroidea), and brittle stars (Ophiuroidea), which can regenerate new arms if existing ones are broken off. An example of echinoderm regeneration frustrated early attempts to keep starfishes from destroying oyster beds; when captured starfishes were chopped into pieces and thrown back into the sea, they actually increased in numbers. So long as a portion of a body, or disc, remained associated with an arm, new starfishes regenerated. Some sea cucumbers (Holothuroidea) can expel their internal organs (autoeviscerate) under certain conditions (*i.e.*, if attacked, if the environment is unfavourable, or on a seasonal basis); a new set of internal organs regenerates within several weeks. Sea urchins (Echinoidea) readily regenerate lost spines, pincerlike organs called pedicellariae, and small areas of the internal skeleton, or test.

Food and feeding habits. Echinoderms feed in a variety of ways. A distinct feeding rhythm frequently occurs, with many forms feeding only at night, others feeding continuously. Feeding habits range from active, selective predation to omnivorous scavenging or nonselective mud swallowing.

Range of habits

Crinoids feed exclusively on floating plankton, capturing the organisms in a network of mucus produced by soft appendages, called tube feet, contained in grooves on the tentacles, or arms. The arms are spread at right angles to the prevailing current, and small animals are passed to the mouth along the grooves by activity of the cilia and the tube feet.

Many asteroids are active predators on shellfishes and even upon other starfishes; other asteroids are mud swallowers. When feeding, some asteroid species extrude their stomach through the mouth onto the prey, which then is partially digested externally, after which the stomach is eventually retracted and digestion is completed inside the body. Most ophiuroids feed on small organisms, either floating or on the bottom, which are captured by the arms and tube feet and passed toward the mouth. Ophiuroids with arms branched in a complex manner may feed similarly to crinoids.

The more primitive, so-called regular, sea urchins (echinoids) are omnivorous or vegetarian browsers, either scraping algae and other small organisms from rocks with their hard teeth or eating seaweed; several deep-sea regular echinoids feed exclusively on plants carried into the sea from the land. The more advanced echinoids (irregular echinoids), which usually lack teeth, are burrowers and pass small organisms to the mouth with the aid of spines and tube feet. Several species of echinoid sand dollars sometimes feed on suspended organisms carried to them by ocean currents as they lie on the sea floor. Some holothurians (sea cucumbers) remain attached to a surface for indefinite periods of time, capturing plankton in a network of branching, sticky tentacles; others select food from the sea floor and push it into the mouth with their tentacles. A large number of holothurians feed by actively swallowing mud and sand, digesting the organic material, and egesting the waste in the form of characteristic castings, in a manner similar to that of earthworms.

Under artificial conditions, as in aquariums, echinoderms can survive apparent starvation for several weeks at a time. After a holothurian has autoeviscerated, it is unable to feed during the several weeks required for gut regeneration. Echinoderms may derive a significant amount of nourishment, at least for the outer cell layers of the body, from organic material dissolved in seawater.

Locomotion. Asteroids and echinoids, which use spines and tube feet in locomotion, may move forward with any area of the body and reverse direction without turning around. The feet may be used either as levers, by means of which the echinoderm steps along a surface, or as attachment mechanisms that pull the animal. Ophiuroids (brittle stars) tend to move by thrashing the arms in one of several possible methods, including a rowing motion in which strokes are taken by two pairs of extended arms; the fifth arm either is extended forward in the direction in which the animal is travelling or trails behind.

Holothurians (sea cucumbers) generally move with the mouth, or oral, end leading, movement being carried out by both the tube feet and contraction and expansion of the body; sluglike movement is common. Holothurians of the family Synaptidae are able to pull themselves across a surface using their sticky arms as anchors.

Stalked crinoids (sea lilies), so-called because they have stems, generally are firmly fixed to a surface by structures at the ends of the stalks called holdfasts. There is, however, evidence to imply that at least some fossil forms could release themselves to move to new attachment areas. The unstalked crinoids (feather stars) generally swim by thrashing their numerous arms up and down in a coordinated way; for example, in a ten-armed species, when arms 1, 3, 5, 7, and 9 are raised upward, arms 2, 4, 6, 8, and 10 are forcibly pushed downward; then the former groups of arms thrashes downward as the latter is raised. Feather stars that do not swim pull themselves across a surface using their arms.

Swimming

Swimming is known to occur in crinoids, ophiuroids, and holothurians. Some holothurians, formerly regarded as strictly bottom-living forms, are capable of efficient swimming; others, with gelatinous or flattened bodies and reduced calcareous skeletons, appear to spend most of their lives swimming in deep water.

Righting response. Among echinoderms a normal position may be with the mouth either facing a surface, as in asteroids, ophiuroids, and echinoids, or facing away from it, as in crinoids and holothurians. When overturned, echinoderms within a short time exhibit a righting response. Starfishes (asteroids) show this response most effectively, using the tube feet and the arms to perform a slow, graceful somersault that restores their normal position. Sea urchins (echinoids) can roll themselves over by concerted action of their tube feet and spines. The flat sand dollar can turn itself over only by burrowing into the sand until its position is vertical, then toppling over. In more agile groups such as holothurians, crinoids, and ophiuroids, righting is performed with relative ease.

Burrowing. Many echinoderms burrow in rock, soft sediments, or sand; crinoids, however, do not burrow because their feeding apparatus must be kept clear of sediment. Some urchins use the combined abrasive actions of their spines and teeth to burrow several inches into rock, usually in areas of severe wave and tidal action. The most characteristic burrowing echinoderms are the more advanced, or irregular, echinoids, which excavate soft sediments to various depths; most sand dollars burrow just below the surface, and some echinoids (spatangids)

may be found at depths of 38 centimetres (15 inches) or more. Holothurians use tentacles and contraction of the body wall in burrowing that generally is related to feeding. Several asteroid species bury themselves in sandy or muddy areas. The characteristic position of several ophiuroid groups involves burying the body into a surface and leaving only the tips of the arms projecting for food gathering.

Ecology. *Habitats.* Echinoderms are exclusively marine animals, with only a few species tolerating even brackish water; among the exceptions are a few tropical holothurians (sea cucumbers) that can withstand partial drying if stranded on a beach by a receding tide. Most echinoderms cannot tolerate marked changes in salinity, temperature, and light intensity and tend to move away from areas where these factors are not optimal. The behaviour of a large proportion of shallow-water species is regulated by light; *i.e.*, individuals remain concealed during the day and emerge from concealment at night for active feeding. Echinoderms are found in the warmest and coldest of the world's seas; those species that can tolerate a broad temperature range usually also have a broad geographic range. The horizontal or vertical distribution of many species is also governed by water temperature. The influence of pressure upon echinoderms has not yet been thoroughly investigated; pressure is of tremendous physiological importance to marine animals generally, however, and deep-sea dwelling echinoderms probably have special adaptations to their environment.

Influence of external factors

Echinoderms occupy a variety of habitats. Along a rocky shore, starfishes (asteroids) and sea urchins (echinoids) may cling to rocks beneath which sea cucumbers (holothurians) and brittle stars (ophiuroids) are concealed. Some sea urchins have special adaptations (*e.g.*, particularly strong skeletons and well-developed tube feet for attachment) for coping with surf pounding against rocks. In sandy areas starfishes, brittle stars, irregular sea urchins, and sea cucumbers may bury themselves or move on the surface. Large populations of all living groups of echinoderms occur in mud and ooze offshore. In some marine areas, echinoderms are the dominant organism; in the deepest ocean trenches, for example, holothurians may comprise more than 90 percent by weight of the living organisms.

Echinoderms frequently use other animals as homes; thousands of brittle stars, for example, may live in some tropical sponges. Sea cucumbers may attach themselves to the spines of sluggish Antarctic echinoids, and one sea cucumber attaches itself to the skin of a deep-sea fish. On the other hand, echinoderms are also hosts to a wide variety of organisms. Various crustaceans and barnacles, for example, cause the formation of galls, or tumourlike growths, in the skeletons of sea urchins, and crinoids are hosts of specialized parasitic worms. Commensal worms, which do no damage, are associated with most groups; an interesting case of commensalism is the association between various tropical sea cucumbers and the slender pearlfish, which often is found in the rectum of the holothurian, head protruding through its anus. Pinnotherid crabs may be found in the rectum of echinoids and holothurians in Peru and Chile, and highly modified parasitic gastropod mollusks are frequently found in the body cavities of holothurians. A conspicuous parasitic sponge grows on two species of Antarctic ophiuroids.

Predation and defense. Although echinoderm populations do not generally suffer from heavy predation by other animals, ophiuroids form a significant part of the diet of various kinds of fishes and some asteroids. In addition, echinoids (sea urchins) are frequently eaten by sharks, bony fishes, spider crabs, and gastropod mollusks; crows, herring gulls, and eider ducks may either peck their tests (internal skeletons) or drop them repeatedly until they break; and mammals, including the Arctic fox, sea otters, and man, also eat them in considerable numbers. Asteroids are eaten by other asteroids, mollusks, and crustaceans, and some holothurians are eaten by fishes and by man. Crinoids appear to have no consistent predators.

Echinoderms can protect themselves from predation in a variety of ways, most of which are passive. The presence of a firm skeleton often deters predators; echinoids, for example, have a formidable array of spines and, in some cases, highly poisonous stinging pincerlike organs (pedicellariae), both of which may cause intense pain and fever in man. Some asteroids use chemical secretions to stimulate violent escape responses in other animals, particularly predatory mollusks. Some holothurians secrete from the anus a sticky mass of white threads, which may entangle or distract predators; others produce the toxin holothurin, lethal to many would-be predators.

Passive defense mechanisms

Aggregation. Echinoderms tend to aggregate in large numbers and evidently also did so in the past; fossil beds consisting almost exclusively of immense numbers of one or a few species are known from as early as the Lower Cambrian. In present-day seas, ophiuroids may cover large areas of the sea floor; vast aggregations of echinoids are also common. Holothurians, crinoids, and some asteroids also often show a tendency to aggregate.

The phenomenon of aggregation is apparently a response to one or more environmental factors, chief of which is availability of food; *e.g.*, large numbers of ophiuroids and crinoids occupy areas in which strong currents carry large amounts of plankton. An ophiuroid raises some arms into the water to capture food, using other arms to hold on to other nearby ophiuroids; in this way, a large aggregation can maintain its position in an environment in which a single ophiuroid or a small clump of them would be swept away. As stated previously, aggregation also enhances possibilities for successful propagation of a species and possibly may increase chances for survival from predators. It has been postulated both that aggregation is a passive phenomenon resulting from interactions between individuals and the environment and also that it is a demonstration of true social behaviour, a result of interactions among individuals; the question of the cause of the phenomenon has not yet been resolved.

FORM AND FUNCTION

External features. *Symmetry and body form.* Many echinoderms of the Paleozoic Era (about 570,000,000–225,000,000 years ago) either lacked symmetry or had bodies whose halves were mirror images of each other (*i.e.*, were bilaterally symmetrical). Bilateral symmetry, which still occurs in all groups, is especially marked in the larval stages. A tendency toward radial symmetry (the arrangement of body parts as rays), which developed early in echinoderm evolution, eventually became superimposed upon the fundamental bilateral symmetry, often obliterating it. Radial symmetry, usually five-rayed (pentamerous), is conspicuous among all groups of living echinoderms. Although the reasons for the success of radial symmetry are not yet completely clear, it has been postulated that a pentamerous arrangement of skeletal parts strengthens an animal's skeleton more than would, for example, a three-rayed symmetry.

Pentamerous structure is evident in the arrangement of the tube feet, which usually radiate from the mouth to the tips of the arms in five bands. Many of the major organ systems, including the water-vascular system, muscles, hemal system (a complex arrangement of spaces), and parts of the nervous system are also pentamerous. The skeleton follows a pentamerous pattern, except in the case of the holothurians.

The four distinct growth patterns among the echinoderms provide the basis for separating the phylum into four subphyla. Among homalozoans, the pattern is asymmetrical. In crinozoans, bands of tube feet radiate from the mouth, cross the theca (*i.e.*, sheath or calyx), and extend onto the arms; in asterozoans the bands of tube feet radiate outward from the mouth onto the arms to produce a star shape; and in the armless echinozoans, the tube feet form five meridians on the spherical or cylindrical body.

Growth patterns

The crinoid (sea lily, feather star) mouth is centrally located on the theca, from which arise a variable number of arms resembling fern fronds. Although five is the primitive number of arms, they branch once or several times in most living forms to produce 10 to 200 arms.

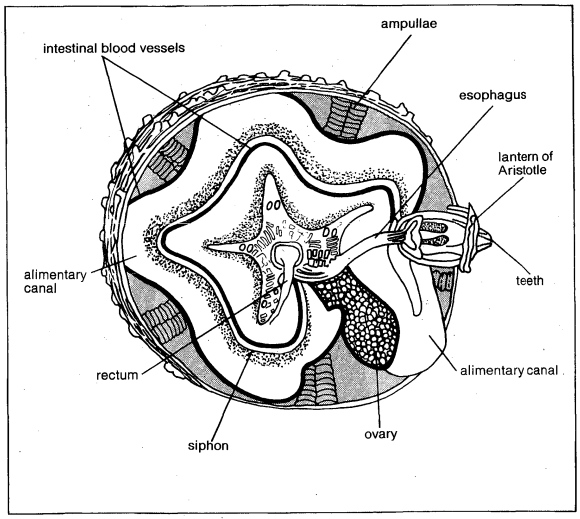


Figure 2: Sea urchin, split so that the oral half is removed, showing alimentary canal and other organs.
From T. Parker and W. Haswell, *A Textbook of Zoology;* St. Martin's Press, Inc., Macmillan & Co., Ltd.

Crinoids either are supported on a stem, or stalk, attached to the underside of the theca or lack stalks, as is the case with most living forms; unstalked crinoids attach themselves by means of slender appendages adapted for grasping (prehensile cirri).

Asteroids (starfishes) have a large central disc from which radiate five or more hollow arms containing parts of the major internal organ systems. The underside (oral surface) of the disc contains a centrally located mouth; the underside of each arm contains five or more bands of tube feet in special grooves called ambulacral furrows. The upper (aboral) surface of the disc has a centrally located anus (often absent) and the sieve plate (madreporite) of the water-vascular system (described below). Seven-armed starfish species are not unusual, and one Antarctic genus may have up to 50 arms.

Ophiuroids (brittle stars) have a small disc from which five arms radiate. The larger internal organs usually are confined to the disc. The centrally located mouth is on the underside of the disc as are the tube feet, which are not arranged in special grooves. Although most ophiuroids have five arms, a few have six or more, and in one group the five arms are branched.

In echinoids (sea urchins) the skeleton forms a rounded, or globular, test of solid plates; tube feet, which emerge through holes in the plates, form five conspicuous bands, or ambulacra. Spaces between bands of feet are termed interambulacra. Regular echinoids are roughly spherical in shape, with a centrally located mouth at the junction of the five bands containing tube feet (ambulacra); the anus is located on the side of the body opposite the mouth (aboral). Irregular urchins are elongated or flattened in shape, with the anus on the oral or aboral surface of the body. In regular and some irregular echinoids, the mouth is equipped with five teeth operated by a complex system of plates and muscles called Aristotle's lantern.

Holothurians (sea cucumbers) are elongated, with mouth and anus at opposite ends of the body. The spaces between the tube feet, which are arranged in five rows, or radii, are known as interradii. The tube feet may be more numerous on the underside of the body than elsewhere, scattered over radii and interradii, or absent. Most holothurians are soft-bodied because the skeleton is reduced and the skeletal units, called ossicles or spicules, are microscopic in size. Holothurians usually show bilateral symmetry outside, radial symmetry inside.

Skeleton. The skeleton is conspicuous in echinoderms, with the exception of the holothurians, forming an effective armour. Each skeletal unit (ossicle) usually consists of two phases, a living tissue (stroma) and a complex lattice (stereom) of mineral calcium carbonate, or calcite, which is derived from the stroma. In living echinoderms, certain properties of calcite are not evident in the stereom because of its latticed structure and the presence of soft stroma; in fossils, however, the stroma may be replaced by secondary calcite, and recognition of fragments of echinoderm skeletons in fossil strata is easier because no other animal group has the same type of skeleton.

Each ossicle is formed from granules which, after secretion from special lime-secreting cells, enlarge, branch, and fuse to build up a three-dimensional network of calcite. Parts of the skeleton enlarge as an animal grows, and resorption and regeneration of the skeleton may occur.

Variety of skeletal structures

Echinoderms have a variety of skeletal structures. In the echinoids, a hollow test (internal skeleton) consisting of ten columns of plates bears large and small spines and the pincerlike organs (pedicellariae) used in defense and to remove unwanted particles from the body. Pedicellariae, also found in the asteroids, are absent from crinoids, ophiuroids, and holothurians. The complex feeding apparatus of echinoids (Aristotle's lantern) consists of 40 ossicles held together by muscles and sutures.

Crinoids (sea lilies) have a hollow sheath (theca or calyx) composed of two or three whorls, each consisting of five skeletal plates; the stalk and the slender appendages (cirri) of unstalked forms consist of a series of drum-shaped ossicles. The asteroid (starfish) skeleton, composed of numerous smooth or spine-bearing ossicles of various shapes, is held together by muscles and ligaments, so that the entire body is flexible. The arms of asteroids are hollow, those of ophiuroids (brittle stars) solid, with the central axis of each arm consisting of elongated ossicles termed vertebrae. The microscopically sized ossicles of holothurians are highly variable in form, ranging from flat lattice plates with holes to exquisitely symmetrical wheels, and are usually numerous; one tropical species, for example, has more than 26,000,000 ossicles in its body wall. A ring of plates, called the calcareous ring, surrounds the tube leading from the mouth to the stomach (*i.e.*, the esophagus) of holothurians. Although located in a similar position to that of the echinoid Aristotle's lantern, the calcareous ring functions as a point of insertion for muscles, not as a feeding apparatus.

Internal features. *Water-vascular system.* The water-vascular system, which functions in the movement of tube feet, is a characteristic feature of echinoderms, and evidence of its existence has been found in even the oldest fossil forms. An internal hydraulic system of canals and reservoirs containing a watery fluid, the system consists of a sieve plate, or madreporite, and a ring vessel, or water-vascular ring, which are connected by a frequently calcified vessel called the stone canal. From the ring vessel, five radial water canals extend outward and give rise to branches that end in the tube feet, which are in contact with the sea. The ring vessel in ophiuroids, asteroids, and holothurians has bulbous cavities called Polian vesicles, which apparently serve to maintain pressure in the system and to hold reserve supplies of fluid; ophiuroids have four or more vesicles, asteroids five, holothurians from 1 to 50. Crinoids lack Polian vesicles, and echinoids have five structures known as either Polian vesicles or spongy bodies.

The madreporite, which is usually located externally, takes in water from outside the body; if internally located, as is the case in many holothurians, fluid is taken from the body cavity. The water or fluid passes from the madreporite to the ring vessel and along the radial canals to the tube feet. The tube feet are extended by contractions of localized muscle areas in the radial canals (ophiuroids) or by contractions of offshoots of the radial canals called ampullae (asteroids, echinoids, and holothurians); the contractions force fluid into the tube feet, which then extend.

The structure of the system varies from group to group; asteroids, for example, frequently have more than one madreporite. Radial canals may lie inward or outward from the skeleton. The tube feet may have well-developed suckers with great holding power, may taper to a point, or be adapted for respiration, feeding, burrow building, mucus production, or sensory perception.

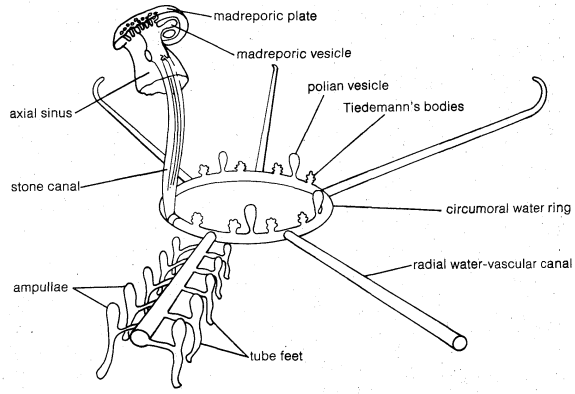


Figure 3: Water-vascular system of an asteroid.
From D. Nichols in R.A. Boolootian (ed.), *Physiology of Echinodermata* (1966); John Wiley & Sons, Inc.

Tube feet

The tube feet of crinoids, which are found in clumps of three on the arms and barblike structures (pinnules) projecting from them, spread a net of sticky mucus that traps small organisms. In ophiuroids, the tube feet are used to gain a hold on a surface and to pass food to the mouth. The numerous tube feet of asteroids are used in locomotion; asteroids with suckerlike tube feet may use them to exert a continuous pull on the valves of shellfish (*e.g.*, oysters, mussels) until muscles holding the valves tire and open slightly, allowing insertion of the asteroid's stomach. Tentacles around the mouth of holothurians are modified tube feet used to capture food; tentacles used to capture plankton are branched and sticky, those used to scoop mud and shovel it into the mouth have a simpler structure.

The tube feet of echinoids serve a variety of functions. The mouth of regular echinoids is surrounded by sensory tube feet, and tube feet further from the mouth are used in locomotion. On the upper side of the body near the anus, the tube feet have respiratory and sensory functions. The tube feet of irregular echinoids, which burrow, are modified in various ways for feeding, burrow construction, and sensory and respiratory functions.

Body wall and body cavity. The body wall (epidermis) contains hairlike projections (cilia) in most echinoderms except ophiuroids; the body wall of crinoids has relatively few. The cilia, capable of producing a waving motion, are used to carry food particles toward the mouth or to remove unwanted particles from the body. The epidermis also contains glandular and sensory cells. The epidermis of skeletal elements such as spines and pedicellariae, which project from the body surface, often is worn away. The extensive body cavity (coelom) is modified to form several specialized regions. Two important subdivisions of the coelom are called the perivisceral coelom and the water-vascular system. The perivisceral coelom is a large, fluid-filled cavity in which the major organs, particularly the digestive tube and sex organs (gonads), are suspended. The water-vascular system was described previously. Other regions of the coelom include the axial sinus (absent from adult holothurians and all echinoids), the madreporic vesicle, and the hyponeural sinus (often called the perihemal system).

Alimentary system and blood systems. The digestive canal consists of a tube, which is almost straight (asteroids and ophiuroids) or coiled in a clockwise direction (crinoids and holothurians). The tube may be divided into esophagus, stomach, intestine, and rectum. Specialized branches of the digestive tube provide enlargement of the digestive surface and may serve other functions; *e.g.*, digestive glands of asteroids, diverticula of echinoids and crinoids, siphons in echinoids, and respiratory trees in holothurians. The anus, absent in ophiuroids and a few asteroids, is present in other groups. The mouth is near the centre of the oral surface, at the point of convergence of the areas containing the tube feet.

The blood system is a complex system of spaces that are neither part of the coelom nor true vessels. A hemal ring and five radial hemal canals surround the esophagus and radial canals of the water-vascular system. A sixth hemal space arises from the hemal ring and enters the axial organ (described below). In addition, a complex network of hemal spaces is associated with the alimentary canal and gonads.

Axial organ. The axial organ, a complex and elongated mass of tissue found in all echinoderms except holothurians, represents the common junction of the perivisceral coelom, the water-vascular system, and the hemal system. Although its functions are not yet well understood, the axial organ plays a part in defense against invading organisms, can contract, is responsible for a circulation of fluids, and may have excretory and secretory activity.

Nervous system and sense organs. The echinoderm nervous system is complex. In all groups, a nerve plexus lies within and below the skin. In addition, the esophagus is surrounded by one to several nerve rings, from which run radial nerves often in parallel with branches of the water-vascular system. Ring and radial nerves coordinate righting activity.

Although echinoderms have few well-defined sense organs, they are sensitive to touch and to changes in light intensity, temperature, orientation, and the surrounding water. The tube feet, spines, pedicellariae, and skin respond to touch, and light-sensitive organs have been found in echinoids, holothurians, and asteroids.

Reproductive system. The masses of sex cells that comprise the gonads of crinoids fill special cavities in the arms or pinnules. Crinoids are the only echinoderms with gonads outside the main body cavity, probably because its volume is reduced. Asteroids typically have ten gonads, two in each arm, which are located near the arm base, appearing as a feathery tuft or a mass of tubules resembling a bunch of grapes. The numerous gonads of some species are arranged in rows along each arm. Ophiuroids have gonads attached to sacs that hang into the body cavity; the sacs, which open outside the body at the bases of the arms, may have one gonad or as many as 1,000 attached.

Number of gonads

Regular echinoids typically have five gonads attached to the interambulacra. A duct from each gonad opens to the exterior near the anus. Most irregular echinoids have four gonads, some have three or five, a few have two; the ducts are on the upper surface of the body. Holothurians differ from all other living echinoderms in having a single gonad, which consists of branching or unbranched tubules; the tubules open into a single duct, which opens to the exterior near the front end of the body. Since many early fossil echinoderms have a single genital opening, or gonopore, it is assumed that these forms also had only one gonad; the condition in holothurians thus is regarded as primitive.

PALEONTOLOGY AND EVOLUTION

Extinct echinoderms. The phylum Echinodermata was already well diversified by the Early Cambrian Period, so that a considerable amount of Precambrian evolution must have taken place. A possible early echinoderm, a Precambrian fossil from Australia with a superficial resemblance to an edrioasteroid (fossil class Edrioasteroidea) had triradial symmetry, from which pentamerous symmetry might have been derived. One problem in the study of echinoderm evolution has been the incomplete nature of the fossil record. Three of the 16 extinct classes of echinoderms now recognized were discovered after 1966, and the decade between 1960 and 1970 saw the discovery of several living fossils; *i.e.*, representatives of groups long thought to be extinct, most notably the somasteroid *Platasterias* and the oegophiuroid *Ophiocanops*. Such data have prompted reassessments and rearrangements of the classification of the phylum.

Relationships among early fossil classes of echinoderms are not clear. In their attachment to the sea floor, certain classes of the subphylum Crinozoa (*e.g.*, eocrinoids, cystoids, paracrinoids, blastoids, and crinoids) are interrelated, and the earliest stem form for these crinozoans probably resembled an eocrinoid. Classes of the fossil subphylum Homalozoa probably represent evolutionary

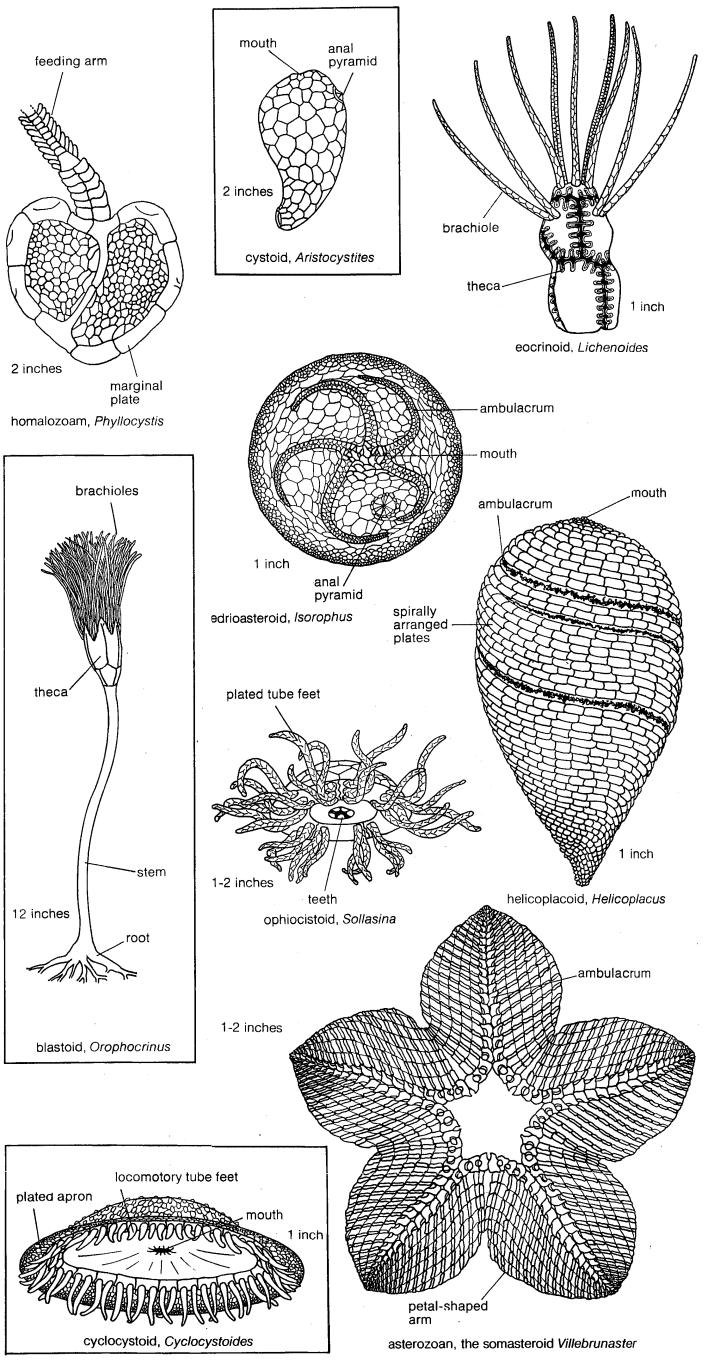


Figure 4: Representative extinct echinoderms. Approximate size is given for each.

From (homalozoan) G. Ubaghs, *Treatise on Invertebrate Paleontology*, part S (1957), University of Kansas and Geological Society of America; (cystoid) J.W. Durham, *Yale Scientific Magazine*, vol. 39, no. 2 (1964); (eocrinoid, blastoid, ophiocistoid, cyclocystoid) D. Nichols in R.A. Boolootian (ed.), *Physiology of Echinodermata* (1966), John Wiley & Sons, Inc.; (edrioasteroid) R.V. Kesling and L. Mirtz, *Contributions from the Museum of Paleontology*, University of Michigan, vol. 15, no. 14 (1960); (helicoplacoid) J.W. Durham and K.E. Caster, "Helicoplacoidea: A New Class of Echinoderms," *Science*, vol. 140, pp. 820–822 (May 17, 1963), copyright 1963 by the American Association for the Advancement of Science; (asterozoan) W.K. Spencer, *Philosophical Transactions of the Royal Society*, B, vol. 235, p. 87 (1951).

end points. The extinct edrioasteroids (subphylum Echinozoa) are possibly related to the echinoids and holothurians; suggested relationships between edrioasteroids and asteroids have not yet been proved.

Conflicting views of relationships

Extant echinoderms. Relationships among the living classes of echinoderms involve two conflicting views. On one hand, many authorities believe that larval stages reflect the evolutionary history of these groups; thus, because echinoids and ophiuroids have pluteus larvae, they appear to form a natural group, holothurians and asteroids forming another. Recent biochemical studies of pigments and other complex substances found in echinoderms also indicate ophiuroid–echinoid and asteroid–holothurian relationships.

On the other hand, comparative anatomy and paleontology contradict these relationships; *i.e.*, early ophiuroids tend to approach asteroids (rather than echinoids) in skeletal structure, while early asteroids show some striking resemblances to crinoids (rather than holothurians) in arm structure. By these criteria, therefore, a crinoid-like ancestor could have given rise to an asteroid–ophiuroid line. Present-day echinoids and holothurians appear to have little in common apart from symmetry; the earliest echinoids with overlapping plates in the test, however, share several features with primitive holothurians, and other fossil groups in the subphylum Echinozoa have structures reminiscent of echinoids and holothurians.

Opponents of the use of larval stages as indicators of evolutionary affinities argue that larvae are unreliable criteria, not only because planktonic larval stages are not typical of echinoderms and are lacking altogether in some species but also because larvae themselves may undergo evolutionary changes. Since paleontological and anatomical evidence found thus far is persuasive, a classification based on adult forms rather than on early larval stages is preferable.

Relationships between echinoderms and other phyla have been studied for more than 100 years, but even with increased knowledge of anatomy and developmental biology the position of echinoderms within the animal kingdom is far from clear. A larval form (tornaria larva) of hemichordates, which resembles the dipleurula larva of echinoderms, has resulted in the conclusion that echinoderms are related to hemichordates and, by assumption, to all chordates. In addition, it has been proposed that some members of the extinct subphylum Homalozoa are chordates, ancestors of the urochordates, cephalochordates, and craniates, but also relatives of the echinoderms. Neither view has been established with certainty.

CLASSIFICATION

Distinguishing taxonomic features. The various subphyla within the Echinodermata are defined in terms of dominant growth patterns. Homalozoa are asymmetrical. Most Crinozoa are attached to the sea floor and have bands of tube feet radiating from the mouth and extending onto the arms. Asterozoa have bands of tube feet radiating from the mouth and extending onto the arms to produce a star shape (divergent growth pattern). Echinozoa lack arms and have tube feet circling the spherically shaped body (meridional growth pattern).

Below the subphylum level, criteria for classification vary, but the basis usually is the skeleton; subdivisions into classes and orders, for example, can be defined in terms of skeletal structure and the arrangement of its component parts.

Annotated classification. The echinoderms once were divided into two great groups, the Pelmatozoa and the Eleutherozoa, the names referring to living habits; *i.e.*, pelmatozoans attach to the sea floor for at least part of their life cycle, unattached eleutherozoans are capable of moving freely over the sea floor. It has been argued that such separation is unnatural because each group contains a mixture of subgroups bearing no relationship to the evolutionary history of the phylum; an alternative classification, presented below, has been established. The terms pelmatozoan and eleutherozoan are now used merely to describe the life habits of echinoderms. Some holothurians (subphylum Echinozoa), for example, have adopted a pelmatozoan habit, attaching themselves to rocks and feeding on plankton; others are eleutherozoan, moving about the sea floor while feeding.

The classification below is essentially that from the *Treatise on Invertebrate Paleontology* (see below *Bibliography*) but has been modified to accommodate recent discoveries. Totally extinct classes, marked with a dagger (†), are known only by fossils.

PHYLUM ECHINODERMATA (echinoderms)

Marine invertebrates worldwide in distribution; living adults with body pattern having structures present in 5s (pentam-

ous); skeleton composed of calcite; internal organs within a body cavity (coelom) that also includes a water-vascular system and tentacles; about 6,000 extant species, about 13,000 extinct species described.

†**Subphylum Homalozoa** (carpoids)

Fossils only (Cambrian to Middle Devonian) about 365,000,000–570,000,000 years ago; without radial symmetry; with fundamentally asymmetrical flattened body.

†*Class Stylophora*

With unique single feeding arm formerly misinterpreted as a stem.

†*Class Homostelea*

With no feeding arm, but with stem composed essentially of 2 series of plates.

†*Class Homoiostelea*

With a feeding arm and a complex stem composed in part of more than 2 series of plates.

†*Class Ctenocystoidea*

With no feeding arm and no stem, but with unique feeding apparatus consisting of a grill-like array of movable plates around mouth.

Subphylum Crinozoa

Fossil and living forms (Lower Cambrian to Recent); radially symmetrical; a partly meridional growth pattern tends to produce a cup-shaped theca (sheath); a partly radially divergent growth pattern produces arms or armlike organs (brachioles) that carry feeding extensions of the radial areas (ambulacra).

†*Class Eocrinoidea*

Lower Cambrian to Silurian about 430,000,000–570,000,000 years ago; body usually consisting of stem, theca, and feeding brachioles.

†*Class Cystoidea*

Lower Ordovician to Devonian about 395,000,000–500,000,000 years ago; with theca and brachioles, with or without stem; numerous plates of theca perforated by numerous thecal pores of a unique type, apparently unconnected with water-vascular system.

†*Class Paracrinoidea*

Middle Ordovician about 460,000,000 years ago; with stem, theca, and arms with barblike structures (pinnules); plates of theca with pore system of unique type.

†*Class Edrioblastoidea*

Middle Ordovician; with stem and theca; no arms or brachioles.

†*Class Parablastoidea*

Lower to Middle Ordovician about 460,000,000–500,000,000 years ago; resemble blastoids (see below, class Blastoidea), but differ in structure of ambulacra and in numbers of thecal plates.

†*Class Lepidocystoidea*

Lower Cambrian; theca and brachioles; no stem; theca covered with numerous overlapping plates.

†*Class Blastoidea*

Silurian to Permian about 280,000,000–430,000,000 years ago; stem, theca with 18–21 plates arranged in 4 rings; numerous feeding brachioles, distinctive infoldings of theca (hydrospires) well developed.

Class Crinoidea (sea lilies and feather stars)

Lower Ordovician to Recent; 700 living species; with, or secondarily without, stem; theca reduced to small cup carrying hollow, usually branching, feeding arms with numerous pinnules; includes fossil orders Inadunata, Flexibilia, and Camerata; living order Articulata; living representatives, stalked sea lilies and unstalked feather stars.

Subphylum Asterozoa

Fossil and living forms (Lower Ordovician to Recent); radially symmetrical with star-shaped body resulting from growth of arms in 1 plane along 5 divergent axes; mouth central, lying on same side of body as tube feet; tube feet, radial water vessels, and other structures restricted to mouth (oral) surface of body.

Class Stelleroidea

Features as subphylum above.

Subclass Somasteroidea. Fossil and living forms (Ordovician to Recent); body flattened; skeleton of arms featherlike (pinnate); transverse growth axes dominant; disc small, arms petal-like; no anus; no pincerlike organs (pedicellariae); 1 suspected living species, *Platasterias latiradiata* in Mexico; rare.

Subclass Asteroidea (starfishes or sea stars). Fossil and living forms (Ordovician to Recent); 1,800 living species; arms broad, hollow; pinnate structure of arms disrupted by dominant longitudinal growth axes; tube feet numerous, carried in deep open grooves roofed by discrete ambulacral ossicles (skeletal units); many groups with suctorial tube feet, pedicellariae, and anus; respiration by interradial gills on aboral surface of body; includes living orders Platyasterida, Paxillosida, Valvatida, Spinulosida, Forcipulatida, Euclasterida, and Zorocallida.

Subclass Ophiuroidea (brittle stars or serpent stars). Fossil and living forms (Ordovician to Recent); 2,000 living species; disc sharply distinct from long, slender arms; longitudinal growth axes dominant; no furrow for tube feet; ambulacral ossicles fused to form vertebrae of arms; no suctorial tube feet; no anus; no pedicellariae; respiration by interradial gills on oral surface of body; includes fossil order Stenurida, living orders Oegophiurida, Phrynophiurida, and Ophiurida.

Subphylum Echinozoa

Fossil and living forms (Lower Cambrian to Recent); radially symmetrical with fundamentally globoid body secondarily cylindrical or discoid; meridional growth patterns, thus projections from body (outspread arms or brachioles) totally absent; nervous, skeletal, muscular, and water-vascular system differentiated into meridional systems but underlying bilateral symmetry generally evident.

†*Class Cyclocystoidea*

Middle Ordovician to Middle Devonian about 375,000,000–460,000,000 years ago; small, disc-shaped; theca composed of numerous plates; ambulacral system with multiple branching.

†*Class Edrioasteroidea*

Lower Cambrian to Lower Carboniferous about 340,000,000–570,000,000 years ago; discoid to cylindrical; five well-developed straight or curved ambulacral food grooves radiate from a central mouth.

†*Class Helicoplacoidea*

Lower Cambrian; pear-shaped or spindle-shaped; spirally pleated contractile flexible internal skeleton (test); columns of plates arranged spirally.

†*Class Ophiocistioidea*

Lower Ordovician to Upper Silurian about 395,000,000–500,000,000 years ago; dome-shaped body partly or completely covered by well-developed test; 5 ambulacral tracts carry plated tube feet relatively enormous in size.

†*Class Camptostromatoidea*

Lower Cambrian; medusa-shaped; body wall plated by several kinds of ossicles; periphery of test with numerous plated arms; symmetry presumed to be radial.

Class Holothuroidea (sea cucumbers)

Fossil and living forms (Ordovician to Recent); 1,100 living species; cylindrical body, elongated orally–aborally, with mouth at or near one end, anus at or near the other; mouth surrounded by conspicuous ring of feeding tentacles; no spines or pedicellariae; single interradial gonad; skeleton usually reduced to form microscopic spicules; includes fossil order Arthrochirotida, living orders Dendrochirotida, Dactylochirotida, Aspidochirotida, Elasipodida, Molpadiida, and Apodida.

Class Echinoidea (sea urchins)

Fossil and living forms (Ordovician to Recent); 900 living species; globular, discoid, or oval in shape, with test of interlocking plates bearing movable spines and pedicellariae; mouth directed downward; anus present; 5 or fewer interradial gonads.

Subclass Perischoechinoidea. Fossil orders Bothriocidaroida, Echinocystoida and Palaechinoida; living order Cidaroida.

Subclass Euechinoidea. Living superorders Diadematacea, Echinacea, Gnathostomata, and Atelostomata. Gnathostomata, and Atelostomata comprise the irregular (more advanced) echinoids; all other groups include regular (more primitive) echinoids.

Critical appraisal. No classification satisfies the requirements of all students of a group. Although many paleontologists and zoologists agree with the higher classification presented here, some disagreement about important points persists. Even though the use of growth patterns as criteria for separating subphyla accommodates most major groups, some problems remain; while crinozoans, for example, are defined as possessing a theca and arms or brachioles, members of class Edrioblastoidea lack both arms and brachioles but have a well-developed theca.

As mentioned previously, several members of the subphylum Homalozoa sometimes are regarded as true chordates rather than echinoderms. Such a classification, how-

ever, would necessitate a reassessment of the entire phylum Echinodermata since the phylum then would not be unique in possessing a latticelike calcite skeleton and a water-vascular system.

In the subphylum Asterozoa, some authors prefer to accord class status to the subclasses. The subclass categories, however, serve to emphasize the close relationships between somasteroids and asteroids on the one hand and somasteroids, asteroids, and ophiuroids on the other. Within the class Asteroidea the orders Paxillosida and Valvatida are often combined into the order Phanerozonida. The recently defined orders Euclasterida and Zorocallida have separated several troublesome genera from the ill-defined order Forcipulatida.

The subphylum Echinozoa includes the classes Cyclocystoidea, Camptostromatoidea, and Helicoplacoidea; although none displays any trace of pentamerous symmetry, the cyclocystoids and camptostromatoids apparently are radially symmetrical. The extinct class Edrioasteroidea includes many members that apparently were permanently attached to the sea floor by a stem, thus differing from other echinozoans and approaching the crinozoans.

Although there is some agreement that the irregular urchins (sand dollars, cake urchins) have been derived from several sources (*i.e.*, are polyphyletic), polyphyly has not yet been proved, and it has been suggested that subdivision of the Echinoidea into two subclasses (Regularia and Irregularia) should be reintroduced.

Study of the higher classification of echinoderms is an active field; undoubtedly the classification will continue to be revised as new emphases and discoveries are developed.

BIBLIOGRAPHY. R.A. BOOLOOTIAN (ed.), *Physiology of Echinodermata* (1966), a comprehensive review of biology and physiology; A.M. CLARK, *Starfishes and Their Relations* (1962), an excellent introductory work, with emphasis on living forms, classification, and biology; L.H. HYMAN, *The Invertebrates*, vol. 4, *Echinodermata* (1955), a classic survey of anatomy, biology, and classification; N. MILLOTT (ed.), *Echinoderm Biology* (1967), a collection of specialized review articles; R.C. MOORE (ed.), *Treatise on Invertebrate Paleontology*, pt. S and U, "Echinodermata," vol. 1 and 3 (1966–67), an authoritative treatment of paleontology and classification, pt. T (crinoids) in prep.; D. NICHOLS, *Echinoderms*, 4th ed. (1969), an excellent up-to-date general work, including fossil forms.

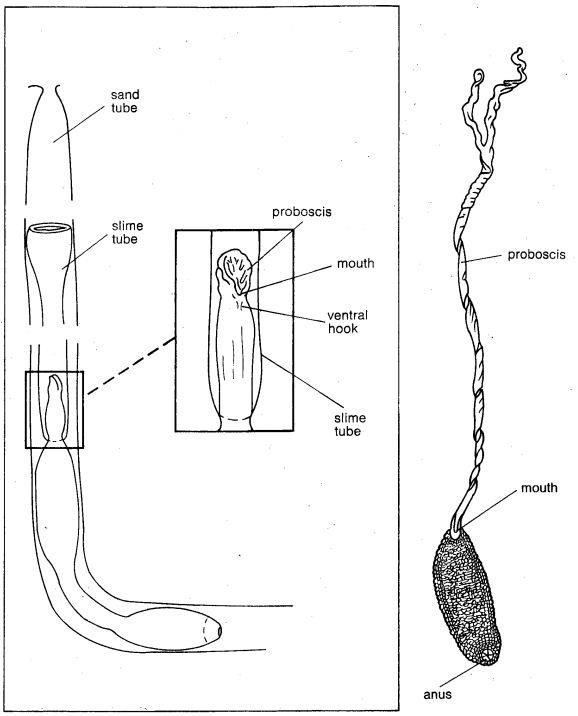
(D.L.P.)

Echiura

The phylum Echiura, also known as Echiuroidea or Echiurida, or the spoonworms, are sausage-shaped organisms with a flattened extension of the "head" that is curved along its lateral edges and sometimes shaped like a scoop or spoon to form a nonretractable anterior proboscis. Adult spoonworms vary from a few millimetres to 600 millimetres in length. In *Urechis* the proboscis is only a fraction of the length of the trunk; however, in *Ikeda* the proboscis may exceed 10 times the trunk length and measure more than one metre.

Apart from a few brackish-water species, echiurids are marine and occur in seabeds throughout the world. Although some species inhabit rock crevices, most live in burrows in the mud. A flow of water is maintained through the burrow by peristaltic contractions of the body. Some respiration probably takes place across the body wall, but the hind gut is often utilized as a respiratory organ by drawing water through the anus.

Life cycle. The sexes are separate and, apart from members of the family Bonelliidae, superficially alike. Ova or spermatozoa produced by an unpaired gonad and shed into the coelomic fluid contained in the body cavity are nourished there until maturation is complete. The mature gametes are taken up by ciliated funnels of the anterior nephridia, which lose their excretory function. Then the gametes pass into thin-wall nephridial sacs and are stored there until their expulsion into the sea where fertilization takes place. Spiral cleavage occurs in the zygote; a free-swimming trochophore larva develops and metamorphoses into a burrowing worm.



Echiurids.
(Left) *Urechis caupo.* (Right) *Bonellia viridis.*
From Paul A. Meglitsch, *Invertebrate Zoology*, copyright © 1967 by Oxford University Press, Inc.; reprinted by permission

Species of the family Bonelliidae are unusual in that the male (only one to six millimetres long) is parasitic within a modified nephridium (or uterus) of a female. A male develops when a larva comes in contact with a female's proboscis, which secretes a hormone that affects larval development. The larva passes to the esophagus and undergoes reduced metamorphosis to a sexually mature male; however, the proboscis and circulatory system fail to develop and only a reduced digestive system forms. Cilia cover the entire body (as with the larva), yet a system of gonoducts develops fully. The male remains within the egg-storage organ of the female, presumably to fertilize the eggs as they are shed into the water.

Form and function. The trunk of spoonworms is sac-like, with a body wall composed of three muscle layers. Anteriorly is a muscular but noncontractile, tactile proboscis, which is highly mobile and capable of great extension in some species. The mouth is situated at the base of the proboscis. Typically, there is a pair of setae on the ventral surface behind the mouth and one or two circles of setae around the posteriorly located anus. The body cavity is filled with coelomic fluid containing two types of cells; in some species, one type of cell contains a respiratory pigment similar to hemoglobin. The alimentary canal is coiled and associated along much of its length with a siphon, or collateral intestine, whose function is obscure. A pair of anal vesicles with an excretory function arises from the end of the intestine. Single to numerous nephridia are attached to the ventral body wall. Most species have a closed vascular system comprising dorsal, ventral, and neurointestinal blood vessels. The nerve cord is ventral. The gonads arise from a mesentary above the nerve cord or from others near the cloaca. They are diffuse, and eggs or spermatozoa are shed into the coelomic fluid.

Feeding

Feeding methods vary according to the size of the proboscis. Species with a long proboscis reach out and pick up detritus around the burrow entrance; the material collected becomes attached to strings of mucous and is transferred to the mouth. In species with a short proboscis, a mucous tube is secreted and forms a lining to the burrow. Food particles drawn into the burrow with the flow of water for respiration become fixed to the mucous tube. When this is filled, the tube and its contents are swallowed.

Classification and phylogeny. *Distinguishing taxonomic features.* Although the form of the proboscis and the number of setae present on the body are important, the most significant features are internal and include musculature, vascular system, and number of nephridia. Spoonworms were earlier linked with the peanutworms (see SIPUNCULA), then Adam Sedgewick assigned them to the phylum Annelida. They were not widely accepted as Annelida; and, after W.W. Newby (1940) studied the embryology and development of *Urechis caupo*, they became established as a separate phylum.

Annotated classification. The following classification from A.C. Stephen and S.J. Edmonds (1972) incorporates earlier views.

PHYLUM ECHIURA (spoonworms)

Unsegmented, coelomate, bilaterally symmetrical marine invertebrates; muscular, saclike, or subcylindrical body with nonretractable muscular proboscis; mouth situated anteriorly near base of proboscis, anus posterior at extremity of body; setae often present, number one or two on ventral surface, occasionally one or two rings form around anus; one, two, or many nephridia in body cavity; pair of anal vesicles arise from cloaca; about 130 species described.

Order Echiuroinea

Longitudinal musculature of body lies between a circular outer layer and an oblique inner layer of muscle; vascular system present; posterior intestine not specialized as respiratory organ; not more than 20 pairs of nephridia.

Family Bonelliidae

Sexually dimorphic; female with saclike trunk, proboscis often bifid; nephridia one or two; ventral setae two or absent, anal setae absent; anal vesicles usually branched; male small, degenerate, parasitic in or on female.

Family Echiuridae

Sexes alike; proboscis never bifid; nephridia paired; one pair ventral setae, anal setae usually absent; anal vesicles mainly saclike.

Order Xenopneusta

Longitudinal musculature of body wall between circular outer layer and oblique inner layer of muscles; vascular system absent; posterior intestine modified as respiratory organ; nephridia, two or three pairs; one family, Urechidae.

Order Heteromyota

Longitudinal musculature of body wall outside both circular and oblique muscle layers; nephridia unpaired, numerous (up to 400); one family, Ikedaidae.

Critical appraisal. The affinities of the spoonworms are obscure. Although they share a few features with the Annelida (*i.e.*, spiral cleavage in embryo, trochophore larva, signs of segmentation in nervous system during early development, setae in all but degenerate bonellid males), the evidence is too slender to suggest a close affinity. The peanutworms (phylum Sipuncula), with which the spoonworms were once united, are bilaterally symmetrical, coelomate, unsegmented worms, but the two groups have few other characters in common. Fossil spoonworms have been recorded from the Cretaceous.

BIBLIOGRAPHY. R.B. CLARK, "Systematics and Phylogeny: Annelida, Echiura and Sipuncula," *Chemical Zoology*, 4:1–68 (1970), physiological approach; C.N. DAWYDOFF, "Classe des Echiuriens," in P.P. GRASSE, *Traité de zoologie*, 5:855–904 (1959), an excellent summary of the range of morphological variation, development, and physiology; W.K. FISHER, "Additions to the Echiuroid Fauna of the North Pacific Ocean," *Proc. U.S. Natn. Mus.*, 99:479–497 (1952), "Echiuroid Worms of the North Pacific Ocean," *ibid.*, 96:215–292 (1948), and "New Genera and Species of Echiuroid and Sipunculoid Worms," *ibid.*, 97:351–372 (1950); G.E. and N. MACGINITIE, *Natural History of Marine Animals*, 2nd ed. (1968), ecological viewpoint; A.C. STEPHEN and S.J. EDMONDS, *The Phyla Sipuncula and Echiura* (1972), a comprehensive monograph containing descriptions of the known species, a discussion of systematics, and an exhaustive bibliography.

(R.W.S.)

Eckehart, Meister

The greatest German speculative mystic, Meister Eckehart was the founder of the school of "Rhineland Mystics" and the ancestor of German Protestantism, Romanticism, Idealism, and Existentialism. He was essentially a preacher, and his most significant works are transcripts from sermons he delivered in medieval German. Some of his bold expressions about the soul's union with God were condemned by the Inquisition. Although his philosophy amalgamates Greek, Neoplatonic, Arabic, and Scholastic elements, it is unique. His doctrine, sometimes abstruse, always arises from one simple personal mystical experience to which he gives a number of names. By doing so, he was also an innovator of the German language, contributing many abstract terms. The National Socialist movement in Germany in the 1930s made him one of its intellectual pioneers, by blatantly falsifying his teaching. Today there is a great interest in Eckehart among some Marxist theorists and Zen Buddhists.

Johannes Eckehart (or Eckhart) was born about 1260 at Hochheim, Thuringia. He entered the Dominican Order when he was 15 and studied in Cologne, perhaps under the Scholastic philosopher Albert the Great. The intellectual background there was influenced by the great Dominican theologian Thomas Aquinas, who had just died. In his mid-30s, Eckehart was nominated vicar (the main Dominican official) of Thuringia. Before and after this assignment he taught theology at Saint-Jacques' priory in Paris. It was also in Paris that he received his master's degree (1302) so that always after he was called Meister Eckehart.

German works

He wrote four works in German (usually called "treatises"). At about the age of 40 he wrote the *Talks of Instruction*, on self-denial, the nobility of will and intellect, and the obedience to God. At the same period, he faced the Franciscans in some famous disputations on theological issues. In 1303, he became provincial (leader) of the Dominicans in Saxony and, three years later, vicar of Bohemia. His main activity, especially from 1314, was preaching to the contemplative nuns established throughout the Rhine valley. He resided in Strassburg as a prior.

The best attested German work of this middle part of his life is the *Book of Divine Consolation* dedicated to the Queen of Hungary. The other two treatises were *The Nobleman* and *On Detachment.* The teachings of the mature Eckehart describe four stages of the union between the soul and God: dissimilarity, similarity, identity, breakthrough. At the outset, God is all, the creature is nothing; at the ultimate stage, "the soul is above God." The driving power of this process is detachment (*Gelassenheit* or *Abgeschiedenheit*).

1. Dissimilarity: "All creatures are pure nothingness. I do not say they are small or petty: they are pure nothingness." Whereas God inherently possesses being, creatures do not possess being but receive it derivatively. Outside God, there is pure nothingness. "The being (of things) is God." The "noble man" moves among things in detachment, knowing that they are nothing in themselves and yet aware that they are full of God—their being.

2. Similarity: Man thus detached from the singular (individual things) and attached to the universal (Being) discovers himself to be an image of God. Divine resemblance, an assimilation, then emerges: the Son, image of the Father, engenders himself within the detached soul. As an image, "thou must be in Him and for Him, and not in thee and for thee."

3. Identity: Eckehart's numerous statements on identity between God and the soul can be easily misunderstood. He never has substantial identity in mind, but God's operation and man's becoming are considered as one. God is no longer outside man, but he is perfectly interiorized. Hence such statements: "The being and the nature of God are mine; Jesus enters the castle of the soul; the spark in the soul is beyond time and space; the soul's light is uncreated and cannot be created, it takes possession of God with no mediation; the core of the soul and the core of God are one."

4. Breakthrough: To Meister Eckehart, identity with God is still not enough; to abandon all things without abandoning God is still not abandoning anything. Man must live "without why." He must seek nothing, not even God. Such a thought leads man into the desert, anterior to God. For Meister Eckehart, God exists as "God" only

when the creature invokes him. Eckehart calls "Godhead" the origin of all things that is beyond God (God conceived as Creator). "God and the Godhead are as distinct as heaven and earth." The soul is no longer the Son. The soul is now the Father: it engenders God as a divine person. "If I were not, God would not be God." Detachment thus reaches its conclusion in the breakthrough beyond God. If properly understood, this idea is genuinely Christian: it retraces, for the believer, the way of the Cross of Christ.

These teachings are to be found in his Latin works too. But the Latin *Sermons, Commentaries on the Bible,* and *Fragments* are more Scholastic and do not reveal the originality of his thought. Nevertheless, Eckehart enjoyed much respect even among scholars. In his 60th year he was called to a professorship at Cologne. Heinrich von Virneburg—a Franciscan, unfavorable to Dominicans, anyway—was the archbishop there, and it was before his court that the now immensely popular Meister Eckehart was first formally charged with heresy. To a list of errors, he replied by publishing a Latin *Defense* and then asked to be transferred to the pope's court in Avignon. When ordered to justify a new series of propositions drawn from his writings, he declared: "I may err but I am not a heretic, for the first has to do with the mind and the second with the will!" Before judges who had no comparable mystical experience of their own, Eckehart referred to his inner certainty: "What I have taught is the naked truth." The bull of Pope John XXII, dated March 27, 1329, condemns 28 propositions extracted from the two lists. Since it speaks of Meister Eckehart as already dead, it is inferred that Eckehart died some time before, perhaps in 1327 or 1328. It also says that Eckehart had retracted the errors as charged.

Heresy charge

BIBLIOGRAPHY. JAMES M. CLARK and JOHN V. SKINNER (eds. and trans.), *Treatises and Sermons* (1958), a good selection and translation from Latin and German, with an introduction and notes; C. DE B. EVANS (trans.), *Meister Eckhart by Franz Pfeiffer,* 2 vol. (1924–31), the most complete translation, based on an obsolete German edition (some of the sermons not authentic); JEANNE ANCELETHUSTACHE, *Maître Eckhart et la mystique rhénane* (1956; Eng. trans., *Master Eckhart and the Rhineland Mystics,* 1957), a short, excellent, illustrated introduction to his life and work; VLADIMIR LOSSKY, *Théologie négative et connaissance de Dieu chez Maître Eckhart* (1960), voluminous and erudite research on Eckehart's Latin and Greek sources; JAMES M. CLARK, *Meister Eckhart: An Introduction to the Study of His Works* (1957), a good exposition of Eckehart's thinking; INGEBORG DEGENHARDT, *Studien zum Wandel des Eckhartbildes* (1967), a résumé of six centuries' outlooks on Meister Eckehart; REINER SCHURMANN, *Maître Eckhart ou la Joie errante* (1971), analyses of three sermons and a comparison of Eckehart and Heidegger.

(R.Sch.)

Eclipse, Occultation, and Transit

Eclipse, occultation, and transit are related astronomical events in which three celestial bodies are aligned; the body passing between the other two obscures some or all of the light of one of these as seen from the other. The difference in terminology is related to the relative sizes of the obscuring and obscured bodies.

The word eclipse comes from the Greek *ekleipsis* "abandonment." The Sun is eclipsed when the Moon comes between it and the Earth; the Moon is eclipsed when it moves into the shadow of the Earth cast by the Sun. Eclipses of natural or artificial satellites of a planet occur as the satellites move into the planet's shadow. The two component stars of an eclipsing binary star move around each other in such a way that their orbital plane passes through or very near the Earth, and each star periodically eclipses the other as seen from the Earth.

When the apparent size of the eclipsed body is much smaller than that of the eclipsing body, the phenomenon is known as an occultation. Examples are the disappearance of a star, nebula, or planet behind the Moon, or the vanishing of a natural satellite or space probe behind some body of the solar system.

A transit occurs when, as viewed from the Earth, a relatively small body passes across the disk of a larger body, usually the Sun or a planet, eclipsing only a very small area: Mercury and Venus periodically transit the Sun, and a satellite may transit its planet.

PHENOMENA OBSERVED DURING ECLIPSES

Lunar eclipse phenomena. The Moon may, when full, enter the shadow of the Earth. The motion of the Moon around the Earth is from west to east relative to the Sun, so that, for an observer facing south, the shadowing of the Moon begins at its left edge (if the Moon were north of the observer, as, for example, in parts of the Southern Hemisphere, the opposite would be true). If the eclipse is a total one and circumstances are favourable, the Moon will pass through the umbra, or darkest part of the shadow, in about two hours (see Figure 1). During this time, the Moon is usually not quite dark. A part of the sunlight, especially the redder light, penetrates the Earth's atmosphere, is refracted into the shadow cone, and reaches the Moon. Meteorological conditions on Earth strongly affect the amount and colour of light that can penetrate the atmosphere. Generally, the totally eclipsed Moon is clearly visible and has a reddish-brown, coppery colour, but the brightness varies strongly from one eclipse to another.

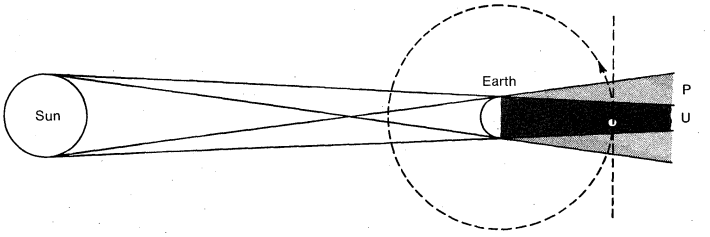


Figure 1: *Eclipse of the Moon.*
The moon revolving in its orbit around the Earth passes through the shadow of the Earth. U (umbra) is the total shadow, P (penumbra) the partial shadow.

After the Moon leaves the umbra, it must still pass through the penumbra of the shadow. When the border between umbra and penumbra is visible on the Moon, the border is seen to be part of a circle, the projection of the circumference of the Earth. This is a direct proof of the spherical shape of the Earth. Because of the Earth's atmosphere, the edge of the umbra is rather diffuse, and the times of contact between the Moon and the umbra cannot be observed accurately. The progress of a lunar eclipse can be adequately followed with the naked eye, field glasses, or a small telescope.

During the eclipse the surface of the Moon cools, at a rate dependent on the constitution of the lunar soil, which is not everywhere the same. Many spots on the Moon sometimes remain brighter than their surroundings during totality—particularly in their output of infrared radiation—possibly because their heat conductivity is less, but the cause is not fully understood.

Cooling of the Moon during eclipses

An eclipse of the Moon can be seen under similar conditions at all places on the Earth where the Moon is above the horizon.

Solar eclipse phenomena. Totality at any particular solar eclipse can only be seen from a relatively narrow belt on Earth. The various phases observable at a total solar eclipse are illustrated in Figure 2A. "First contact" designates the moment when the disk of the Moon, invisible against the bright sky background, just touches the disk of the Sun. The partial phase of the eclipse then begins, as a small indentation in the western rim of the Sun becomes noticeable. The dark disk of the Moon now gradually moves across the Sun's disk, and the bright area of the Sun is reduced to a crescent. The sunlight, shining through gaps in foliage and other small openings, is then seen to form little crescents of light that are images of the light source, the Sun. Toward the beginning of totality, the direct light from the Sun diminishes very quickly and the colour changes. The sky becomes dark but, along the horizon, the Earth's atmosphere still appears bright because the umbra of the Moon's shadow on the Earth extends over a rather narrow region. The scattered light coming in from a distance beyond this region produces weird effects. Men, animals, and birds re-

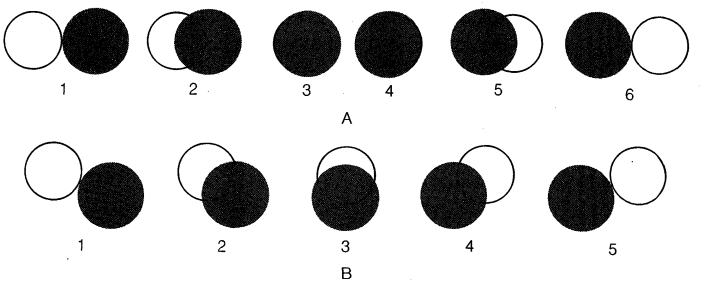


Figure 2: *Successive phases of a solar eclipse.* The dark disk of the Moon gradually moves across the disk of the Sun from west (right) to east (left). (A) Total eclipse: (1) first contact; (2) partial phase; (3) second contact, beginning of totality; (4) third contact, end of totality; (5) partial phase; (6) fourth contact. (B) Partial eclipse: (1) first contact; (3) maximum phase; (5) last contact.

act with fear; birds may go to roost as they do at sunset.

As the tiny, narrow crescent of sunlight disappears, little bright specks remain where depressions in the Moon's edge, the limb, are last to obscure the Sun's limb. These specks are known as Baily's beads, after the 18th-century English astronomer Francis Baily, who first drew attention to them. The beads vanish at the moment of second contact, when totality sets in. This is the climax of the eclipse. The reddish prominences and chromosphere of the Sun, around the Moon's limb, can now be seen. The brighter planets and stars become visible in the sky. The white corona extends out from the Sun to a distance of more than a sun's diameter, at which point it fades into invisibility. The temperature in the path of totality falls by some degrees. The light of totality is considerably brighter than that of the Full Moon, but it is quite different (see below).

The moment of third contact approaches, at which time many of the phenomena of second contact appear again in reverse order. Suddenly the first Baily's bead appears, now on the other side of the Moon. More beads of light follow, the Sun's crescent grows again, the corona disappears, and daylight brightens. The thin crescent of the Sun gradually widens, and about one and a quarter hours later the eclipse ends with fourth contact, when the last encroachment made by the Moon on the Sun's rim disappears.

During the partial phase, both before and after totality, it is necessary to protect the eyes against injury by the intense brilliance of the Sun. It should be viewed directly only through strong filters, a dark smoked glass, or a heavily fogged photographic plate or film.

Shadow bands

When totality is imminent and only a small crescent of the Sun remains, the so-called shadow bands can often be seen on plain light-coloured surfaces, such as open floors and walls. These are striations of light and shade, moving and undulating, several centimetres (or inches) wide. Their velocity and direction depend on air currents at various heights, as they are caused by refraction of sunlight by small inhomogeneities in the Earth's atmosphere. A similar phenomenon is the projection of water waves on the bottom of a sunlit swimming pool or bath.

THE GEOMETRY OF ECLIPSES, OCCULTATIONS, AND TRANSITS

Eclipses of the Sun. An eclipse of the Sun takes place when the Moon comes between the Earth and the Sun so that the Moon's shadow sweeps over the face of the Earth (see Figure 3). This shadow consists of two parts: the umbra, or total shadow, a cone into which no direct sunlight penetrates; and the penumbra, or half shadow, which is reached by light from only a part of the Sun's disk.

To an observer within the umbra, the Sun's disk appears completely covered by the disk of the Moon; such an eclipse is called total. To an observer within the penumbra, the Moon's disk appears projected against the Sun's disk so as to overlap it partly; the eclipse is then called partial for that observer. The umbra cone is narrow at the distance of the Earth, and a total eclipse is observable only within the narrow strip of land or sea over which the umbra passes. A partial eclipse may be seen from places within the large area covered by the penumbra. Sometimes the Earth intercepts the penumbra of the Moon but is missed by its umbra; only a partial eclipse of the Sun is then observed anywhere on the Earth.

By a remarkable coincidence, the sizes and distances of the Sun and Moon are such that they appear as very nearly the same angular size (about 0.5°) at the Earth, but their apparent sizes depend on their distances from the Earth. The Earth revolves around the Sun in an elliptical orbit, so that the distance of the Sun changes slightly during a year, with a correspondingly small change in the apparent size, the angular diameter, of the solar disk. In a similar way, the apparent size of the Moon's disk changes somewhat during the month because the Moon's orbit is also elliptical. When the Sun is nearest to the Earth and the Moon is at its greatest distance, the apparent disk of the Moon is smaller than that of the Sun. If an eclipse occurs at this time, the Moon's disk passing over the Sun's disk cannot cover it completely but will leave the rim of the Sun visible all around it. Such an eclipse is said to be annular. Total and annular eclipses are called central.

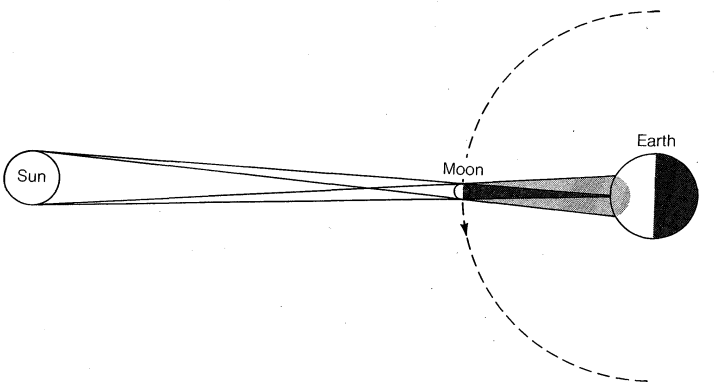


Figure 3: *Eclipse of the Sun.* The shadow of the Moon sweeps over the surface of the Earth. In the darkly shaded region (umbra) the eclipse is total; in the lightly shaded region (penumbra) the eclipse is partial. The shaded region on the opposite side of the Earth indicates the darkness of night. (Dimensions of bodies and distances are not to scale.)

In a partial eclipse (Figure 2B), the centre of the Moon's disk does not pass across the centre of the Sun. After the first contact, the visible crescent of the Sun decreases in width until the centres of the two disks reach their closest approach. This is the moment of maximum phase, and the extent is measured by the ratio between the smallest width of the crescent and the diameter of the Sun. After maximum phase, the crescent of the Sun widens again until the Moon passes out of the Sun's disk at the last contact.

Eclipses of the Moon. When the Moon moves through the shadow of the Earth (see Figure 1) it loses its bright direct illumination by the Sun, although its disk still remains faintly visible. As the shadow of the Earth is directed away from the Sun, a lunar eclipse can occur only at the time of Full Moon—that is, when the Moon is on the side of the Earth opposite to that of the Sun. A lunar eclipse appears much the same at all points of the Earth from which it can be seen. When the Moon enters the penumbra, a penumbral eclipse occurs. The dimming of the Moon's illumination by the penumbra is so slight as to be scarcely noticeable, and penumbral eclipses are rarely watched. After a part of the Moon's surface is in the umbra and thus darkened, the Moon is said to be in partial eclipse. After about an hour, when the whole disk of the Moon is within the umbra, the eclipse becomes total. If the Moon's path leads through the centre of the umbra, the total eclipse lasts about an hour and three-quarters.

Penumbral eclipse

Eclipses, occultations, and transits of satellites. These phenomena are conveniently illustrated by the four larg-

est satellites of Jupiter, whose eclipses provide a frequently occurring and fascinating spectacle to the telescopic observer. The three innermost moons (Io, Europa, and Ganymede) disappear into the shadow of Jupiter at each revolution, though the fourth (Callisto) is not eclipsed every time. Because of the sizeable dimensions of these bodies, some minutes elapse between first contact with the shadow and totality. The orbits of these satellites lie nearly in the same plane as Jupiter's orbit around the Sun, and, at practically every revolution of each satellite, the following four eclipse phenomena take place: (1) eclipse of the satellite when it passes through Jupiter's shadow; (2) occultation of the satellite when it disappears behind the planet, as seen from the Earth; (3) transit of the satellite across the disk of Jupiter; (4) transit of the shadow of the satellite across the planet's disk.

Figure 4 illustrates these phenomena; it shows Jupiter

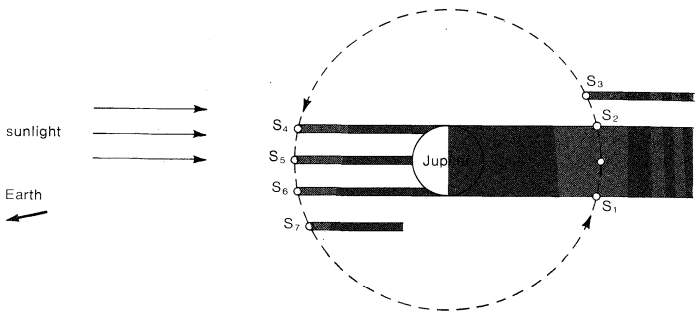


Figure 4: *Eclipses of the satellites of Jupiter.* S_1–S_7 mark successive positions of one of these satellites as it revolves in its orbit around Jupiter.

and the orbit of one of its satellites, the direction of the sunlight illuminating the system, and the direction toward the Earth, from where the observation is made. When the satellite arrives at the point S_1 of its orbit, it enters Jupiter's shadow (eclipse) and vanishes. At S_2 it comes out of the shadow, but, to the terrestrial observer, it is now hidden behind the planet (occultation) until at S_3 it reappears at the limb. When the satellite reaches the position S_4, its shadow falls on Jupiter causing a small dark spot on its surface. Seen from the Earth, the satellite is to the left of Jupiter approaching Jupiter's limb, at the time that its shadow spot passes across the planet's disk (transit of shadow). At S_5 the satellite starts to pass in front of the planet (transit of satellite), following its shadow spot. Both Jupiter and the satellite must have their illuminated sides facing the Earth. They differ little in total surface brightness; near the limb the satellite is somewhat brighter than the planet's surface on which it appears projected, but near the middle of the disk it is hardly distinguishable. At S_6 the shadow leaves the planet, and at S_7 the satellite emerges at the limb.

Proof of finite speed of light

Historically, the eclipses of Jupiter's satellites are important, for they provided one of the earliest proofs of the finite speed of light. It is possible to calculate with considerable precision the times of disappearance and reappearance of a satellite undergoing eclipse. The Danish astronomer Ole Rømer in 1675 noticed discrepancies between the observed and calculated times of such eclipses, which he correctly explained as being due to the difference in the travel time of light when the Earth is nearest to Jupiter or farther away from it.

A related phenomenon is the occultation of a space probe by a planet. During the beginning and the end of such an occultation, signals sent out by the spacecraft and received on Earth have penetrated the planet's atmosphere and can yield information about atmospheric density and composition.

THE FREQUENCY OF SOLAR AND LUNAR ECLIPSES

A solar eclipse, especially a total one, can be seen from only a limited part of the Earth, while the eclipsed Moon can be seen at the time of the eclipse wherever the Moon is above the horizon.

In most calendar years there are two lunar eclipses; in some years one or three or none occur. Solar eclipses occur two to five times a year, five being exceptional; there were five in 1935 and will be again in 2206. The average number of total solar eclipses in a century is 66 for the Earth as a whole.

Numbers of solar eclipses predicted to take place during the 20th to 25th centuries are:

1901–2000:	228	eclipses,	of which	145	are	central
2001–2100:	224	"	"	144	"	"
2101–2200:	235	"	"	151	"	"
2201–2300:	248	"	"	156	"	"
2301–2400:	248	"	"	160	"	"
2401–2500:	237	"	"	153	"	"

Any point on Earth may, on the average, experience no more than one total solar eclipse in three to four centuries. The situation is quite different for lunar eclipses. An observer remaining at the same place (and granted cloudless skies) can see 19 or 20 lunar eclipses in 18 years: three or four total eclipses and six or seven partial eclipses may be visible from beginning to end, and five total eclipses and four or five partial eclipses at least partially visible. All these numbers can be worked out from the geometry of the eclipses. A total lunar eclipse can last as long as an hour and three-quarters, but for a solar total eclipse maximum duration of totality is only 7½ minutes. This difference results from the fact that the Moon is much smaller in cross-section than the extension of the Earth's shadow but can be only a little greater in apparent size than the Sun.

Cycles of eclipses. The eclipses of the Sun and Moon occur at New Moon and Full Moon, respectively, so that one basic time period involved in the occurrence of eclipses is the synodic month; *i.e.*, the time of one revolution of the Moon around the Earth with respect to the Sun.

A solar eclipse does not occur at every New Moon, because the Moon's orbit plane is inclined to the ecliptic, the plane of the orbit of the Earth around the Sun. The angle between the planes is about 5°; thus the Moon can pass well above or below the Sun. The line of intersection of the planes is called the line of the nodes, being the two points where the Moon's orbit intersects the ecliptic plane. The ascending node is the point where the Moon crosses the ecliptic from south to north, and the descending node that where it goes from north to south. The nodes move along the orbit from west to east, going completely around the ecliptic in about 19 years. The Moon's revolution from one node to the same node again (called the draconic month, 27.212220 days) takes somewhat less time than a revolution from Full Moon to Full Moon (the synodic month, 29.530589 days). For an eclipse to occur, the Moon has to be near one of the nodes of its orbit. The draconic month, therefore, is the other basic period of eclipses.

The saros period

Resonance between these two periods produces what is called the saros period, after which time Moon and Sun return very nearly to the same relative positions. The saros was known to the ancient Babylonians (see below *Use of eclipses for astronomical purposes: In ancient astronomy*). It comprises 223 lunations (the time period from New Moon to New Moon)—that is, 6,585.321124 days, or 241.9986 draconic months. This is nearly a whole number, so that the Full Moon is in almost the same position (*e.g.*, very near a node) at the beginning and end of a saros, a lapse of time equalling 18 years and 10⅓ days or 11⅓ days if five leap years fall within the period. Thus, there is usually a close resemblance between any eclipse and the one taking place 18 years and 11 days earlier or later. Since the date differs by only about 11 days in the calendar year, the latitudes on Earth of the two eclipses will be about the same, and so will the relative apparent sizes of Sun and Moon. The saros period also comprises 238.992 anomalistic months, again nearly a whole number. In one anomalistic month, the Moon describes its orbit from perigee to perigee, in which point it is nearest to the Earth. So the Moon's distance from the Earth is the same after a whole number of anomalistic months and very nearly the same after one

saros. The saros period is, therefore, extremely useful for the prediction of both solar and lunar eclipses.

Because of the extra one-third day in the saros, the eclipse recurs each time approximately 120° farther west on the surface of the Earth. After three saroses, or 54 years and about a month (a period known to the ancient Greeks as *exeligmos*), the longitude is repeated.

There is a regular shift on Earth to the north or to the south of successive eclipse tracks from one saros to the next. The eclipses occurring when the Moon is near its ascending node shift to the south, those happening when it is near its descending node shift to the north. A saros series of eclipses begins its life at one pole of the Earth and ends it at the other. Every saros series lasts about 1,300 years and comprises 73 eclipses. About 42 of these series overlap at any time.

The inex period

Two consecutive saros series are separated by the inex period, 29 years minus 20 days—that is, 358 synodic months—after which time the New Moon has come from one node to the opposite node. The lifetime of an inex group is about 23,000 years, 70 groups coexisting, each comprising 780 eclipses. All other cycles in eclipses are combinations of the saros and the inex.

Prediction and calculation of solar and lunar eclipses. The problem may be divided into two parts. The first is to find out when an eclipse will occur, the other to determine when and where it will be visible.

For this purpose it is convenient first to consider the Earth as fixed and to suppose the observer looking out from its centre. To this observer, O in Figure 5, the Sun

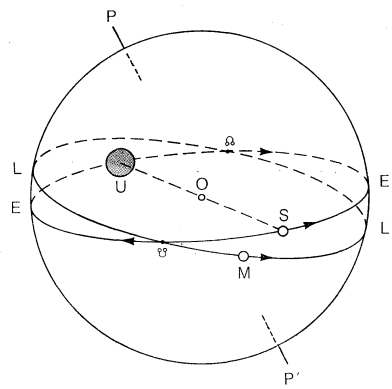


Figure 5: Apparent motions of the Sun and the Moon on the celestial sphere (see text).

and Moon appear projected on the celestial sphere. While this sphere appears to him to rotate daily, as measured by the positions of the stars, around the line PP′ (the Earth's axis of rotation), the Sun's disk, S, appears to travel slowly along the great circle EE′ (the ecliptic), making a complete revolution in one year. At the same time the Moon's disk, M, revolves along the circle LL′ once during a lunar month. The angular diameters of the two disks S and M are each about 0.5° but vary slightly.

Every month the Moon's disk revolving along LL′ will overtake the more slowly moving Sun once, at the moment of New Moon. Usually the Moon's disk will pass above or below the Sun's disk. Overlapping of the two results in an eclipse of the Sun; which can happen only when the New Moon occurs at a moment when the Sun is near the points ☊ or ☋; these signs denote the ascending and descending nodes of the Moon's orbit.

The crosscut of the umbra, the shadow cone of the Earth, at the distance of the Moon (as shown in Figure 5), may be projected like a disk U onto the celestial sphere. It subtends an angle of about 1.4°; its centre will always be opposite to the Sun's disk and will travel along EE′. An eclipse of the Moon takes place whenever the Moon's disk overlaps the shadow disk; this happens only when the shadow disk is near one of the nodes or the Sun is near the opposite node. The passage of the Sun through the lunar nodes is therefore the critical time for both solar and lunar eclipses. The Moon's orbit plane, represented by the circle LL′, is not fixed, and its nodes move slowly along the ecliptic in the direction indicated by the arrow, making a complete revolution in about 19 years. The interval between two successive passages of the Sun through one of the nodes is termed an "eclipse year," and, since the Moon's node moves so as to meet the advancing Sun, this interval is about 18.6 days less than a tropical (or ordinary) year.

The eclipse year

In Figure 6 the region of the ascending node as seen

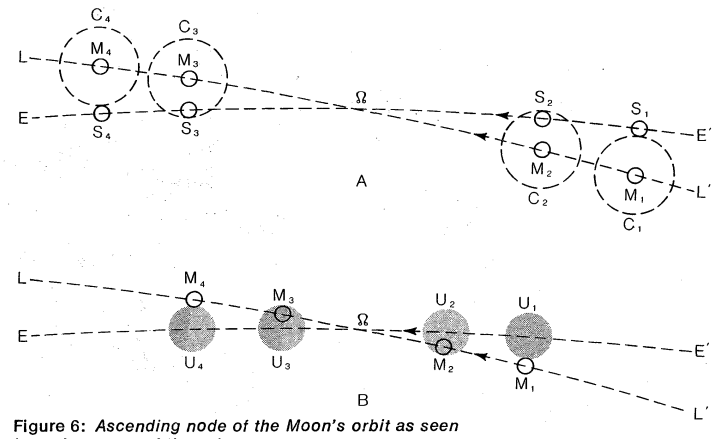


Figure 6: *Ascending node of the Moon's orbit as seen from the centre of the sphere.* Conditions necessary for (A) a solar eclipse and (B) a lunar eclipse (see text).

from the centre of the sphere is much enlarged. Here the node is kept fixed and the apparent motions of the Sun and the Moon are shown relative to the node. To the imaginary observer at the centre of the Earth, the Sun's disk will travel along the circle EE′, the Moon's disk along LL′. The Sun is so distant compared with the size of the Earth that, from all places on the Earth's surface, the Sun is seen nearly in the same position as from the centre. But the Moon is relatively near and its projected position on the celestial sphere is different for various observing stations on the Earth; it may be displaced as much as 1° from the position in which it is seen from the centre of the Earth. If the radius of the Moon's disk is enlarged by 1°, a circle, C, is obtained that encloses all possible positions of the Moon's disk seen from anywhere on the Earth. Conversely, if any circle of the Moon's size is drawn inside this "Moon circle," C, there is a place on the Earth from which the Moon is seen in that position.

Accordingly, there will be an eclipse of the Sun somewhere on the Earth whenever the Moon overtakes the Sun in such a position that the Moon circle, C, passes over the Sun's disk; when the latter is entirely covered by the Moon circle, the eclipse will be central (total or annular). From Figure 6A it is evident that a solar eclipse will take place if a New Moon occurs while the Sun moves from S_1 to S_4. This period is called the eclipse season; it starts 19 days before the Sun passes a node and ends 19 days after. Since there is a New Moon every month, at least one solar eclipse, and sometimes two, occurs during every eclipse season—of which there must be two in each calendar year. A fifth solar eclipse during a calendar year is possible because part of a third eclipse season may occur in the beginning of January or the end of December. When the New Moon falls within 11 days of the node passage (between S_2 and S_3), the eclipse will be central.

The eclipse season

Figure 6B illustrates the condition necessary for a lunar eclipse. If a Full Moon occurs within 13 days of a node passage of the Sun (when the shadow disk, U, passes the ascending node), the Moon will be eclipsed. Most eclipse seasons, but not all, will thus also contain a lunar eclipse. When three eclipse seasons fall in a calendar year, there may be three lunar eclipses in that year, as will next happen in 1982. Eclipses of the Sun are evidently more frequent than those of the Moon, but the former are visible only from a very limited region of the Earth, the latter from a whole hemisphere.

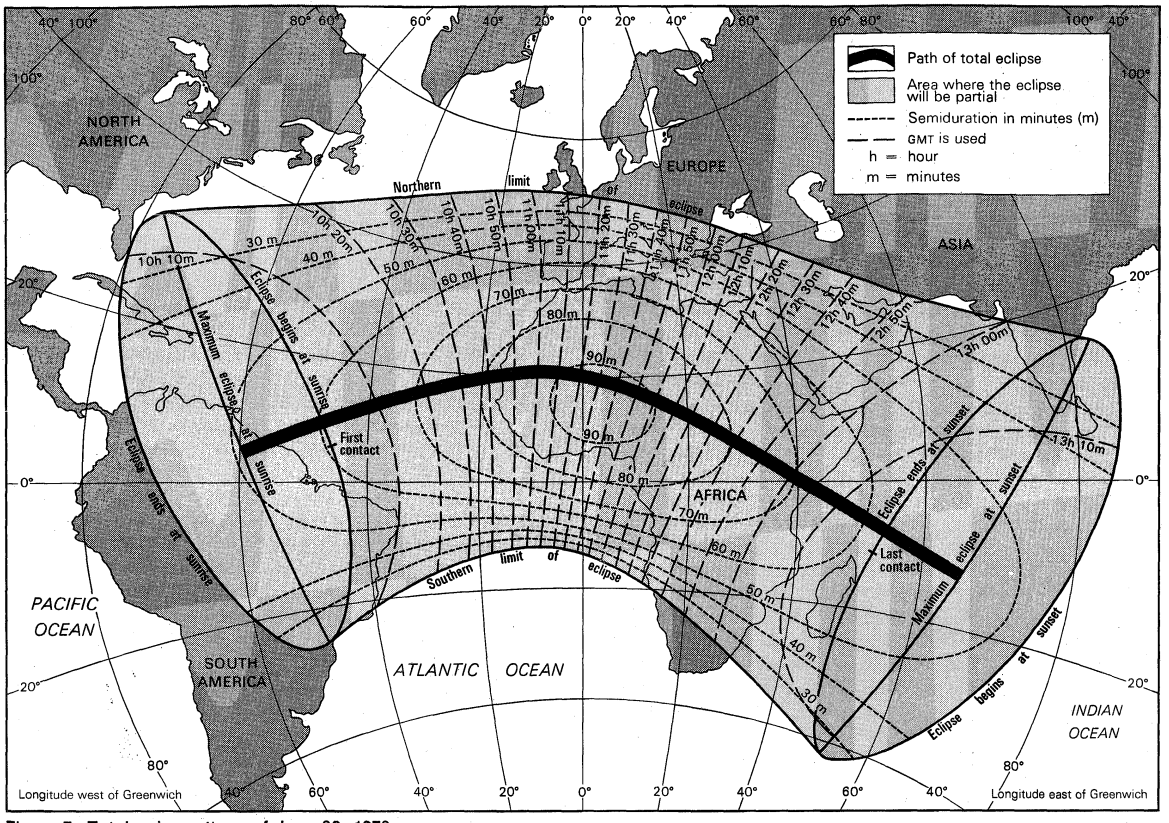


Figure 7: Total solar eclipse of June 30, 1973.
By courtesy of the U.S. Naval Observatory

During a solar eclipse the shadow cones (umbra and penumbra) of the Moon sweep across the face of the Earth (Figure 3), while at the same time the Earth is rotating on its axis. Figure 7 illustrates the circumstances of the total solar eclipse of June 30, 1973. The heavy black line marks the path along which the umbra travels from South America across the Atlantic Ocean—and Africa—into the Indian Ocean. Within the area of this lane, the eclipse is total. The wide surrounding region bounded by the full line is that covered by the penumbra, where the eclipse is partial. Within the area at its extreme left, the Sun is already in partial eclipse when it rises; within the area at the right, the Sun sets during the partial eclipse.

The astronomical ephemerides, or tables, published for each year give such maps for the more important eclipses as well as data for accurate calculation of the times of contact at any observing station. Calculations are made some years ahead in ephemeris time (ET), which is defined by the orbital motion of the earth and the planets. At the time of the eclipse, the correction is made to universal time (UT), which is defined by the rotation of the Earth and is not rigorously uniform.

It is possible with the aid of modern tables to predict solar eclipses several years ahead with an accuracy of a few seconds. For predictions of longer range, the main uncertainty is that of the Moon's motion. Eclipses can, of course, be "predicted backward" as well as forward, and the calculation of ancient eclipses has been of value in historical research (see below).

ECLIPSE RESEARCH ACTIVITIES

Solar research. During a solar eclipse, the Moon acts as a screen outside Earth's atmosphere. As seen from Earth, the Moon's dark projection on the Sun crosses the Sun at about 350 kilometres (220 miles) per second. For a few seconds, the bright, ordinarily visible disk of the Sun, called the photosphere, is eclipsed, while the chromosphere, the Sun's lower atmosphere, remains visible at the Sun's edge. In this brief time light emitted from the chromosphere can be studied and its decrease in intensity with height above the photosphere can be measured. Discrimination between layers within the chromosphere is possible to some extent. An instrument, the coronagraph, has been developed to obscure the brilliant photosphere artificially at times outside of eclipses, but the study of thin layers of the solar chromosphere and corona is still best done during an eclipse.

Study of solar chromosphere

The first photograph of a solar eclipse was taken in 1851, and the first of scientific importance by the British scientist and inventor Warren de la Rue and the Italian astronomer Angelo Secchi in 1860.

Spectroscopic observations. The ordinary spectrum of the Sun contains a brilliant many-coloured background—the continuum. This radiation emanates chiefly from the lower layers of the solar atmosphere. Cooler atoms higher up, however, selectively absorb the radiations, so that the solar spectrum consists of the bright rainbow background with many narrow gaps—dark lines —where the light has been absorbed.

Spectroscopy was first applied to the eclipsed Sun in 1868, when the path of totality passed over India and Malaya and the spectrum of bright lines originating in the solar prominences was observed. One of these lines was supposed by the British astronomer Joseph Norman Lockyer to be emitted by an unknown chemical element, which he called helium (from Greek *hēlios*, "sun"); helium was not identified on Earth until 1895. During the same eclipse, the French astronomer P.-J.-C. Janssen was inspired to find a means of observing the solar prominences without an eclipse, in which he succeeded the next day.

At the moment of totality, when the Moon obliterates the last trace of the bright photosphere, with its dark-line spectrum, the light from the upper tenuous layers of the atmosphere (which is usually lost in the much brighter photospheric light) flashes into view with the characteristic bright-line spectrum of a luminous gas. This spectrum, first observed in 1870, disappears within four to five seconds, as the Moon moves across the Sun. Because of its evanescent character, it has been called the flash spectrum. A second flash occurs at the end of totality.

Analysis of flash spectra has led to some surprising results. The spectrum matches the dark-line spectrum only roughly. The lines of the neutral metals are of comparable strength in the two spectra, but those of the ionized metals (*i.e.*, those made up of atoms that have lost one or more electrons) are markedly enhanced in the flash spectrum. The difference is attributed, in part, to lower pressures in the upper layers, but high temperature appears to contribute to the increased excitation. This is especially true for the flash lines of ionized helium, which do not appear at all in the ordinary dark-line spectrum. These require excitation temperatures of at least 25,000° K (about 45,000° F) for their production, whereas the temperature required to produce the observed quality and quantity of bright emission from the surface is only 6,000° K (about 10,800° F).

Other observations. Extending upward from the chromosphere, and closely related to it, are the so-called prominences, one of the striking features of a total eclipse, which project outward into space, sometimes as far as several hundred thousand miles. They appear as rose-coloured patches of flame, projecting well beyond the limb of the Moon, and consist of long interlacing filaments of incandescent gas.

Solar corona

During totality can be seen one of the most beautiful of natural phenomena, the solar corona, which shines like finely etched white frost against the deep blue of the eclipse-darkened sky. The form of the corona presented at different eclipses is almost infinitely variable. On occasion, usually when sunspots are near a minimum, long streamers extend four or five solar diameters away from the Sun. At other times, especially close to sunspot maximum, the corona is more nearly circular, but with jagged, petal-like extensions. The corona is faint, 500,000 times or so less brilliant than the Sun itself. Consequently the sky glare surrounding the Sun ordinarily hides the coronal details, and, before the development of the coronagraph, it was believed to be impossible to record the corona except during an eclipse.

Also interesting are observations of the radio emission of the Sun during a solar eclipse. When the Moon crosses different parts of the Sun, inhomogeneities in the generation of solar radio waves can be localized.

For geodetic purposes, measuring the Earth, the exact timing of the moments of the contacts of Moon and Sun are important, because these depend on the site of observation and so on the shape of the Earth.

The Sun's ultraviolet radiation and X-radiation helps to create and influences the ionosphere, a region in Earth's upper air where many atoms are ionized. The ionosphere changes during an eclipse as the ionizing radiation is cut off.

So that observations of the ionosphere made during an eclipse can be interpreted correctly, it is necessary to predict the times and phases of the eclipse as it reaches the high layers of Earth's atmosphere, where they may be substantially different from times and phases on the surface.

In addition to light and other electromagnetic radiation, the Sun continually emits charged particles or corpuscles; these particles form what is called the solar wind. This radiation is also screened from Earth during an eclipse of the Sun. Because the velocity of the particles is much less than that of light, and because of the movement of the Earth–Moon system around the Sun, the corpuscular eclipse should come some hours before the optical eclipse and over a different place on the surface of the Earth. No certain evidence of a corpuscular eclipse has been found, probably because observations of the solar wind are difficult and can be made only outside the atmosphere.

Observation from rockets

Since about 1960, rockets have been sent high into Earth's atmosphere to gather data during eclipses. Their trajectories are chosen to lie largely in the region from which the chromosphere and inner corona are visible, while the photosphere is eclipsed, and they gather data on X-rays and ultraviolet emission that cannot penetrate Earth's lower air.

The search for comets in the vicinity of the Sun during total eclipses has been rewarded on a few occasions.

Tests of relativity during solar eclipse. One of the predictions of the general theory of relativity, as presented by Einstein in 1915, is the deflection of light rays by a gravitational field, for example that of the Sun. According to this theory the deflection, which causes the image of a star to appear slightly too far from the Sun's image, amounts to 1.75 seconds of arc (a small fraction of a degree) at the limb of the Sun and decreases in proportion to the apparent distance from the centre of the solar disk of the star whose light is deflected. This is twice the amount given by the older Newtonian dynamics if light is assumed to have inertial properties. If light does not have such properties, as is generally accepted now, the Newtonian deflection is zero. To test the theory, extremely precise measurements are needed, of as many stars as possible around the Sun, and preferably close to it. The brightness of the corona prevents stars closer to the Sun than one solar diameter from being observed. During totality the sky around the Sun is photographed with long-focus cameras to give the largest possible scale. Then the equipment is left in place, and about half a year later the same region of the sky with the same stars in it can be photographed again, now during the night and without the disturbing gravity field of the Sun. Comparison of the two sets of photographs shows the amount of gravitational displacement of starlight by the Sun and serves as a test of theory. Results of a number of eclipse observations made since 1918 have verified Einstein's prediction, though the deduced values are somewhat uncertain as a consequence of the very small displacement.

Favourable conditions for observing solar eclipses. The most important conditions for observation of a solar eclipse are the duration of totality, the weather, and the position of the eclipse site.

When all circumstances are as favourable as possible, the duration of totality can be as long as seven minutes and 40 seconds. Among the determining factors are the excess of the apparent diameter of the Moon over that of the Sun, the relative velocities of the Sun and Moon across the sky as seen from the Earth, and the position of the intersection of the shadow cone with the surface of the Earth. To permit the longest observations, this intersection, the zone of totality, should be as close as possible to the Moon, so that the zone will be larger, and should fall at the region on Earth where Earth's rotational speed is at a maximum, so that an observer will remain longer in the moving shadow. Both these conditions are best met when the zone of totality falls near the Equator. In spite of all care, however, clouds at the moment of the eclipse can ruin everything. Clouds form more readily than usual because of the drop in temperature within the shadow, and the chances of good weather are increased when the eclipse takes place in desert regions or in dry weather conditions.

Likely weather conditions and accessibility of the site are considered important in choosing a site for an eclipse expedition. The expected scattering of light in the atmosphere and the quality of seeing (atmospheric steadiness) are also taken into account, so that, in spite of transportation difficulties, a site in the mountains may be preferred. A less turbulent atmosphere is also to be found over water, so that a coastal site from which the eclipsed Sun can be observed over the sea is often chosen. Sometimes observations are planned from a site near the limit of the zone of totality, where the Moon's edge can be seen apparently moving grazingly through the chromosphere.

Since 1963 the use of specially equipped aircraft as observation platforms above the clouds has made possible great increases in the observation time. Flying in the direction of the motion of the Moon's shadow, the aircraft can remain within the umbra for a long time, and this extends the effective duration of totality. Supersonic aircraft moving at about 1,200 miles (1,930 kilometres) per hour could keep pace with the shadow so that totality would be prolonged to an hour or more for observers on board.

Lunar research. Lunar eclipses can yield information about the cooling of the Moon's soil when the Sun's radi-

ation is suddenly removed and, therefore, about the soil's conductivity of heat and its structure. Infrared and radio-wavelength radiation from the Moon decline in intensity more slowly than does visible light emission during an eclipse because they are emitted from below the surface, and measurements indicate how far the different kinds of radiation penetrate into the lunar soil. Infrared observations show that at many "bright spots" the soil retains its heat much longer than in surrounding areas. Because of the absence of a lunar atmosphere the solid surface is exposed to the full intensity of ultraviolet and corpuscular radiation from the Sun, which, in some rock materials, may give rise to fluorescence. Observations during lunar eclipses have given positive results for this phenomenon with the appearance of abnormal bright regions in the obscured parts of the Moon.

Lunar "bright spots"

TRANSITS OF MERCURY AND VENUS

At the time of inferior conjunction (*i.e.*, when moving between the Earth and the Sun) the planet Mercury, seen from the Earth, usually passes north or south of the Sun because of the inclination of its orbit. But if the conjunction occurs when Mercury is very near one of the nodes of its orbit, the planet crosses the disk of the Sun as a small black circular spot, visible only with a telescope. Since the Earth passes Mercury's nodes on May 7 and November 9, transits can occur only near those dates. The shortest possible interval between two successive transits is seven years.

Transits of Venus can be seen without a telescope if the eyes are properly protected. When the transit is central it takes about eight hours. The phenomenon is rare and can happen only within a day of two of the dates when the Earth passes the nodes of Venus' orbit—that is, on June 7 and December 8. The transits come in pairs, with an interval of eight years between members of a pair; between the pairs, more than 100 years elapse.

Transits of Venus are helpful in finding the parallax and from it the distance of the Sun, as first pointed out by the British astronomer Edmond Halley in 1679. Parallax is the apparent difference in direction of an object when observed from different positions. The transits of June 1761 and 1769 and those of December in 1874 and 1882 were, therefore, extensively observed. The next pair of transits of Venus will occur on June 8, 2004 and June 6, 2012. It is not likely that so much trouble will be taken to observe them, since other, more accurate methods of determining the solar parallax have now been developed.

One kind of observation for determining parallax consists in fixing the times of the contacts of the disks of the planet and the Sun from different points on the Earth. The observers at past transits became aware of a few remarkable phenomena. When Venus was partially overlapping the disk of the Sun, the part of the limb of the planet that extended beyond the Sun was seen to be surrounded with a radiant aureole, which observers of the transit in 1761 ascribed to the presence of an atmosphere on Venus. A second phenomenon was seen just after second and before third contact (see Figure 1), when Venus just touched the Sun's limb on the inside; this consisted in the development of a little dark connection—the so-called black drop—between Venus and the limb. Because of the black drop, the times of contact could not be sharply defined. Presumed causes of the black drop are diffraction, atmospheric agitation, and also, perhaps, instrumental and physiological factors.

The black drop

The amount of sunlight intercepted during a transit depends on the diameter of the planet, and measuring this amount of sunlight may be one of the most accurate ways of determining the planet's diameter.

Other modern investigations possible during transits pertain to the screening of solar particles from Earth by the inner planets when they are close to conjunction. These electrically charged particles, forming part of the solar wind, have a velocity far less than that of light and are not only screened by the planet's disk but also by any magnetic field the planet may have, so that their influence on the Earth's magnetic field need not coincide with the time of the optical occultation.

OCCULTATIONS BY THE SUN AND MOON

The occultation of the Crab Nebula by the solar corona, the extensive outer atmosphere of the Sun, takes place each year in June. The Crab Nebula is a radio source (Taurus A), and, when the occultation occurs, its pattern of radio emission is observed to broaden significantly. The broadening is attributed to scattering of radio-frequency radiation by the density irregularities in the extended corona. The nearest approach of Sun and Crab Nebula is at about five solar radii; the scattering is brought into evidence up to 60 solar radii.

The Moon sometimes occults a planet but very often occults a star. As a consequence of the inclination of the Moon's orbit with respect to the ecliptic (the Sun's apparent annual path) and the movement of the nodes of the Moon's orbit, all the stars in a belt of 10° around the ecliptic are occulted at some time during a period of about nine years. Among these stars are the bright stars Aldebaran, Regulus, Spica, and Antares and the star clusters Pleiades, Hyades, and Praesepe, and the Crab Nebula.

Since the Moon always moves eastward, an occulted star disappears at the Moon's eastern limb and reappears at the western. These phenomena can be best observed at the dark limb of the Moon.

Efforts have been made to determine the apparent diameter of stars by accurately estimating the time they take to disappear behind the Moon; with modern optics and electronic devices this would not be impossible in some cases. But there are complications of two kinds: first, many stars are close binaries (doubles); and, second, the decline of the light of the occulted star shows an optical phenomenon called a diffraction pattern. But, for the purpose of yielding basic data on the perturbations (disturbances) in the Moon's orbit and on irregularities in the rotation of the Earth, the exact times of star occultations are still of value.

The disappearance or reappearance of Jupiter takes more than one minute, but the time for a minor planet is under one second, and that for even a giant star like Aldebaran or Antares is less still, by a factor of ten. Other stars disappear nearly instantaneously, indicating that the Moon has no sensible permanent atmosphere.

Occultation of the Crab Nebula by the Moon has attracted much interest because of the possibility it affords of finding out whether the nebula's X-ray emission comes from a small source like a star or from a more extended source like a cloud of gas. Observations on July 7, 1964, indicated that the X-ray source in the central region of the nebula had a width of about 1′ of arc, which is roughly one light-year at the distance of the Crab Nebula—that is, a relatively extended source.

Lunar occultation of the Crab Nebula

The Crab Nebula can also be observed as a radio source during an occultation by the Moon. Because the radio waves are subject to refraction by electrons, an estimate of the density of the lunar atmosphere can be made. The density has proved to be below the possibility of measurement; the upper limit of density obtained by this method is very much lower than that determined by optical means.

ECLIPSING BINARY STARS

When, in a binary system (a double star whose components are in orbit around a common centre), one component comes between the other and the Earth, an eclipse occurs and the amount of light received from both stars together is reduced; a total eclipse gives the greatest reduction, and the light is reduced to a minimum. Two eclipses take place during each revolution. The first star shown to have light that varied through eclipses in this way was Algol (Beta Persei). Its variability was discovered in 1667 by the Italian astronomer Geminiano Montanari, but the explanation as eclipsing variable was not proposed until a century later by the English astronomer John Goodricke. At present, about 3,000 eclipsing binaries are known.

The many possible sizes of the two stars, both in relation to each other and to the size and tilt of their orbits, combined with the eccentricity (deviation from the circu-

lar) of the orbits, lead to a large variety of ways in which the light of the system changes during a revolution: the so-called light curve.

If the orbits of the two stars are not circular, then the motion in the orbits is not uniform but is most rapid near periastron (stars closest together) and least rapid near apastron (stars farthest apart). Therefore, the eclipses may not be equally spaced in time, as they are in the case of circular orbits. Study of the spacing and durations of the light minima will yield both the eccentricity and the orientation of the orbits.

A steady change in the spacing of the minima as well as in their relative duration means that the line of apsides (major axis of the relative orbit) is rotating; the most rapid such rotation known is that of the eclipsing binary GL Carinae (see also STAR).

If the two stars are unequal in size, they may wholly overlap for some time. When the star of higher surface brightness is behind, the resulting eclipse is darker than that resulting when the brighter star cuts off the light of the dimmer one. Limb darkening, the tendency for a star to appear less brilliant near the edge of its apparent disk, also influences the shape of the light curves, and the amount of limb darkening present may be determined from those curves.

Spectroscopic studies of the light of a double star allow determination of the ratio of the masses of the components to each other. The sum of the masses, however, can only be determined if the inclination of the orbit is known. In the case of an eclipsing binary, the plane of the orbit is known to be directed to the Earth. Therefore, eclipsing binaries can be made to reveal complete information about their masses; this is one reason why they are so energetically studied.

A rapidly rotating star bulges at its equator. Also, many double stars are so close together that they distort each other tidally; in the case of very close pairs, this distortion has caused them to turn always the same sides toward each other. If such stars are eclipsing binaries, the amount of distortion can, in some cases, be determined from the light curve. For some stars the ratio of the long axis of the star to the short has been found to be as great as 5:3.

The star Zeta Aurigae

There are a few star systems like Algol, notably Zeta Aurigae, in which the difference in the sizes of the components is so great that the crossing of the smaller, brighter star before the disk of the supergiant companion should really be called a transit rather than an eclipse. This particular star has the remarkably long orbital period of 972 days and consists of an orange supergiant having a diameter of almost 300,000,000 kilometres (186,000,000 miles, about equal to the diameter of Earth's orbit) and a much smaller, brighter star, only three times larger than the Sun. The secondary minimum lasts for 38 days, and a very sensitive photometer must be employed to detect the extremely slight fading of light when the smaller star is shining through the extended, tenuous atmosphere of the large one (see also STAR).

ECLIPSES IN HISTORY

In ancient times eclipses were regarded as portents; hence it is not surprising that many eclipses are mentioned in history and in literature, in addition to those recorded in astronomical writings.

The value of these records may be classified as follows: (1) literary and historical, depending on the interest that they aroused and their connection with historical events; (2) chronological, insofar as they make it possible to verify chronological systems resting on other evidence and to supply dates for events connected with the eclipses; (3) astronomical, including the determination by ancient astronomers of the periods and motions of the Sun and Moon and by modern astronomers of the mean rate of change of those motions.

Literary and historical references. *Chinese.* In the *Shu Ching* ("Classic of History") it is stated that Hsi and Ho (semidivine figures sent by the legendary emperor Yao to the four corners of the Earth to keep the Sun on its course and to prevent eclipses) had neglected the duties of their office and had abandoned themselves to drink in their own cities. The king, Chung K'ang, placed the Marquess of Yin in command of an army with instructions to punish them. The Marquess issued a florid address to his troops in which he referred to Hsi and Ho as having neither heard nor known anything on a recent occasion when "the Sun and Moon did not meet harmoniously in Fang." This phrase is taken by Chinese scholars, ancient and modern, to mean an eclipse. It would therefore appear that a military expedition was sent to punish Hsi and Ho for failing to observe or to prevent an eclipse. There are good reasons for believing that this part of the *Shu Ching* is a literary restoration dating from the 4th century AD, but enough is found in a quotation in the *Tso Chuan* ("Commentary of Tso," compiled between 430 and 250 BC) and in statements in the *Bamboo Annals* to prove that the reference is to an authentic eclipse, though its relation to the supposed offense of Hsi and Ho is doubtful. Attempts have been made to identify this eclipse, but the necessary data are wanting. The accepted Chinese chronology places the accession of Chung K'ang in 2159 BC. The *Bamboo Annals*, whose chronology is probably nearer to the truth, place it in 1952 BC. The *Shih Ching*, or "Classic of Poetry," contains a lamentation caused by an eclipse of the Moon followed by an eclipse of the Sun. The dates are clearly defined and are found to agree with events in 734 BC. The earliest eclipses that have been definitely identified are (for the Moon) that of 1361 BC and (for the Sun) that of 1217 BC, both derived from inscriptions on ancient oracle bones. The eclipses recorded in the *Ch'un Ch'iu*, "Spring and Autumn" annals, possess chronological and astronomical rather than literary interest. Some evidence for a total eclipse of the Sun in 1375 BC has been found in the form of a clay tablet record excavated at ancient Ugarit (now in Syria).

Assyrian. The Assyrian eponym canon, which preserves the names of the annual magistrates who gave their names to the years, records under the year that corresponds to 763–762 BC: "Insurrection in the city of Ashur. In the month Sivan the sun was eclipsed." The reference must be to the eclipse of the Sun on June 15, 763 BC. A reference to the same eclipse has been found in the Bible: " 'And on that day,' says the Lord God, 'I will make the sun go down at noon, and darken the earth in broad daylight' " (Amos 8:9).

The Homeric references

Greek. Homer, in the *Odyssey*, twice makes the unrecognized Odysseus predict that Odysseus will return "as the old moon wanes, and the new appears"; that is, at New Moon, when alone an eclipse of the Sun is possible. On the day when Odysseus was to become known and slay his wife's suitors, the seer Theoclymenus notes among other portents of gloom that "the sun has perished out of heaven, and an evil mist has spread over all." This was interpreted by Plutarch and Eustathius as a total eclipse of the Sun. Some modern scholars have regarded it as merely a vision of the seer. An identification with the solar eclipse of April 16, 1178 BC, which was total in or near Ithaca, has been suggested. But it is not improbable that this story belongs to legend rather than to history.

In a fragment of a lost poem by Archilochus occur the words:

> Nothing there is beyond hope, nothing that can be sworn impossible, nothing wonderful, since Zeus, father of the Olympians, made night from mid-day, hiding the light of the shining sun, and sore fear came upon men.

This phenomenon has been identified as the total eclipse of the Sun on April 6, 648 BC.

Small fragments survive of other poetic descriptions of eclipses, and the ninth paean of Pindar, addressed to the Thebans, takes an eclipse of the Sun as its theme, as follows:

> Beam of the sun! O thou that seest afar, what wilt thou be devising? O mother of mine eyes! O star supreme, reft from us in the daytime! Why has thou perplexed the power of man and the way of wisdom, by rushing forth on a darksome track?

Pindar then proceeds to speculate on the meaning of the

eclipse as a portent. The poem probably refers to the solar eclipse of April 30, 463 BC, which was nearly total at Thebes.

The most famous of ancient eclipses was a total eclipse of the Sun that happened, according to Herodotus, during a battle between the Lydians and the Medes. The portent induced them to conclude peace. Herodotus further says that Thales of Miletus predicted this eclipse to the Ionians for the very year in which it happened. Babylonians were certainly predicting eclipses at about that time by means of the saros (see above). But, as that cycle gives eclipses for every year, the statement about Thales must, if true, mean that he predicted that an eclipse of that particular year would attain a great magnitude and that it would take place in Asia Minor. Such a prediction can be made by the saros without any mathematical computation. The eclipse was certainly that of May 28, 585 BC, and must have been predicted by means of the eclipse of May 18, 603 BC. Several ancient writers have recorded 585 BC or some neighbouring year as the date of the eclipse.

The historian Thucydides comments on the frequency of eclipses during the Peloponnesian War. The most interesting of these were the solar eclipse of August 3, 431 BC, when "the sun assumed the shape of a crescent and became full again, and during the eclipse some stars became visible" (a statement that agrees well with modern computation), and the lunar eclipse of August 27, 413 BC. That date had been selected by the Athenian commanders Nicias and Demosthenes for the departure of their armies from Syracuse. All preparations were ready, but the signal had not been given when the Moon was eclipsed. The soldiers and sailors clamoured against departure, and Nicias, in obedience to the soothsayers, resolved to remain thrice nine days. This delay enabled the Syracusans to capture or destroy the whole of the Athenian fleet and army.

Eclipses in Thucydides

August 15, 310 BC, is the date of a total eclipse of the Sun said to have been seen by the commander Agathocles and his men the day after he had sailed from Syracuse on his way to Africa. Modern computations of the eclipse track render it probable that he sailed to the north of Sicily.

In Plutarch's dialogue concerning the features of the Moon's disk, one of the characters, called Lucius, deduces from the phases of the Moon and the phenomena of eclipses a similarity between the Earth and the Moon and illustrates his argument by means of a recent eclipse of the Sun, "which, beginning just after noon, showed us plainly many stars in all parts of the heavens, and produced a chill in the temperature like that of twilight." This eclipse has been identified with that of March 20, AD 71. Further on, Lucius refers to a brightness that appears round the Moon's rim in total eclipses of the Sun. There are numerous other references to eclipses in Greek literature.

Roman. A large number of eclipses was recorded in Roman history. One that has attracted the attention of students of astronomy and of the Roman calendar alike is stated by Cicero to have occurred in what may have been the 350th year from the foundation of Rome. He also says that it was described by the poet Quintus Ennius: "On the Nones of June the sun was covered by the moon and night." This would appear to have been the solar eclipse of June 21, 400 BC, which reached a total or almost total phase at Rome a few minutes after sunset. It seems to show that in that year the calendar month of June began 16 days later than it did after the Julian reform. The eclipse of the Moon on June 21–22, 168 BC, has attracted much attention. The Romans were at that time at war with Macedonia, and Polybius says that this eclipse was interpreted as an omen of the eclipse of a king and thus encouraged the Romans and discouraged the Macedonians.

Use of eclipses for chronological purposes. No one system of dating has been continuously in use since ancient times. Formerly, dates were frequently expressed in terms of a king's reign; years were also named after officials of whom lists have been preserved (eponym canons, see above). In these cases it is important to be able to equate certain specific years thus defined with years before the Christian Era (BC). This can be done whenever the date of an eclipse or other identifiable and calculable astronomical phenomenon is given in an ancient record.

Chinese chronology can be confirmed accurately by eclipses from the Shang dynasty (*c.* 1766–*c.*1122 BC) onward. The chronology of Ptolemy's canon list of kings, which gives the Babylonian series from 747 to 539 BC, the Persian series from 538 to 324 BC, the Alexandrian series from 323 to 30 BC, and the Roman series from 30 BC onward, is confirmed throughout by eclipses. The eclipse of 763 BC, recorded in the eponym canon, makes it possible to carry the chronology back with certainty through the period covered by that canon to 893 BC. Assyrian, Babylonian, and Persian records provide a reliable chronology for the countries with which Assyria, Babylon, and Persia came into contact, but there is no check from eclipses on Greek or Egyptian historical chronology before the Persian period. Identifiable eclipses recorded under named Roman consuls extend back to 217 BC. The dated solar eclipse of Ennius, the lunar eclipse seen at Pydna in Macedonia on June 21–22, 168 BC, and an eclipse of 190 BC can be used to determine the position of Roman calendar months in the natural year; and occasionally eclipses help to fix the precise dates of a series of events, such as those connected with the Athenian disaster at Syracuse.

Use of eclipses for astronomical purposes. *In ancient astronomy.* At least from the time of King Nabonassar (747 BC), a dated canon of astronomical observations was preserved at Babylon and included all eclipses. This rendered it possible to determine accurately the intervals between eclipses and must have facilitated the discovery of the 18-year cycle (more exactly the cycle of 6,585⅓ days that the 10th-century Greek lexicographer Suidas named the saros) and its multiple, the 54-year cycle of 19,756 days (see above). It has been seen that Thales probably used such a cycle. In a Babylonian observation tablet of 568 BC, mention is made of failure to observe a predicted eclipse of the Moon, which is found by computation to have been real but invisible at Babylon. It was doubtless predicted by cycle. Tables based on the saros go back to the 4th century BC. Lunar eclipses give more accurately than any other observation the actual time when Sun and Moon are in opposition. From an early date the Babylonian astronomers must have deduced from them not only the mean interval between two conjunctions (closest apparent approaches) of Sun and Moon but also the principal inequality (change of speed) in the motion of the Moon and the similar inequality in the motion of the Earth (or, as according to their geocentric theory they conceived it, of the Sun). They were able to define the periods of these inequalities (the cause of which lies in the ellipticities of the orbits), which astronomers call the anomalistic month and year.

Babylonian record of observations

In the same way, since eclipses happen only when the Sun and Moon are at the intersections of their orbital planes called nodes, and since the path of the shadow in a lunar eclipse depends on the position of the centre of the Sun in relation to the node, the Babylonians were also able to determine the position and motion of the nodes. By assuming, as is approximately true, that the saros of 6,585⅓ days contained an exact number: (1) of synodic months, or revolutions of the Moon measured from the Sun, (2) of anomalistic months, or revolutions of the Moon measured from its apogee or perigee (*i.e.*, from its farthest distance from and closest approach to the Earth), and (3) of draconic months, or revolutions of the Moon measured from its node, astronomers, perhaps as early as the 6th century BC, computed the relative motions of the Sun and Moon, the lunar perigee and apogee, and the nodes. About 500 BC the Chaldean astronomer Nabu-rimmani, apparently from a more accurate study of eclipse observations, obtained improved values that: for the motion of the Moon relative to the Sun were 10″ of arc per annum too small; for the Moon's perigee motion 20″ of arc per annum too great; and for the motion of its

node 5″ of arc too small. Still more accurate values were obtained by Kidinnu about 383 BC, from whom they passed to the Greek astronomer Hipparchus. In the system of Nabu-rimmani the distance of the Moon from its node was used for the prediction of the magnitude of lunar eclipses.

In modern astronomy. Ancient eclipses are of the highest value for the determination of "secular accelerations"—*i.e.*, for the determination of the progressive, as distinct from the periodic, changes in celestial motions. In 1693 the British astronomer Edmond Halley asserted his belief that the Moon's motion was subject to acceleration but did not specify the amount. In 1749 the British astronomer Richard Dunthorne demonstrated from a comparison of the recorded with the computed times of eclipses distributed over 2,000 years that such an acceleration existed and assigned to it the value of 10″ of arc a century. By this is meant that the effect of the acceleration is to produce an advance of 10″ of arc in the Moon's longitude in the course of a century. The French mathematician and astronomer P.-S. Laplace showed in 1787 that such an acceleration should be the gravitational effect of the secular diminution in the eccentricity of the Earth's orbit; but, according to J.C. Adams in 1853, the acceleration due to this cause amounted to 6″ of arc only in a century. The residue must be explained by other causes, mainly tidal friction.

BIBLIOGRAPHY. D.H. MENZEL, *Our Sun* (1949), a semi-scientific description of solar radiation and phenomena, with many plates; S.A. MITCHELL, *Eclipses of the Sun*, 5th ed. (1951), the development of solar science and a discussion of the writer's experiences and results during ten eclipse expeditions; F.J.M. STRATTON, *Modern Eclipse Problems* (1927), masterly, understandable presentation of knowledge at that time; SIR F.W. DYSON and R. VAN DER R. WOOLLEY, *Eclipses of the Sun and Moon* (1937), history, methods, and results of eclipse observations; W.M. SMART, *Text-Book on Spherical Astronomy*, 5th ed. (1962), presentation of the underlying, not too difficult mathematical tools for the general calculations of occultations and eclipses; F. LINK, *Eclipse Phenomena in Astronomy* (1969), treatment of modern developments in eclipse problems, excluding solar eclipses and eclipsing variables; J.K. FOTHERINGHAM, *Historical Eclipses* (1921); T. VON OPPOLZER, *Canon der Finsternisse* (1887; Eng. trans., *Canon of Eclipses*, 1962), astronomical data of all solar eclipses between 1207 BC and AD 2161 and eclipses of the moon from 1206 BC to AD 2163 with maps of the central lines of the solar eclipses over the earth: G. VAN DEN BERGH, *Eclipses in the Second Millennium B.C. (—1600 to —1207)* (1954), demonstration of a method to compute each of these eclipses with simple arithmetic, and *Periodicity and Variation of Solar (and Lunar) Eclipses* (Eng. trans. 1955), arrangement of all eclipses of Oppolzer's *Canon* into the saros and the inex periods in a newly developed panorama; J. MEEUS, C.C. GROSJEAN, and W. VANLERLEEN, *Canon of Solar Eclipses* (1966), more accurate data and maps of central lines over the whole earth of all solar eclipses between AD 1898 and 2510; J.K. FOTHERINGHAM, "A Solution of Ancient Eclipses of the Sun," *Mon. Not. R. Astr. Soc.*, 81:104–126 (1920), computation of the accelerations of the Moon and Sun in 11 eclipses between 1062 BC and AD 364, with quotations from ancient records; F. BOLL, "Finsternisse," *Pauly-Wissowa, Real-Encyclopaedie*, 6:2329–2364 (1909); O. NEUGEBAUER, "The History of Ancient Astronomy: Problems and Methods," *J. Near Eastern Studies*, 4:1–38 (1945); J. NEEDHAM, *Science and Civilisation in China*, vol. 3 (1959), discussion of eclipses, placed in the context of Chinese astronomy, with extensive references to original literature; A. PANNEKOEK, *De groei van ons wereldbeeld* (1951; Eng. trans., *A History of Astronomy*, 1961), critical consideration of astronomical knowledge of the peoples in the ancient world and the place of astronomy in their cultures; SIR B. LOVELL (ed.), *Astronomy*, 2 vol. (1970), discourses in the physical sciences from 1851 to 1939, many of which are devoted to solar eclipses; *Explanatory Supplement to the Astronomical Ephemeris and the American Ephemeris and Nautical Almanac* (HMSO, 1961), a full presentation of all data for the calculation of astronomical phenomena.

(J.H.)

Ecology

Long unfamiliar to the public, and relegated to a second-class status by many in the world of science, ecology emerged in the 1960s as one of the most popular and most important aspects of biology. It has become painfully evident that the most pressing problems in the affairs of men—expanding populations, food scarcities, environmental pollution, and all the attendant sociological and political problems—are to a great degree ecological.

The word ecology was coined by a German zoologist, Ernst Haeckel, who applied the term *oekologie* to the "relation of the animal both to its organic as well as its inorganic environment." The word comes from the Greek *oikos*, meaning "household, home, or place to live." Thus ecology deals with the organism and its environment. The word environment includes both other organisms and physical surroundings. It involves relationships between individuals within a population and between individuals of different populations. These interactions between individuals, between populations, and between organisms and their environment form ecological systems, or ecosystems. Ecology has been defined variously as "the study of the interrelationships of organisms with their environment and each other," as "the economy of nature," and as "the biology of ecosystems."

HISTORICAL BACKGROUND

Ecology had no firm beginnings. It evolved from the natural history of the Greeks, particularly Theophrastus, a friend and associate of Aristotle. He first described the interrelationships between organisms and between organisms and their nonliving environment. Later foundations for modern ecology were laid in the early work of plant and animal physiologists.

Chronological sequence. In the early and mid-1900s two groups of botanists, one in Europe and the other in America, studied plant communities from two different points of view. The European botanists concerned themselves with the study of the composition, structure, and distribution of plant communities. The American botanists studied the development of plant communities, or succession. Both plant and animal ecology developed separately until American biologists emphasized the interrelation of both plant and animal communities as a biotic whole.

During the same period interest in population dynamics developed. The study of population dynamics received special impetus in the early 19th century, after a British economist, Thomas Malthus, called attention to the conflict between expanding populations and the capability of the earth to supply food. R. Pearl (1920), A.J. Lotka (1925), and V. Volterra (1926) developed mathematical foundations for the study of populations, and these studies led to experiments on the interaction of predators and prey, competitive relationships between species, and the regulation of populations. Investigations of the influence of behaviour on populations was stimulated by the recognition in 1920 of territoriality in nesting birds. Concepts of instinctive and aggressive behaviour were developed by K. Lorenz and N. Tinbergen, and the role of social behaviour in the regulation of populations was explored by V.C. Wynne-Edwards.

While some ecologists were studying the dynamics of communities and populations, others were concerned with energy-budgets. In 1920, August Thienemann, a German freshwater biologist, introduced the concept of trophic, or feeding, levels, by which the energy of food is transferred through a series of organisms, from green plants (the producers) up to several levels of animals (the consumers). An English animal ecologist, C.E. Elton (1927), further developed this approach with the concept of ecological niches and pyramids of numbers. Two American freshwater biologists, E. Birge and C. Juday, in the 1930s, in measuring the energy budgets of lakes, developed the idea of primary production, *i.e.*, the rate at which food energy is generated, or fixed, by photosynthesis. Modern ecology came of age in 1942 with the development, by R.L. Lindeman of the United States, of the trophic-dynamic concept of ecology, which details the flow of energy through the ecosystem. Quantified field studies of energy flow through ecosystems were further developed by Eugene and Howard Odum of the United

States; similar early work on the cycling of nutrients was done by J.D. Ovington of England and Australia.

The study of both energy flow and nutrient cycling was stimulated by the development of new techniques—radioisotopes, microcalorimetry, computer science, and applied mathematics—that enabled ecologists to label, trace, and measure the movement of particular nutrients and energy through the ecosystems. These modern methods encouraged a new stage in the development of ecology—systems ecology, which is concerned with the structure and function of ecosystems.

Conceptual development. Until recent years ecology lacked a strong conceptual base. Modern ecology, however, is now focused on the concept of the ecosystem, a functional unit consisting of interacting organisms and all aspects of the environment in any specific area. It contains both the nonliving (abiotic) and living (biotic) components through which nutrients are cycled and energy flows. To accomplish this cycling and flow, ecosystems must possess a number of structured interrelationships between soil, water, and nutrients, on the one hand, and producers, consumers, and decomposers on the other. Ecosystems function by maintaining a flow of energy and a cycling of materials through a series of steps of eating and being eaten, of utilization and conversion, called the food chain. Ecosystems tend toward maturity, or stability, and in doing so they pass from a less complex to a more complex state. This directional change is called succession. Whenever an ecosystem is used, and that exploitation is maintained—as when a pond is kept clear of encroaching plants or a woodland is grazed by domestic cattle—the maturity of the ecosystem is effectively postponed. The major functional unit of the ecosystem is the population. It occupies a certain functional niche, related to its role in energy flow and nutrient cycling. Both the environment and the amount of energy fixation in any given ecosystem are limited. When a population reaches the limits imposed by the ecosystem, its numbers must stabilize or, failing this, decline from disease, starvation, strife, low reproduction, or other behavioural and physiological reactions. Changes and fluctuations in the environment represent selective pressure upon the population to which it must adjust. The ecosystem has historical aspects: the present is related to the past and the future to the present. Thus the ecosystem is the one concept that unifies plant and animal ecology, population dynamics, behaviour, and evolution.

The unifying concept of ecology

SCOPE

Of necessity, ecology is a multidisciplinary science. It involves plant and animal biology, taxonomy, physiology, genetics, behaviour, meteorology, pedology, geology, sociology, anthropology, physics, chemistry, mathematics, and electronics. Often it is difficult to draw a sharp line between ecology and any of these, for all impinge on it. The same situation exists also within ecology. In understanding the interactions between the organism and the environment or between organisms, it is often difficult to separate behaviour from population dynamics, behaviour from physiology, adaptation from evolution and genetics, animal ecology from plant ecology.

Ecology developed along two lines: the study of plants and the study of animals. Plant ecology focuses on the relationships of plants to other plants and their environment. The approach is largely descriptive of the vegetational and floristic composition of an area and usually ignores the influence of animals on the plants. Animal ecology concerns the study of population dynamics, distribution, behaviour, and the interrelationships of animals and their environment. Because animals depend upon plants for food and shelter, animal ecology cannot be fully understood without a considerable background of plant ecology. This is particularly true in applied areas of ecology—wildlife and range management.

Both plant and animal ecology may be approached as the study of the interrelations of an individual organism with its environment, called autecology, or as the study of groups of organisms, called synecology.

Autecology, in many ways the classical study of ecology, is experimental and inductive. Because it is usually concerned with the relationship of an organism to one or more variables such as humidity, light, salinity, or nutrient levels, it is easily quantified and lends itself to experimental design both in the field and the laboratory. It has borrowed techniques from chemistry, physics, and physiology. Autecology has contributed at least two important concepts: the constancy of interaction between an organism and its environment, and the genetic adaptability of local populations to local environmental conditions.

Autecology and synecology

Synecology, on the other hand, is philosophical and deductive. It is largely descriptive and not easily quantified and contains a bewildering array of terminology. Only recently, since the advent of the electronic and atomic ages, has synecology developed the tools to study complex systems and enter an experimental phase. Important concepts developed by synecology are those concerned with nutrient cycling, energy budgets, and ecosystem development. Synecology has strong ties with pedology, geology, meteorology, and cultural anthropology.

Synecology may be subdivided according to environmental types, as terrestrial or aquatic. Terrestrial ecology, which may be further subdivided into forest, grassland, arctic, and desert ecology, concerns such aspects of terrestrial ecosystems as microclimate, soil chemistry, soil fauna, hydrologic cycles, ecogenetics, and productivity. Terrestrial ecosystems are more influenced by organisms and subject to much wider environmental fluctuations than are aquatic ecosystems. Aquatic ecosystems are affected more by the condition of the water and resist such environmental variables as temperature. Because the physical environment is so important in controlling aquatic ecosystems, considerable attention is paid to the chemical and physical characteristics of the ecosystem, such as the currents and the chemical composition of the water. By convention, aquatic ecology, called limnology, is limited to freshwater stream ecology and lake ecology. The former concerns life in flowing waters; the latter, life in relatively still water. Marine ecology deals with life in the open sea and in estuaries.

Other ecological approaches concern specialized areas. The study of the geographic distribution of plants and animals is ecological plant and animal geography. The study of population growth, mortality, natality, competition, and predator–prey relations is population ecology. The study of the genetics and ecology of local races and distinct species is ecological genetics. The study of the behavioral responses of animals to their environment, and of social interactions as they affect population dynamics, is behavioural ecology. Investigations of interactions between the physical environment and the organism fall under ecoclimatology and physiological ecology. The study of groups of organisms is community ecology (although it is difficult to separate it from studies of bioenergetics, biogeochemical cycles, and trophic-dynamic aspects of the community or ecosystem ecology). That part of ecosystem ecology concerned with the analysis and understanding of the structure and function of ecosystems by the use of applied mathematics, mathematical models, and computer programs is systems ecology. Systems ecology, concentrating on input and output analysis, has stimulated the rapid development of applied ecology, concerned with the application of ecological principles to the management of natural resources, agricultural production, and problems of environmental pollution.

Systems ecology

METHODOLOGY AND RESEARCH

Because ecologists work with living systems possessing numerous variables, the techniques used by physicists and chemists, mathematicians and engineers, require modification; they are not easily applied nor are the results as precise as those obtained in other sciences. It is relatively simple, for example, for a physicist to measure gain and loss of heat from metals or other inanimate objects, which possess certain constants of conductivity, expansion, surface features, and the like. To determine the heat exchange between an animal and its environment, how-

ever, a physiological ecologist is confronted with an array of almost unquantifiable variables and has the formidable task of both gathering the numerous data and analyzing them. Ecological measurements probably never will be as precise or as subject to the same ease of analysis as measurements in physics, chemistry, or certain quantifiable areas of biology.

In spite of these problems, various aspects of the environment can be determined by physical and chemical means, ranging from simple chemical identifications and physical measurements to the use of sophisticated mechanical apparatus. The development of biostatistics and proper experimental design, and the improvements in methods of sampling, permit a quantified statistical approach to the study of ecology. Because of the extreme difficulties of controlling environmental variables in the field, studies involving the use of experimental design are largely confined to the laboratory and to controlled field experiments designed to test the effects of only one variable or several variables. The use of statistical procedures, and the application of computer science to mathematical models based on data obtained from the field, are providing new insights into population interactions and ecosystem function. Mathematical programming is becoming increasingly important in applied ecology, especially in the management of natural resources and agricultural problems having an ecological basis.

Controlled environmental chambers enable experimenters to maintain plants and animals under known conditions of light, temperature, humidity, and daylength so that the effects of each variable (or combination of variables) on the organism can be studied. Biotelemetry and other electronic tracking equipment, products of the space age, permit the rapid and nondestructive sampling of plant and animal populations. Such tools enable ecologists to follow from a distance the movements and behaviour of a free-ranging animal by radio signals beamed from a sender attached to the organism. Radioisotopes are used for tracing the pathways of nutrients through ecosystems, for determining the time and extent of transfer of energy and nutrients through the different components of the ecosystem, and for the determination of food chains. The use of laboratory microcosms—aquatic and soil micro-ecosystems, consisting of biotic and nonbiotic material from natural ecosystems, held under conditions similar to those found in the field—are useful in determining rates of nutrient cycling, ecosystem development, and other functional aspects of ecosystems. Microcosms enable the ecologist to duplicate experiments and to perform experimental manipulation on them.

APPLICATIONS

Man, the dominant organism on earth, has stood apart from the interaction of living things; yet, as he is discovering, he is inextricably tied to his environment. The air he breathes, the water he drinks, the food he consumes, and the products he uses and throws away bind him to the functions of local and global ecosystems.

Man-influenced ecosystems

Agriculture, forestry, wildlife, and fisheries all represent some aspect of man's modification and exploitation of natural ecosystems. To achieve maximum food production, man has eliminated stable, complex ecosystems and replaced them with fragile, simplified single-species ecosystems, or monocultures, in which energy flow is funneled into crops. He has shortened food chains by selecting and breeding domestic stock that produces meat, milk, or wool by converting forage grasses directly into products useful to man. Successful agriculture depends upon the ecological adaptation of crop plants and domestic animals to environmental conditions and upon the genetic capability of plants to respond to cultural practices and fertilization. Fertilization alters the rate and amount of nutrients entering into ecosystems, and in some instances has undesirable side effects, as when excess phosphates stimulate algal blooms in ponds and lakes. Simplified ecosystems also attract pest species, both plant and animal, requiring costly control methods. Overgrazing of rangelands and mismanagement of crop fields result in wind and water erosion that carries topsoil to rivers and lakes and impoverishes agricultural lands.

Management of forests involves the manipulation of natural stands of trees by means of cutting practices and controlled burning to maintain the forest in a stage of development that favours those species most economically desirable. Management and exploitation of the fish populations for sustained yields requires a thorough knowledge of fish biology and of the essential physical and biological aspects of the environments. Failure to give careful consideration to these points, especially to the dynamics of populations, threatens to extinguish certain oceanic species, thus destroying the fishing industry and removing an important source of protein for man.

Environmental pollution represents a short-circuiting of the biogeochemical cycles, the chemical, or nutrient, cycles in nature. More materials are being poured into the global ecosystem, or biosphere, than can be recycled naturally. As these materials accumulate in the biosphere they have adverse, if not lethal, effects on various organisms. Too much sulfur dioxide in the atmosphere, for example, can kill and injure plants and animals; too much nitrogen and phosphorus in water can cause oxygen-depleting algal blooms that suffocate fish and other aquatic organisms.

Unnatural substances also enter biogeochemical cycles. Chlorinated hydrocarbon pesticides, such as DDT, are circulated by atmospheric and oceanic currents to all parts of the earth, where they are taken up, transferred through, and concentrated in several trophic levels. In strong enough doses they threaten the survival of organisms, such as predatory birds, by impairing metabolism and reproduction. Nuclear contaminants often follow, and become taken up in, normal metabolic cycles, thus presenting a health hazard.

Other man-induced changes in the ecosystem can create equally severe public health problems. In certain tropical regions, for example, the replacement of simple age-old irrigation practices with large-scale dam projects and warm, slow-moving water in miles of irrigation ditches has created ideal living conditions for snails that serve as secondary hosts for parasitic disease organisms. Atmospheric pollutants are responsible for a distinct rise in cases of chronic bronchitis; of nasal, pharyngeal, and optic irritations; and of lung cancer. The health and physical well being of man and the continued existence of his own environment depends on how he applies the principles of ecology toward the solution of many urgent environmental problems.

SOCIETIES AND PUBLICATIONS

Ecology is represented by a number of societies and journals throughout the world. The Ecological Society of America, founded in 1915, publishes two quarterly journals, *Ecology* and *Ecological Monographs*, and a quarterly *Bulletin*. The British Ecological Society, founded in 1913, publishes three journals: the *Journal of Ecology* (devoted to plants), the *Journal of Animal Ecology*, and the *Journal of Applied Ecology*. The Ecological Society of Australia, founded in 1960, publishes an annual *Proceedings of the Ecological Society of Australia*. The Ecological Society of Japan, founded in 1954, sponsors the *Japanese Journal of Ecology*. There are other important journals, with international boards of editors, published for ecologists by major publishing houses. One is *Oikos*, published by Munksgaard International for ecologists in Denmark, Finland, Iceland, and Scandanavia, although the papers it presents are of worldwide interest. *Oecologica*, published by Springer-Verlag, in Germany, in cooperation with the International Association for Ecology (Intecol), contains papers dealing with all aspects of ecology, including theoretical and behavioural ecology and epidemology. The *Journal of Experimental Marine Biology*, published by North Hollands Publishing Company, contains papers on all aspects of marine ecology, experimental work with ecosystems, and general ecological problems. *Limnology* and *Oceanography*, published bv the American Society of Limnology and Oceanography,

present papers on freshwater and marine ecology. Excellent papers on ecology can also be found in the *Journal of Wildlife Management*, published by the Wildlife Society.

BIBLIOGRAPHY. W.C. ALLEE *et al.*, *Principles of Animal Ecology* (1949), a somewhat dated but still important source book in animal ecology, with an excellent early history of ecology; H.G. ANDREWARTHA, *Introduction to the Study of Animal Populations* (1961), a concise summary of an important and controversial earlier text, *The Distribution and Abundance of Animals* by Andrewartha and L.C. BIRCH (1954); PETER FARB, *Ecology* (1963), an enjoyable introduction to ecology, heavily illustrated with outstanding photographs; E.B. FORD, *Ecological Genetics* (1964), a basic reference on adaptation and adjustments of wild populations to their environment; E.J. KORMONDY, *Concepts of Ecology* (1969), a short, interesting introduction to major concepts of ecology, and (ed.), *Readings in Ecology* (1965), a collection of annotated research papers that illustrate many important concepts of ecology, and ed. with R.S. LEISNER, *Ecology* (1971), a collection of essays; G.K. REID, *Ecology of Inland Waters and Estuaries* (1961), a good summary of ecological principles as applied to fresh water and estuarine ecosystems; P. SHEPARD and D. MCKINLEY (eds.), *The Subversive Science: Essays Toward an Ecology of Man* (1969), a collection of thoughtful and provocative papers written by contemporary students of the ecology of human populations; R.L. SMITH, *Ecology and Field Biology* (1966), a comprehensive survey of all aspects of ecology; T.R.E. SOUTHWOOD, *Ecological Methods* (1966), an advanced book on the study of populations, with emphasis on insects and quantitative analysis; G.M. VANDYNE (ed.), *The Ecosystem Concept in Resource Management* (1969), a technical review of the application of ecological theory to forest range, watershed, and wildlife management; K.E. WATT, *Ecology and Resource Management* (1968), a highly technical but important text relating mathematical models and computer programming to resource management; C.B. WILLIAMS, *Patterns in the Balance of Nature and Related Problems in Quantitative Ecology* (1964), an advanced treatise. See also the *Annual Review of Ecology and Systematics.*

(R.L.Sm.)

Econometrics

Econometrics, in the literal meaning of the term, is concerned with measurement in economics. But this is too broad to be descriptive of the subject as it is now studied and practiced, for the fact is that most of economics is concerned with measurement, and many economists are practicing statisticians. The difference is that econometrics seeks to determine and use relationships suggested by general economic analysis but put into mathematical form. It may, for example, seek to determine the relationship between food prices and the sales of food over a period of time (given other relevant variables) and to express that relationship in an equation that can be depicted as a demand curve, as that concept is understood in formal economic analysis. Or the econometrician may examine the relationship between costs and output in a given industry. He may try to find out what sort of mathematical relationship generally prevails between costs and output. Students of business cycles, to take another example, have attempted to discover whether cyclical fluctuations generated from systems of economic equations will throw light on the underlying mechanism of economic cycles. Econometric relationships are not precise: they are intended to demonstrate how actual economic processes evolve in time, but they are estimated as approximate statistical relationships in recognition of their basic probabilistic nature.

Studies of consumption data

The statistical analysis of demand. Many of the earliest studies in econometrics were attempts to quantify the relationship between the prices of a commodity and the amounts of it sold. In theory, the quantity demanded by individual consumers of particular goods and services depends on their incomes and on the relative prices of items the consumers intend to buy. Theory suggests also that changes in prices and income should have fairly predictable effects on the quantities sold.

Econometricians who first applied themselves to analyzing demand used collections of market statistics, compiled over time, known as time series, and studied the relationship between changes in price and changes in demand. Others turned to family-budget statistics that showed how families at various income levels spend their money. These data were used to estimate relationships between income and expenditure. Some studies have combined time series and family budget statistics.

At the practical level, such studies are useful in analyzing price behaviour. They show which commodities are elastic in demand (*i.e.*, have sales relatively responsive to changes in their prices) and which are inelastic (have sales less responsive to price changes). Such information has been useful in determining farm-price-support policy, rate structures for regulated industries, tariff policy, foreign-exchange-rate policy, rationing regulations in wartime, and tax policies. It has also been useful to private business in making price, inventory, and production decisions.

The analysis of production and supply. Econometric analysis has been applied not only to consumer behaviour but also to the behaviour of producers. The approach involves the examination of so-called production functions, cost functions, and supply functions.

Production functions. The production function is a mathematical expression of the technical relationship between the output of a firm and its various inputs (or factors of production). The earliest statistical analyses of the production function were made in an effort to test the theory that labour and capital are reimbursed according to their marginal productivities—*i.e.*, the amounts added to production by the "last" man hired or the "last" unit of capital employed. Studies by the University of Chicago economist Paul Douglas and his collaborator Charles Cobb, using econometric techniques, gave support to the theory. Their studies used a theoretical relationship expressed by an equation:

Studies of labour and capital inputs

$$X = AL^{\alpha}K^{\beta},$$

in which X is total output, L is labour input measured in man-hours, and K is the stock of capital equipment. The values of X, L, and K can be obtained from statistical data for different firms or industries or for all manufacturing industries at different time periods. The values of A, α, and β are then estimated.

If the Cobb–Douglas production function corresponds to reality, labour's share of manufacturing output, according to the marginal productivity theory, will be α, and capital's share will be β. Thus, a comparison of the estimated value for α with the share actually going to labour in the form of wages will provide a test of the theory. Cobb and Douglas found that α ranged between two-thirds and three-fourths in many countries and time periods, while values for β were between one-third and one-fourth. This implied that the sum of α and β tended to be unity, or, to put it another way, that labour and capital divide the total output between them (see DISTRIBUTION, THEORY OF).

More recent work on the production function has been based on a generalized form of the Cobb–Douglas equation, with the implication that the wage rate, adjusted for price changes, is related to labour productivity.

Cost and supply functions. Econometric work in the field of cost functions was originally undertaken to test the theory that marginal cost—the addition to total cost resulting from an increment in output—first declines as production expands but ultimately begins to rise. Econometric studies, however, have indicated that marginal cost often remains more or less constant. Such studies have been made for single plants, for firms, and for whole industries.

Marginal cost curves

Work in estimating supply functions has been confined for the most part to agriculture. Here the problem is to distinguish the effects of exogenous factors, such as temperature, rainfall, and pestilence, from those of endogenous factors, such as changes in prices and inputs. Time is an important element in agricultural supply functions. In tree crops such as cocoa, tea, and coffee, time lags between changes in market price and ensuing changes in production have often been estimated at more than two or three years. Field crops are estimated to have lags of one or two years.

The distribution of income and wealth. One of the well-known empirical regularities of statistical economics, observed in many societies over many generations, is that a small proportion of the population receives most of the income. Attempts have been made to describe this statistical regularity in mathematical terms. These attempts began with the work of an Italian economist and sociologist, Vilfredo Pareto. Efforts have also been made to explain why the distribution of income is so consistent, through the development of a probability process showing how incomes of different sizes are generated and how individuals move from one income class to another with the passage of time. No convincing probability process has yet been established to justify one particular distribution function to the exclusion of all others, but the analysis has enriched the understanding of the methods by which types of distribution functions come into being. This approach has also been applied to the distribution of wealth, the relative size of firms, and other economic phenomena. Distributional measures are frequently important in analyzing the effect of income inequality on aggregate demand.

Studies of the entire economy

Macro-econometric models. *Types of models.* After the mid-1930s the development of national income accounting and of macro-economic theory opened the way for macro-economic model building. This involved attempts to describe an entire economy in mathematical and statistical terms. The first models were those of the Dutch economist Jan Tinbergen in studies of The Netherlands and the United States. These contained equations describing input and output for the economy as a whole (aggregate-production functions), equations describing the share of the factors of production in the distribution of output, and import–export equations. In order to provide a mathematically complete system, equations were added for determining price levels, wage rates, interest rates, and other variables.

A model developed by L.R. Klein and A.S. Goldberger in the United States after World War II was the forerunner of a large family of macro-econometric models. It was constructed on an annual basis and has been elaborated upon in a form known as the Michigan model. A later generation of models was based on quarterly data to permit analysis of short-term movements of the economy and better estimation of the lags between different variables. The Wharton-Econometric Forecasting Unit model at the University of Pennsylvania and the model of the Office of Business Economics (OBE) of the U.S. Department of Commerce differ from the Michigan model in their more detailed lag structures (incorporating delayed time relationships among variables), many separate price equations for various economic sectors, and more monetary equations. The Wharton model, in addition, is divided into manufacturing and nonmanufacturing sectors. Both the Wharton and the OBE models have undergone considerable development and improvement.

A model at the Brookings Institution in Washington, D.C., is distinctive in being a quarterly system with eight separate production sectors. Each sector has its own set of equations for the price level, the wage rate, output, employment, and capital stock. This makes it a large system containing more than 300 simultaneous equations. It is also unique in having a set of equations that describe relations between industries. Work has been done on the Brookings model to enlarge its monetary sector, provide a complete balance-of-payments sector, and expand the set of interindustry equations. The methods developed for the Brookings model to combine interindustry relationships, known as input–output relationships, have been extended to the construction of a larger version of the Wharton model, which has equations for 50 input–output sectors together with more than 100 macro-economic equations.

A model jointly constructed by the U.S. Federal Reserve Board, the Massachusetts Institute of Technology, and the University of Pennsylvania, called the FRB–MIT–PENN model, is specially designed to handle the entire monetary sector. It has a large number of financial equations with a detailed lag structure and enough other equations to show the main directions of monetary influence on the economy.

Similar models have been developed in a number of advanced industrial countries, including Japan, Canada, Great Britain, The Netherlands, and Belgium. Many models have been constructed for developing economies as well, including those of Mexico, Israel, Brazil, and India.

The uses of models. A major purpose in the development of macro models has been improved economic forecasting and the analysis of public policy. They have also been applied to the analysis of economic fluctuations and economic growth.

The impact of the computer

The construction and use of models has, of course, been greatly aided by the electronic computer. In the earliest studies, all calculations had to be done laboriously by mechanical means. This limited the choice of methods to the simplest; it also restricted the size of the systems built and limited their applications. Modern computers enable econometric researchers to be much more experimental in testing ideas; they also enable them to deal with many nonlinear processes and to follow up a wide range of practical applications. Systems can now be solved in seconds. It is possible to consider the response of a whole set of dependent variables to dynamic patterns of change in a set of independent variables. The changes can be introduced with a great deal of flexibility, and the econometrician can make many forecasts, varying them according to different assumptions about alternative events. He can study policy alternatives in a much more complex fashion than in former days, when he was restricted to measuring a single change in one dependent variable resulting from a policy change expressed in a single independent variable.

At any point of time, historical values can be inserted into the equations of an econometric system. Values for the exogenous variables, representing policy commitments or the judgments of the forecaster, can also be inserted into the equations. The system can then be solved for values of the dependent variables—for example, gross national product in the coming year or components of gross national product. Many of the econometric models in the United States, Canada, Japan, The Netherlands, Great Britain, and other countries are regularly used in this way.

The growth of computer technology has also made it easier to study the dynamic behaviour of large systems of equations. It has been found that these systems generate movements that closely conform to historical economic fluctuations. Dynamic analysis by econometric models has thus thrown much light on the underlying mechanism of economic cycles.

BIBLIOGRAPHY. A first-generation general work on modern econometric methodology is TRYGVE HAAVELMO, *The Probability Approach in Econometrics*, supplement to *Econometrica*, vol. 12 (1944). Second-generation general books on methodology are *Studies in Econometric Method*, ed. by W.C. HOOD and TJALLING C. KOOPMANS (1953); and HENRI THEIL, *Economic Forecasts and Policy*, 2nd ed. (1961). These two books also contain some applied studies of substantive interest and are in the form of collections rather than pure systematic development of the subject. LAWRENCE R. KLEIN, *An Introduction to Econometrics* (1962), presents much of the second-generation material at an elementary level, with many examples based on modern micro-econometric studies and the earlier classics. A third generation of general textbooks that systematically develop the whole econometric methodology begins with J. JOHNSTON, *Econometric Methods* (1963); and continues with ARTHUR S. GOLDBERGER, *Econometric Theory* (1964); EDMOND MALINVAUD, *Statistical Methods in Econometrics* (1966); PHOEBUS J. DHRYMES, *Econometrics* (1970); and HENRI THEIL, *Principles of Econometrics* (1971). Three major macro-econometric model studies are described in LAWRENCE R. KLEIN and ARTHUR S. GOLDBERGER, *An Econometric Model of the United States, 1929–1952* (1955); J.S. DUESENBERRY *et al.* (eds.), *The Brookings Quarterly Econometric Model of the United States* (1965); and M.K. EVANS and LAWRENCE R. KLEIN, *The Wharton Econometric Forecasting Model* (1968). These studies deal with the construction, analysis, and application of macro models and are not primarily methodological.

(L.R.K.)

Economic Development

National income per capita is the conventional criterion of economic development and underdevelopment. This criterion may be applied to the level or to the rate of growth of per capita income. The "underdeveloped" countries are usually defined as those with a much lower level of per capita income than the "developed" countries. The income per capita definition of underdevelopment is based on the selection of an arbitrary level of per capita income (for example, $500 a year); the countries with per capita income below this level are put in the underdeveloped category.

It is necessary to consider some of the statistical and conceptual difficulties of using the conventional definition of underdevelopment before analyzing the causes of underdevelopment. The statistical difficulties are well-known. To begin with, there are the awkward borderline cases. Even if analysis is confined to the underdeveloped countries in Asia, Africa, and Latin America, there are rich oil countries, such as Kuwait or Venezuela, that have per capita incomes well above the demarcation line of $500 but that are otherwise underdeveloped in their general economic characteristics. Second, there are a number of technical difficulties that make the per capita incomes of many underdeveloped countries (expressed in terms of an international currency, such as the U.S. dollar) a very crude measure of their per capita real income. These difficulties include the defectiveness of the basic national income and population statistics, the inappropriateness of the official exchange rates at which the national incomes in terms of the respective domestic currencies are converted into the common denominator of the U.S. dollar, and the problems of estimating the value of the noncash components of real incomes in the underdeveloped countries. Finally, there are conceptual problems in interpreting the meaning of the international differences in the per capita income levels, discussed below.

Difficulties of measurement

The theoretical background

MEANINGS OF UNDERDEVELOPMENT

Material poverty and subjective discontent. When people talk about the problem of poverty of the underdeveloped countries, they seem to have in mind not one problem but two distinct problems, implying two different interpretations of the low per capita incomes of these countries. There are those who are concerned with the underdeveloped countries on humanitarian grounds; that is, with the problem of helping the people of these countries to attain certain minimum material standards of living in terms of such factors as food, clothing, shelter, and nutrition. For them, low per capita income is the measure of the problem of poverty in a material sense: the aim of economic development is to improve the material standards of living by raising the absolute level of per capita incomes. There are others who are concerned with the underdeveloped countries from the standpoint of international relations and in terms of reducing international tensions. They tend to see the low per capita incomes of the underdeveloped countries in relative terms; that is, in relation to the high per capita incomes of the developed countries. For them, even if an underdeveloped country is able to improve its material standards of living through a rise in the level of its per capita income, it may still be faced with the more intractable subjective problem of discontent created by the widening gap in the relative income levels between itself and the rich underdeveloped countries and the developed countries.

The interpretation of a low per capita income level as an index of poverty in a material sense may be accepted with two qualifications. First, the level of material living depends not on per capita income as such but on per capita consumption. The two may differ considerably when a large proportion of the national income is diverted from consumption to other purposes; for example, through a policy of forced saving. Second, the poverty of a country is more faithfully reflected by the representative standard of living among the great mass of its people. This may be well below the simple arithmetic average of per capita income or consumption when national income is unequally distributed and there is a wide gap in the standard of living between the rich and the poor.

The widening gap in incomes. More complicated issues are raised by the attempt to interpret the low per capita income level or, more specifically, the widening gap in the per capita incomes between the developed and the underdeveloped countries as an index of the subjective problem of discontent of the latter group of countries. To begin with, it should be noted that the "widening gap" may refer (1) to the gap in the rates of growth in per capita incomes and (2) to the gap in the levels of per capita incomes in a process of growth.

Broadly speaking, the growth rate in the total gross domestic product (GDP) of the underdeveloped countries has not lagged behind that of the developed countries. What has pulled down the rate of growth of their per capita product is not their slower rate of growth in output but their faster population growth. Thus, during the period 1960–67 the growth rate of the GDP of the underdeveloped countries was 5 percent, but their population growth was 2.5 percent, giving them a growth rate in per capita product of 2.5 percent. The growth rate of the GDP of the developed countries, on the other hand, was 4.8 percent, but their population growth was 1.2 percent, giving them a higher growth rate in per capita product of 3.6 percent. One of the fears behind the notion of the widening gap is that the underdeveloped countries will be held down by the vicious circle of low income, low savings, and economic stagnation. The available evidence casts considerable doubts on this theory. The savings rate of the underdeveloped countries has risen much higher than was thought possible in the 1950s, when the target saving rate was put at 10–12 percent of the GDP. During the period 1960–67 the gross saving rate among the Asian, African, and Latin American underdeveloped countries ranged between 11 and 16 percent of their GDP (excluding the special group of south European underdeveloped countries where both the saving ratio and the growth rate were higher). Further, among the Asian, African, and Latin American underdeveloped countries there does not seem to be an inevitable connection between a low level of per capita income as such and a slower rate of economic growth. During the period 1960–66, for example, very high rates of growth in total product of 5–7 percent or above have been attained by the relatively poor underdeveloped countries, such as South Korea, Thailand, the Philippines, Pakistan, and Nigeria, with per capita incomes between $100 and $200. On the other hand, rates of growth below 3 percent have been recorded for the relatively rich underdeveloped countries, such as Argentina and Uruguay, with per capita incomes above $500. This suggests that, given the appropriate economic policies, including population-control policies, there need not be any immutable reason why the growth rates of the underdeveloped countries should always lag behind those of the developed countries. There is, however, a further and arithmetically more intractable aspect of the widening gap. This arises simply from the operation of the arithmetic of growth on the large initial gap between the income levels of the developed and the underdeveloped countries. As an example, an underdeveloped country with a per capita income of $100 and a developed country with a per capita income of $1,000 may be considered. The initial gap in their incomes is $900. Let the incomes in both countries grow at 5 percent. After one year, the income of the underdeveloped country is $105, and the income of the developed country is $1,050. The gap has widened to $945. The income of the underdeveloped country would have to grow by 50 percent to maintain the same absolute gap of $900.

Effect of population growth

The crucial question is how far this sort of arithmetic can be related to the complex phenomenon of the subjective problem of the discontent of the underdeveloped countries with their economic status relative to that of the developed countries. First, the underdeveloped coun-

tries' sense of dissatisfaction and grievance arises not only from measurable differences in national incomes but also from the less easily measurable factors, such as their reaction against the colonial past and their complex drives to raise their national prestige and achieve equality in the broadest sense with the developed countries. Thus, it is not uncommon to find their governments using a considerable proportion of their resources in prestige projects, ranging from steel mills, hydroelectric dams, universities, and defense expenditure to international athletics. These symbols of modernization may contribute a nationally shared satisfaction and pride but may or may not contribute to an increase in the measurable national income. Second, it is possible to argue that in many cases the internal gap in incomes within individual underdeveloped countries may be a more potent source of the subjective level of discontent than the international gap in income. Faster economic growth may help to reduce the internal economic disparities in a less painful way, but it must be remembered that faster economic growth also tends to introduce greater disruption and the need for making bigger readjustments in previous ways of life and may thus increase the subjective sense of frustration and discontent. Finally, it is difficult to establish that the subjective problem of discontent will bear a simple and direct relationship to the size of the international gap in incomes. Some of the apparently most discontented countries are to be found in Latin America, where the per capita incomes are generally higher than in Asia and Africa. A skeptic can turn the whole approach to a *reductio ad absurdum* by pointing out that even the developed countries with their high and rising levels of per capita income have not been able to solve the subjective problem of discontent and frustration among various sections of their population.

Toward a workable criterion

Two conclusions may be drawn from the above points. First, the subjective problem of discontent in the underdeveloped countries is a genuine and important problem in international relations. But economic policy acting on measurable economic magnitudes can play only a small part in the solution of what essentially is a problem in international politics. Second, for the narrower purpose of economic policy there is no choice but to fall back on the interpretation of the low per capita incomes of the underdeveloped countries as an index of their poverty in a material sense. This can be defended by explicitly adopting the humanitarian value judgment that the underdeveloped countries ought to give priority to improving the material standards of living of the mass of their people. But, even if this value judgment is not accepted, the conventional measure of economic development in terms of a rise in per capita income still retains its usefulness. The governments of the underdeveloped countries may wish to pursue other, nonmaterial goals, but they would be able to make clearer decisions if they knew the economic cost of their decisions. The most significant measure of this economic cost can be expressed in terms of the foregone opportunity to raise the level of per capita income.

A SURVEY OF DEVELOPMENT THEORIES

The hypothesis of underdevelopment. If the underdeveloped countries are merely low-income countries, why call them underdeveloped? The use of the term underdeveloped in fact rests on a general hypothesis on which the whole subject matter of development economics is based. According to this hypothesis, the existing differences in the per capita income levels between the developed and the underdeveloped countries cannot be accounted for purely in terms of differences in natural conditions beyond the control of man and society: that is to say, the underdeveloped countries are underdeveloped because, in some way or another, they have not yet succeeded in making full use of their potential for economic growth. This potential may arise from the underdevelopment of their natural resources, of their human resources, or from the "technological gap." More generally, it may arise from the underdevelopment of economic organization and institutions, including the network of the market system and the administrative machinery of the government. The general presumption is that the development of this organizational framework would enable an underdeveloped country to make a fuller use not only of its domestic resources but also of its external economic opportunities, in the form of international trade, foreign investment, and technological and organizational innovations.

Growth economics and development economics. The task of development economics is to study how the underdeveloped countries may be enabled to make a fuller use of their potential opportunities for economic development. This may be contrasted with another branch of study, called growth economics, which is concerned with the study of the long-run, or steady-state, equilibrium growth paths of the economically developed countries, which have long overcome the problem of initiating economic development.

Growth theory assumes the existence of a fully developed modern capitalist economy with a sufficient supply of entrepreneurs responding to a well-articulated system of economic incentives to drive the growth mechanism. Typically, it concentrates on macroeconomic relations, particularly the ratio of savings to total output and the aggregate capital–output ratio (that is, the number of units of additional capital required to produce an additional unit of output). Mathematically, this can be expressed (the Harrod–Domar growth equation) as follows: the growth in total output (g) will be equal to the savings ratio (s) divided by the capital–output ratio (k); *i.e.*, $g = \frac{s}{k}$. Thus, suppose that 12 percent of total output is saved annually and that three units of capital are required to produce an additional unit of output: then the rate of growth in output is $\frac{12}{3}\% = 4\%$ per annum. This result is obtained from the basic assumption that whatever is saved will be automatically invested and converted into an increase in output on the basis of a given capital–output ratio. Since a given proportion of this increase in output will be saved and invested on the same basis, a continuous process of growth is maintained.

Misapplication of growth theory

Growth theory, particularly the Harrod–Domar growth equation, has been frequently applied or misapplied to the economic planning of an underdeveloped country. The planner starts from a desired target rate of growth of perhaps 4 percent. Assuming a fixed overall capital–output ratio of, say, 3, it is then asserted that the underdeveloped country will be able to achieve this target rate of growth if it can increase its savings to $3 \times 4\% = 12\%$ of its total output. The weakness of this type of exercise arises from the assumption of a fixed overall capital–output ratio, which assumes away all the vital problems affecting the underdeveloped country's capacity to absorb capital and invest its saving in a productive manner. These problems include the central problem of the efficient allocation of available savings among alternative investment opportunities and the associated organizational and institutional problems of encouraging the growth of a sufficient supply of entrepreneurs; the provision of appropriate economic incentives through a market system that correctly reflects the relative scarcities of products and factors of production; and the building up of an organizational framework that can effectively implement investment decisions in both the private and the public sectors. Such problems, affecting the underdeveloped country's absorptive capacity for capital and other inputs, constitute the core of development economics. Development economics is needed precisely because the assumptions of growth economics, based as they are on the existence of a fully developed and well-functioning modern capitalist economy, do not apply.

A "new orthodoxy" in development economics. The underdeveloped countries, even excluding those in southern Europe, still remain a very mixed collection of countries. They differ widely in area, population density, and natural resources. They are also at different stages in the development of market and financial institutions and of an effective administrative framework. These differences

Reasons for the change in theoretical models

are sufficient to warn against wide-sweeping generalizations about the causes of underdevelopment and all-embracing theoretical models of economic development. But when development economics first came into prominence in the 1950s, there were powerful intellectual and political forces propelling the subject toward such general theoretical models of development and underdevelopment. First, many writers who popularized the subject were frankly motivated by a desire to persuade the developed countries to give more economic aid to the underdeveloped countries, on grounds ranging from humanitarian considerations to considerations of cold war strategy. Second, there was the reaction of the newly independent underdeveloped countries against their past "colonial economic pattern," which they identified with free trade and primary production for the export market. These countries were eager to accept general theories of economic development that provided a rationalization for their deep-seated desire for rapid industrialization. Third, there was a parallel reaction, at the academic level, against older economic theory, with its emphasis on the problem of the efficient allocation of scarce resources and a striving after new and "dynamic" approaches to economic development.

All these combined to produce a crop of theoretical approaches that soon developed into a "new orthodoxy" with its characteristic emphasis on "crash" programs of investment in both material and human capital, on domestic industrialization, and on government economic planning as the standard ingredients of development policy. These theories still continue to have a considerable influence on the conventional wisdom in development economics, although in retrospect most of them have turned out to be partial theories. A broad survey of these theories, under three main heads, is given below. It brings out their interrelationships and is particularly relevant to the debate over whether the underdeveloped countries should seek economic development through domestic industrialization or through international trade. The limitations of these new theories—and how they have led to a gradual revival of a more pragmatic approach to development problems which falls back increasingly on the older economic theory of efficient allocation of resources—are then traced.

The missing-component approach. First, there are the theories that regard the shortage of some strategic input (such as the supply of savings, foreign exchange, or technical skills) as the main cause of underdevelopment. Once this missing component was supplied—say, by external economic aid—it was believed that economic development would follow in a predictable manner based on fixed quantitative relationships between input and output. The overall capital–output ratio, mentioned above, is the most well-known of these fixed technical coefficients. But similar fixed coefficients have been assumed between the foreign-exchange requirements and total output and between the input of skilled manpower and output.

Savings. Given the broad relationship between capital accumulation and economic development, it was plausible to argue in the 1950s that the underdeveloped countries were held back mainly by a shortage in the supply of capital. These countries were then saving only about 5–7 percent of their total product, and the raising of the saving ratio to 10–12 percent was frequently regarded as the central problem of economic development. But the saving ratios in the underdeveloped countries have been rising in recent decades until they have reached or passed this target level, and the saving-shortage theory has thus become less plausible. During the same period it has increasingly come to be recognized that there is much wastage of capital resources in the underdeveloped countries for various reasons, such as wrong choice of investment projects, inefficient implementation and management of these projects, and inappropriate pricing and costing of output. These faults are particularly noticeable in public-sector investment projects and are one of the reasons why the Pearson Commission Report of the International Bank for Reconstruction and Development

Aid and growth

(1969) finds that "the correlation between the amounts of aid received in the past decades and the growth performance is very weak." But even in the private sector there has been a considerable distortion in the direction of investment induced by policies designed to encourage import substitution. Thus, in most underdeveloped countries, a considerable part of private expansion investment, both foreign and domestic, has been diverted into the expansion of the manufacturing sector, catering to the domestic market through various inducements, including tariff protection, tax holidays, cheap loans, and generous foreign-exchange allocations granting the opportunity to import capital goods cheaply at overvalued exchange rates. As a consequence, there has developed a very considerable amount of excess capacity in the manufacturing sector of the underdeveloped countries pursuing import-substitution policies.

Foreign exchange. Since many of these manufacturing industries are high-cost industries with no reasonable prospects of becoming internationally competitive, their situation may be taken as evidence of the underdeveloped countries' inability to absorb capital. But it has frequently been argued that their lack of economic viability is caused primarily by a shortage of foreign exchange to buy essential industrial inputs. There is thus a second version of the missing-component approach, according to which the underdeveloped countries are held back because they are unable to expand their export receipts sufficiently to pay for the technologically irreducible minimum imports of capital goods and other industrial inputs required to sustain a reasonable rate of economic growth. This concept of technologically determined foreign-exchange requirements may now be examined.

The import-substitution approach to economic growth defines "industrialization" rather narrowly as the expansion of the modern manufacturing sector based on capital-intensive technology. Capital is therefore identified with durable capital equipment in the form of complex machinery and other inputs that the underdeveloped countries are not able to produce domestically. Thus, foreign-exchange requirements are calculated on the basis of the fixed technical input–output coefficients of the manufacturing sector.

The expanded view of industrialization

But an entirely different perspective on capital and foreign-exchange requirements is arrived at if a broader view of industrialization, more relevant to economic development, is adopted. For this purpose, industrialization may be regarded as the application of modern science, technology, and management to the task of raising the productivity of resources, not only in the manufacturing sector but in the economy as a whole, including the agricultural sector. If the aim of economic development is to increase the total national output, the concept of capital cannot be restricted to durable capital goods and complex forms of machinery. In the setting of the underdeveloped countries there are other forms of capital that are likely to contribute more to an increase in total output. The underdeveloped countries are characterized (1) by a large share of agriculture in their total output; (2) by the inadequacy of their social-overhead capital, particularly transport, communications, and irrigation; and (3) by the relatively large share of "nonreproducible" capital or natural resources in their total stock of productive equipment. In this setting, it must be expected that investment will improve the productivity of the inherited natural resources by the construction of a more adequate system of social-overhead capital, particularly transport and irrigation. This type of capital formation consists largely of construction work that has to be performed on the spot, as contrasted with the importation of machinery in prefabricated form. In particular, it requires large quantities of two types of inputs that the underdeveloped countries can advantageously produce at home. First, it requires various building materials, such as brick, cement, and tiles, that because of their weight, bulk, and heavy transport costs offer a natural advantage to local production. Second, and even more important, construction work requires the employment of a large number of workers engaged in the production of future

output instead of present output; and these workers must be supported by a subsistence fund of consumers' goods to maintain them during the waiting period before their work yields a return in the form of an increased quantity of consumable products. It can thus be seen why the older classical notion of capital as a subsistence fund is likely to be more important for the underdeveloped countries than the modern concept of capital as durable capital goods. The modern concept of capital is a reflection of the setting in the advanced countries in which agriculture forms only a small proportion of the total output and in which reproducible, or man-made, capital forms the bulk of productive equipment.

Weaknesses of the foreign-exchange-shortage theory

If this analysis is correct, the foreign-exchange-shortage theory as a species of the missing component suffers from two serious weaknesses. First, by emphasizing too narrow a concept of industrialization and capital, it distracts attention from the considerable potentialities of economic development open to the underdeveloped countries; namely, by import substitution in the production of the simpler tools and construction materials that they are likely to require in large quantities in the construction of their social overhead capital and by improving the productivity of the agricultural sector, which is the chief supplier of capital as subsistence fund. Second, if a large number of the manufacturing industries suffering from excess capacity have no reasonable prospects of increasing their efficiency and reducing their costs to become internationally competitive, then it is a waste of foreign-exchange resources to keep them running at their full technical capacity. Even if there were enough foreign exchange available to do so, it would be more economic to scrap these industries and divert the foreign exchange saved in this way to more productive uses, both in the domestic manufacturing sector and in the agricultural and export sectors.

Education. The third version of the missing-component approach regards the underdeveloped countries as being held back mainly by a lack of education and skills. If education and skill are defined as everything that is required to raise the productivity of the people in the underdeveloped countries by improving their skills, enterprise, initiative, adaptability, and attitudes, this proposition is true but is an empty tautology. This aspect of the missing-component approach, however, is formulated in terms of specific skills and educational qualifications that can be supplied by crash programs in formal education through massive expansion of schools, universities, and technical-training centres. The usual method of manpower planning thus starts from a target rate of expansion in output and tries to estimate the numbers of various types of skilled personnel that would be required to sustain this target rate of economic growth on the basis of fixed technical coefficients between inputs of skill and national output.

This approach was plausible enough in many underdeveloped countries immediately after their political independence, when there were obvious gaps in various branches of the administrative and technical services created by the departure of the nationals of the former colonial powers. It was then obviously important for the underdeveloped countries to train their own nationals as quickly as possible to fill these gaps. But most underdeveloped countries have already passed through this initial phase. In the meantime, as the result of programs in education expansion, their schools and colleges are annually producing large numbers of fresh graduates at much faster rates than their general rate of economic growth. This creates a growing problem of educated unemployment, which is one of the most potent causes of the subjective problem of discontent in the underdeveloped countries. An important factor behind the rapid educational expansion has been the expectation that students would obtain well-paid white-collar jobs at salary levels many times the prevailing per capita income of their countries. Thus, the underdeveloped countries' inability to create jobs to absorb their growing armies of graduates is an explosive element in what has come to be called the revolution of rising expectations.

There is a parallelism between the narrow concept of industrialization as the expansion of the manufacturing sector and the narrow concept of education as the academic and technical qualifications that can be supplied by the expansion of the formal educational system. If a broader concept of education, relevant for economic development, is needed, it is necessary to seek it in the pervasive educational influence of the economic environment as a whole on the learning process of the people of the underdeveloped countries. This is a complex process that depends, among other less easily analyzable things, on the system of economic incentives and signals that can mold the economic behaviour of the people of the underdeveloped countries and affect their ability to make rational economic decisions and their willingness to introduce or adapt to economic changes. Unfortunately, the economic environment in many underdeveloped countries is dominated by a network of government controls that tend not to be conducive to such ends. Indigenous entrepreneurs often find it more rewarding to devote their energy and ingenuity to the task of procuring the necessary government import licenses and other permits and exploiting the loopholes in government regulations than to the problem of raising the efficiency and productivity of resources.

Surplus resources and disguised unemployment. Next to be considered are theories that emphasize the surplus resources of the underdeveloped countries as the principal means of promoting economic development. There are two contrasting theories. The first concentrates on the countries with relatively abundant natural resources and low population densities and argues that a considerable amount of both surplus land and surplus labour may still exist in these countries because of inadequate marketing facilities and lack of transport and communications. Economic development is pictured as a process whereby these underutilized resources of the subsistence sector are drawn into cash production for the export market. International trade is regarded as the chief market outlet, or vent, for the surplus resources. This theory will be considered later. For the moment, it is useful to concentrate on the second theory, which is concerned with the thickly populated countries and the possibility of using their surplus labour as the chief means of promoting economic development. According to this theory, because of heavy population pressure on land, the marginal product of labour (that is, the extra output derived from the employment of an extra unit of labour) is reduced to zero or to a very low level. But the people in the subsistence sector are able to enjoy a certain customary minimum level of real income because the extended-family system of the rural society shares the total output of the family farm among its members. A considerable proportion of labour in the traditional agricultural sector thus contributes little or nothing to total output and is really in a state of disguised unemployment. This labour may therefore be drawn into other uses without any cost to society.

The concept of disguised unemployment

It is necessary to clear up a number of preliminary points about the concept of disguised unemployment before considering its applications. First, it is highly questionable whether the marginal product of labour is actually zero even in densely populated countries, such as India or Pakistan. Even in these countries, with existing agricultural methods, all available labour is needed in the peak seasons, such as harvest. The most important part of disguised unemployment is thus what may be better described as seasonal unemployment during the off-seasons. The magnitude of this seasonal unemployment, however, depends not so much on the population density on land as on the number of crops cultivated on the same piece of land through the year. There is thus little seasonal unemployment in countries such as Taiwan or South Korea, which have much higher population densities than India, because improved irrigation facilities enable them to grow a succession of crops on the same land throughout the year. But there may be considerable seasonal unemployment even in sparsely populated countries growing only one crop a year.

The main weakness in the proposal to use disguised unemployment for the construction of major social-overhead-capital projects arises from an inadequate consideration of the problem of providing the subsistence fund to maintain the workers during the waiting period before these projects yield consumable output. This may be managed somehow for small-scale local-community projects when the workers are maintained *in situ* by their relatives. But when it is proposed to move a large number of surplus workers away from their home villages for major construction projects taking a considerable time to complete, the problem of raising a sufficient amount of subsistence fund to maintain the labour becomes formidable. The only practicable way of raising such a subsistence fund is to encourage voluntary saving and the expansion of a marketable surplus of food that can be purchased with the savings to maintain the workers. The mere existence of disguised unemployment does not in any way ease this problem.

Critical minimum-effort theories. Third are those theories based on the proposition that the underdeveloped countries are firmly held down to their low per capita incomes by a series of interlocking vicious circles. Thus, a gradual or step-by-step effort to raise incomes would be of no avail: only "a big push" exceeding a certain critical minimum effort would give sufficient power to lift the underdeveloped countries out of their low-income equilibrium trap.

The population trap

One version of this theory holds that any gradual increase in total income would be swallowed up by an increase in population without raising the per capita income. Thus, a critical minimum effort is necessary to raise the rate of growth in total income above the maximum possible rate of growth in population, so that per capita incomes can be raised to generate a margin of savings above the minimum subsistence level, thus providing a means of escape from the low-level equilibrium trap. But this theory, as originally expounded, has been overtaken by events. Total national incomes in the underdeveloped countries have been rising at the average of 5 percent, compared with a population growth of 2.5 percent, giving them an average rise in per capita income of 2.5 percent per annum during most of the postwar period. In countries where the level of per capita income has remained constant, such as Argentina or Indonesia, the main cause seems to have been internal political instability rather than simple population pressure. This theory nevertheless serves a useful purpose in emphasizing the burden of population growth to the underdeveloped countries. The drawback of a rapidly increasing population is not only that it depresses the per capita level of income and consumption: it also requires a much larger amount of savings to equip the rapidly growing number of workers with the same amount of capital per head. But this may be more usefully interpreted as an argument for controlling population growth than as an argument for a critical minimum effort in increasing output.

A more elaborate and influential argument for a critical minimum effort is based on the need to overcome the smallness in the size of the domestic market resulting from the low incomes and purchasing power of the consumers in the underdeveloped countries. It is argued that this prevents the underdeveloped countries from taking advantage of the economies of large-scale production and the economies generated by a complementary group of industries. Thus, a single modern factory may fail because of the smallness of the market outlet for its output. What is needed is the simultaneous setting up of a group of factories producing consumer goods, so that among them they provide a sufficient volume of additional employment and purchasing power to provide a sufficiently large market for each other's products. The output of the consumers' goods produced by each factory should be expanded in a balanced way so that the supply matches the increased demand for the product generated by the expansion in incomes. The more ambitious versions of the theory then extend this framework to include balanced-growth relationships between consumer-goods industries and capital-goods industries and balanced-growth relationships between the whole of the manufacturing sector and the construction of new social-overhead capital. The labour supply for the new industries is assumed to be freely available from the disguised unemployment from agriculture. What is required is a big push in the form of a massive injection of external capital resources to finance the balanced-growth program of investment.

Objections to the big-push and balanced-growth approaches

Two main objections may be laid against the big-push and balanced-growth versions of the critical minimum-effort argument.

First, even if it is assumed that the raw labour from agriculture can be trained in sufficient numbers, it is still necessary to solve the problem of finding an army of entrepreneurs and managers to run a whole flock of brand-new industries. Further, there is the crucial problem of coordinating a comprehensive balanced-growth program. Planning, in the sense of drawing up on paper a formally consistent set of projects, is only the beginning. The really difficult part lies in the implementation of the various projects according to a planned timetable and keeping the plans in each sector in line with the rest. The greater the interdependence among the different components of the program, the greater is the likelihood that the delay or failure in one part of the program will throw the whole plan out of gear. The balanced-growth program is supposed to be carried out in a setting of a mixed economy (that is, an economy with substantial public and private sectors) with the government agencies taking a leading part. But the complex task of coordination required by the balanced-growth approach is well beyond the administrative capabilities of most of the underdeveloped countries. Further, the drive to make the investment programs as large as possible frequently strains the available resources and pushes the government continually to a position of having to take "crisis" measures, usually in the form of a sudden tightening of controls, particularly foreign-exchange controls. This adds to the uncertainties that inhibit private investment. In most underdeveloped countries the diseconomies generated by faulty coordination of investment plans and the shortage of competent entrepreneurs and managers would in practice outweigh the theoretical advantages claimed for a large-scale, integrated balanced-growth program.

Second, the balanced-growth approach is implicitly based on an autarkic policy of matching the expansion of domestic production to the expansion of the domestic market. It leaves out the balancing of demand and supply through exports and imports. Such an autarkic approach may be plausible for a very large underdeveloped country, such as India, although even there it has been a major factor responsible for slow economic growth. But the majority of the underdeveloped countries are very small countries. Out of some 90 of them, 72 have fewer than 15,000,000 population and 51 have fewer than 5,000,000 population. For the smaller countries, balanced growth within their national economy makes little economic sense. The economies of large-scale production must be sought in specialization for the export market.

INDUSTRIALIZATION VERSUS TRADE

The central issue of development policy may now be considered; that is, the relative importance of domestic industrialization and international trade as the principal factor in growth for the underdeveloped countries. The case for domestic industrialization, which underlies much of the new orthodoxy in development economics, is made up of negative arguments against international trade combined with positive arguments for the expansion of manufacturing as the leading sector in economic development.

Antitrade arguments

There are three main antitrade arguments. First, it is maintained that the pattern of international trade of the 19th century, in which foreign investment in mines and plantations led to an expansion of primary exports from the underdeveloped countries, did not adequately transmit the benefits of export expansion to the indigenous

people. The mines and plantations are said to be enclaves of foreign capital and economic organization creating a dualism between the modern export sector and the traditional sector of the rest of the country. Second, it is argued that the world-market demand for primary products is subject to violent fluctuations; thus, the underdeveloped countries, which specialize in a few primary products, are vulnerable to short-run fluctuations in their export receipts and need to insulate themselves against these fluctuations by encouraging domestic industries. Third, it is argued that, because of low-income elasticity of demand for primary products (*i.e.*, because less is spent on them as national incomes rise) and the substitution of synthetic materials for natural raw materials, there is a declining long-run trend in the world-market demand for primary products. Thus, export expansion in these products no longer offers an "engine of growth" to the underdeveloped countries.

The argument for industrialization

The positive argument for industrialization is based on the assumption that capital-intensive modern technology in manufacturing would raise labour productivity above its low level in the traditional-handicraft and the agricultural sectors. Because of this, it is argued, the modern manufacturing sector would be able to earn a high rate of profits; the re-investment of these profits provides the mainspring of economic development, which draws the surplus labour from agriculture. The existence of disguised unemployment in the agricultural sector means that a virtually unlimited supply of labour is available to the manufacturing sector at a constant wage rate, enabling the manufacturing sector to continue earning a high rate of profit for further investment.

The underdeveloped countries undoubtedly suffered the traumatic experience of a drastic fall in their export receipts during the Great Depression of the 1930s. The question arises of how far a delayed response to this experience is relevant in the postwar period; for international trade since World War II has been generally buoyant, and a second great depression has been avoided by the full-employment policies pursued in the advanced countries. The only notable recession in the postwar period has been the decline in raw-material prices from the exceptionally high levels they reached during the Korean War boom. As to whether the export earnings of the underdeveloped countries producing primary products are really more unstable than those of the developed countries in the postwar period, systematic empirical studies by a number of economists have cast considerable doubts on this once widely accepted proposition. Further, there is a greater recognition of the fact that fluctuations in export earnings depend on both the fluctuations in prices determined by the world-market factors and the fluctuations in quantities determined by the domestic factors in the underdeveloped countries. These include harvest fluctuations, political instability, and government marketing policies affecting the supply of exports. This leads to a general assumption underlying much of the argument for domestic industrialization and increasing national self-sufficiency. While the underdeveloped countries cannot control the world-market factor governing their international trade, they have an effective control over the internal factors governing their domestic economic activities. This is questionable on a closer analysis. The internal economic organization in many underdeveloped countries is extremely "brittle" and subject to frequent breakdowns and speculative instability of prices in the face of unforeseen changes, such as a bad harvest threatening famine conditions in a particular locality. The reasons for this arise from poor transport and communications, inadequate inventories of goods, lack of confidence aggravated by domestic inflation, and a generally underdeveloped market system. Thus, it is possible to argue that the underdeveloped countries' vulnerability to internal shocks can be reduced by freer international trade offering an elastic world-market supply of goods on which it can rely in its frequent emergencies. To take advantage of this, the underdeveloped countries would have to use a larger part of their export earnings to build up their foreign-exchange reserves instead of committing their foreign-exchange receipts up to the hilt in imported machinery and durable capital goods. This is another point confirming the significance of the classical notion of capital as circulating capital as contrasted with the modern concept of capital as fixed capital.

Factors affecting export performance

World-market demand. Global projections about future trends in the world-market demand for primary products based on broad secular factors are somewhat speculative and open to disagreement. The fact that the income elasticity of demand for these products is likely to be low because people demand proportionately less of food and basic consumer goods as their income rises can be countered by the broad factors working in the opposite direction, including the growth in world population and the worldwide depletion of natural resources. From the point of view of an underdeveloped country seeking to expand its exports, however, the more directly relevant consideration is not the size of aggregate world-market demand for primary products but its own share in the total market in a particular product. This depends on the relative costs and prices at which it can sell its exports compared with the rival sources of supply; that is to say, on the price elasticity of demand rather than the income elasticity of demand. A country's competitiveness in the export market is determined by two main factors: first, by the economic factors determining its comparative costs; and second, by the economic factors affecting the degree of its domestic inflation and the overvaluation of its currency at fixed exchange rates. Here again, detailed case by case studies of the underdeveloped countries' scope for increasing their exports have cast considerable doubts on the older pessimism about the unfavourable world-market-demand factors inhibiting their ability to expand exports. Depending on random factors, some countries may find it difficult to expand exports because they are already producing a large share of the total world supply or because the world-market demand for these particular products is stagnant. But, in the buoyant postwar conditions of international trade, this still leaves a wide range for expanding exports that do not suffer from these handicaps. Appropriate policies, such as the control of inflation and correction of the overvalued currency, combined with a diversion of capital and entrepreneurial resources from domestic industrialization to export expansion, can still do much to increase the foreign-exchange earnings of an underdeveloped country. Further, it should be remembered that the tendency to substitute synthetic for natural raw materials is not merely a matter of autonomous technical innovations; it is determined by the relative cost and availability of rival products. A country such as Malaysia has thus been able to maintain its exports of natural rubber in the face of competition by synthetic rubber by appropriate policies of research and investment that have increased the yield of natural rubber. Export expansion would still seem to provide an engine of growth for those countries that make appropriate efforts to increase the competitiveness of their exports.

Promotion of manufacturing industry as a source of growth

Development of domestic industry. The positive case for the expansion of the manufacturing sector to supply the domestic market may now be considered. It is based on the general assumption that by adopting modern capital-intensive technology, the manufacturing sector would raise the productivity of labour and become the leading sector, drawing in the disguised unemployment from the traditional agricultural sector. This familiar two-sector growth model is a formalization of the historical experience of countries such as England and Japan, which have successfully passed through the development process. How far it can be applied to the underdeveloped countries of the post-World War II period is a matter of some conjecture. In writings on the subject, the main emphasis is placed on the availability of an unlimited supply of labour at a constant wage to the manufacturing sector, enabling the latter to continue earning high profits for re-investment. But this distracts attention from the fact that the labour supply can play merely a passive role in the process. The dynamic mainspring of the development process has to be found in the ability of the manufactur-

ing sector to raise labour productivity and earn high profits. The crucial question, therefore, is how far the adoption of modern capital-intensive technology, by itself, would enable the manufacturing sector in the underdeveloped countries to raise productivity and economic efficiency. This returns to the fundamental problem of the underdeveloped countries' capacity to absorb capital and modern technology. As has been shown, most underdeveloped countries have tried to expand their domestic manufacturing sector and to substitute for imports by a variety of policies granting protection and subsidies. Yet a large proportion of the manufacturing industries have remained economically inefficient in the sense that they cannot withstand free competition with the imported goods. Another way of describing the situation is that these industries would make no profit or show losses if their outputs and inputs were valued at world-market prices. Yet within the closed-economy (that is, isolated from the world market) situation of many underdeveloped countries, these manufacturing industries seem to enjoy high profits in terms of the local currency, frequently in spite of a substantial excess capacity. These profits seem to arise not from a genuine increase in productivity through the adoption of modern technology but from the artificial economic distortions created by government policies. On the one hand, high protection, quantitative import controls, and domestic inflation combine to create a captured market yielding easy profits for the manufacturing industries. On the other hand, these industries enjoy subsidies in the form of low-interest loans, tax concessions, and a liberal allowance of foreign exchange, enabling them to import capital goods cheaply at overvalued exchange rates. The real economic resources that have gone into the expansion of the modern manufacturing sector in the underdeveloped countries thus seem to have been obtained not so much from a self-generating process of increase in productivity creating an expanding stream of re-investable resources but from the transfer of resources from other sectors, particularly from the export and the domestic agricultural sector.

Detrimental effects of import-substitution policies

As noted above, one of the arguments used against the 19th-century pattern of the expansion of primary exports was that foreign investment in the mines and plantations introduced an economic dualism between the export sector and the traditional sector. In many ways, the import-substitution policies pursued by the underdeveloped countries have aggravated this dualism between the manufacturing sector in big, modern cities and the traditional sector in the countryside. The diversion of limited capital funds in the form of cheap loans to the big economic units in the manufacturing sector means a greater scarcity of capital and higher rates of interest for the peasants and small economic units in the traditional sector. The diversion of limited government funds to the construction of large-scale public utilities catering to the needs of the urban manufacturing sector means that the rural districts are deprived even of the elementary public services. On top of all this, the wage level in the organized labour market in the manufacturing sector is very much higher than that in the unorganized labour market of the traditional sector. This is for a variety of reasons, ranging from trade-union pressures to the organizational need to retain an experienced labour force on a regular basis. This wage gap attracts an increasing flood of migrant labour from the countryside to towns. But, given the capital-intensive technology of the modern manufacturing sector, its employment-creating capacity is very limited. There is thus a dangerous process that transforms the disguised unemployment of the agricultural sector into the open unemployment of the shantytowns. The growing economic disparity created by domestic industrialization policies that discriminate in favour of big economic units in the urban manufacturing sector and against the peasants and small-scale economic units in the rest of the country is a powerful factor adding to the problem of subjective discontent and tensions in the underdeveloped countries.

One way out of these difficulties is to orient the manufacturing industries to the export market rather than to the domestic market. Some of the smaller Asian countries—notably Korea, Taiwan, Hong Kong, and Singapore—have successfully taken this escape route by exporting a variety of labour-intensive manufactured goods. They have been helped in this by an abundant supply of labour that is industrious, disciplined, and adaptable. Abundance of suitable raw materials may also provide a basis for the export of semimanufactured and processed products. The expansion of manufactured goods can be encouraged by the use of subsidies instead of tariff protection. But fundamentally it requires the drastic weeding out of inefficient manufacturing industries, retaining only those that have genuine prospects of becoming competitive in the world market.

The postwar record

LESSONS OF POSTWAR EXPERIENCE

The economic performance of the underdeveloped countries during the postwar period may now be examined. Broadly speaking, the GDP of the underdeveloped countries as a group grew at an average rate of 5 percent per annum during most of the postwar period. Out of a group of 54 countries, approximately half exceeded the 5 percent growth rate, and approximately half fell short of it during the period 1955–65. Looking more closely at the individual differences in economic performance during the period 1960–66, there were 18 high-growth countries (growth rate of 6 percent or higher), 21 medium-growth countries (growth rates between 4 and 6 percent), and 15 low-growth countries (growth rates less than 4 percent). These differences in economic performance are to some extent due to fortuitous causes, such as differences in the endowment of resources, which are not related to the development effort. But the figures also indicate several important lessons in development policy.

Agriculture. The first of these is the close connection between the rate of growth in the output of the agricultural sector and the general rate of economic development. The high rates of economic growth are associated with rapid expansion of agricultural output and low rates of economic growth with the slow growth of agriculture. This is to be expected, since agriculture forms a large part of the total domestic product and of the exports of the underdeveloped countries. What is more interesting is that the expansion of agricultural output has by no means been confined to those countries with an abundant supply of unused land to be brought under cultivation. Taiwan and South Korea, with some of the highest population densities in the world, have been able to expand their agricultural output rapidly by a vigorous pursuit of appropriate policies. These include the provision of adequate irrigation facilities, enabling a succession of crops to be grown on the same piece of land throughout the year; the use of high-yielding seeds and fertilizers, which raise the yields per acre in a dramatic fashion; and improvements in credit and marketing facilities and a general improvement in the economic organization of the agricultural sector.

Agricultural and capital formation

As has been shown, the importance of agricultural expansion for economic development derives from the vital role that the increase of the subsistence fund plays in the capital formation of the underdeveloped countries. This is particularly important in relation to social-overhead capital. It frequently develops that large-scale social-overhead projects such as multipurpose hydroelectric dams and superhighways are underutilized because of the lack of smaller projects, such as irrigation ditches or feeder roads. It is in the construction of this network of small-scale projects filling in the gaps in social-overhead capital that the increase in agricultural output as a source of capital formation is likely to play its most important role. Furthermore, agricultural development, by raising the incomes of the mass of the people in the countryside, has two important functions. First, it increases the size of the domestic market for the manufacturing sector, thus contributing significantly to balanced growth. Second, it reduces internal economic disparities between the urban centres and the rural districts.

Exports. A second conclusion to be drawn is the close connection between export expansion and economic development. The high-growth countries have been characterized by rapid expansion in exports. Here again it is important to note that export expansion has not been confined to those countries fortunate in their natural resources, such as the oil-exporting countries. Some of the underdeveloped countries have been able to expand their exports in spite of limitations in natural resources by initiating economic policies that shift resources from inefficient domestic manufacturing industries to export production. Nor has export expansion from the underdeveloped countries been confined to primary products. There has been a considerable expansion of exports of labour-intensive manufactured goods and semiprocessed goods.

Effects of trade barriers

It is in this context that the removal of trade barriers that developed countries have erected against underdeveloped countries can be seen as a more important way of helping these countries than simply increasing economic aid. Trade barriers are many. They include restrictions on temperate-zone agricultural products and sugar; restrictions on the simpler labour-intensive manufactured goods (which can be produced more cheaply by the underdeveloped countries); and tariff escalation, or higher rates of duties on processed products as compared with raw materials, which discourages the growth of processing industries in the underdeveloped countries. The removal of these trade barriers can help those underdeveloped countries that have already shown their capacity to take advantage of the available external economic opportunities, whereas a considerable part of aid flows into those underdeveloped countries that have not shown their capacity to absorb capital.

The conventional explanation of the connection between export expansion and economic development is that the foreign-exchange receipts from exports enable the underdeveloped countries to purchase their requirements of capital goods from abroad. This may be accepted without accepting the assumption of a fixed relationship between the foreign-exchange requirements and output. But in order to analyze the longer run impact of export expansion on an underdeveloped country, it is not sufficient to concentrate only on foreign-exchange receipts. Some of the countries that have obtained "windfall" receipts through the discovery of oil and other mineral products may or may not achieve a sustained rate of economic growth, depending on how productively they can invest their resources. From a longer run point of view, the increasing capacity of an underdeveloped country to adapt its resources and internal economic organization to the pressures of world-market demand and international competition may thus turn out to be a more important connecting link between export expansion and economic development. It is important in this connection to stress the educative effect of freer international trade in keeping an underdeveloped country open to the new ideas, new wants, and new techniques of production and methods of organization from abroad. This needs to be stressed all the more when the existing manufacturing sector of the underdeveloped countries—fostered by heavy protection—has failed in its function as a leading sector in transmitting a dynamic stimulus to the traditional sector of the economy.

Population. The third lesson to be learned is the need to control the rapid population growth of the underdeveloped countries. The average rate of population growth of the underdeveloped countries is 2.5 percent per year, but there are some countries where population growth is 3 percent or more. If the aim of economic development is to raise the level of per capita incomes, it is obvious that this can be achieved both by increasing the rate of growth of total output and by reducing the rate of growth of population. Development economists of the 1950s tended to neglect population-control policies. They were partly seduced by theories of dramatically raising total output through crash investment programs and partly by the belief that population growth could be controlled slowly, through gradual changes in social attitudes and values. But it is now increasingly recognized that changes in social values and attitudes are equally necessary in development programs aimed at increasing total output. More importantly, there have been great technical advances in methods of birth control that lend themselves to mass dissemination in the underdeveloped countries at very low cost. Furthermore, it must not be imagined that the benefits of birth-control policies are to be reaped only in the long run. As soon as some births are prevented, the relative increase in population in the working-age groups and the decline in the number of children immediately starts to release resources for increasing consumption and saving. Taking realistic rates of return on the investment of capital in the underdeveloped countries, a given amount of resources devoted to retarding population growth would be capable of contributing many times more to a rise in per capita incomes than the same amount of resources devoted to increasing total output.

Immediate advantages of population control

Growth potential. The fourth and the most general lesson to be learned is the importance of improving the absorptive capacity of the underdeveloped countries, not only for capital but also for all their opportunities for economic growth, including international trade, foreign investment, and technical innovation. This is their basic problem if one accepts the hypothesis that they are underdeveloped because they have not as yet succeeded in making effective use of their potential economic opportunities. The central element in this problem is one of the efficient allocation of the available resources. But efficient allocation of resources is not merely a matter of the formal optimum conditions of economic theory. It requires the building up of an effective institutional and organizational framework to carry out the allocation of resources. In the private sector this requires the development of a well-articulated market system that embraces the markets for final products and the markets for the factors of production. The vent for surplus theory explains how the exchange economy spreads by bringing the underutilized land and labour of the subsistence economy into export production. But at a later stage there is a need to concentrate not only on the spread but also on the intensification of the exchange economy through the growth of more specialized markets. In the public sector the development of the organizational framework requires improvements in the administrative machinery of the government, especially in its fiscal machinery.

Long-range planning of resource allocation

In the setting of the underdeveloped countries, one is concerned not only with the once for all problem of efficient allocation of resources but also with improving the capacity of these countries to make a more effective use of their resources over a period of time. That is to say, one is concerned not only with the static problem of the efficient allocation of given resources with the given organizational framework but also with dynamic problems of improving the capability of this framework. From this point of view, there is no conflict, as some people would like to maintain, between the static, or the short-run, considerations and the dynamic, or longer run, considerations. The two sets of requirements move in the same direction.

The problem of the efficient allocation of investable funds in the underdeveloped countries may be taken as an example. Static rules would require the underdeveloped countries to have higher rates of interest to reflect their greater capital scarcity. But many underdeveloped countries, under the influence of dynamic theories of economic development, have used a variety of direct and indirect controls to divert large sums of capital to the manufacturing sector in the form of loans at interest rates well below the level required to equate the demand and supply of capital funds. This has resulted not only in a wasteful use of scarce capital resources but also in a retardation of the development of a domestic capital market. Instead of developing a unified capital market for the whole country, this aggravates the financial dualism characterized by low rates of interest in the modern sector and high rates in the traditional sector. The policy

of keeping the official rate of interest below the equilibrium rate of interest also results in an excess demand for loans, leading to domestic inflation and pressure on the balance of payments and to a discouragement of the growth of domestic savings. Few private individuals are prepared to buy government securities when they frequently carry rates of interest below the rate of depreciation in the value of money. Through the pursuit of "cheap money" policies that contradict the real facts of capital scarcity, the governments of underdeveloped countries have failed to make use of the opportunity of building up a domestic capital market based on an expanding volume of transactions in government securities.

DEVELOPMENT IN A BROADER PERSPECTIVE

Modern economic development started in Great Britain, which in the 1780s accounted for a little over 1 percent of the total world population at that time. Since then, economic development has spread in widening circles to other parts of the world, spurred on by a series of technological innovations, particularly in the form of improvements in transport and communications. In the early decades of the 19th century, the circle of the developed countries was limited to western Europe. By the late 19th century, the circle had widened to include North America, Australia and New Zealand, and Japan. By the early 1970s about 34 percent of the total world population belonged to the developed countries, which among them have 87.5 percent of the total world gross national product. What are the prospects of the present-day underdeveloped countries of Asia, Latin America, and Africa joining this circle of economic development?

On the negative side there are a number of factors that have added to their difficulties. First, the level of per capita product in the present-day underdeveloped countries is much lower than in the developed countries in their preindustrialization phase (with the exception of Japan). Second, the present-day underdeveloped countries have large population bases and are handicapped by much faster rates of population growth. Third, they have generally a much weaker social and political framework to cope with the more explosive forces of discontent engendered by their reaction against their colonial past and by their present internal economic disparities.

Factors favouring economic development

On the positive side, the present-day underdeveloped countries can draw upon a greater store of scientific and technical knowledge from the advanced countries. The potential opportunities to exploit the "technological gap" are not confined to manufacturing. Modern science and technology can make immense contributions to agriculture, as illustrated by the Green Revolution created by the introduction of improved seeds and fertilizers in some Asian and Latin American countries. Modern methods of birth control can make a decisive contribution in the race for raising per capita incomes. In addition, as the circle of the developed countries widens, they are bound to exert an increasing upward pull on the underdeveloped countries.

The economic growth of the advanced countries has generally resulted in an expanding demand for the products and sometimes for direct labour services from the underdeveloped countries. But there are also the stronger localized pulls, such as the pull of the United States economy on Mexico and the pull of western Europe on the underdeveloped countries of southern Europe. The spectacular economic growth of Japan since World War II may also exert a similar pull on neighbouring countries in East Asia.

In conclusion, views as to whether underdeveloped countries are helped or hindered by being latecomers depend to a great extent on how the basic goal of economic development is defined. If the goal of development is to improve the material standards of living of the mass of the people in the underdeveloped countries, then there would seem to be some grounds for qualified optimism. Provided the underdeveloped countries are prepared to pursue appropriate economic policies, including population policies, there do not appear to be insuperable difficulties in achieving the goal of economic development in a narrower sense. But, if one views the goal of economic development in a wider sense of closing the widening gap in income levels, the prospects are less bright. The arithmetic of growth, operating on the initial differences in income, is bound to make it exceedingly difficult (though not impossible) for a presently underdeveloped country to make "a big leap" to overtake the leaders in economic development. But, as has been suggested, this second interpretation of the meaning of economic development virtually turns it into a problem in international power politics. Development economics has still a long way to go before it can begin to approach this problem.

BIBLIOGRAPHY. The statistics of national incomes and rates of growth are taken from the PEARSON COMMISSION REPORT, *Partners in Progress* (1969); and G.M. MEIER, *Leading Issues in Economic Development* (1970). The discussion of conceptual problems of national income measurement is based on S. KUZNETS, *Modern Economic Growth: Rate, Structure and Spread* (1966) and on H. MYINT, *The Economics of the Developing Countries*, 3rd ed. (1967). For a good exposition of growth economics, see R.M. SOLOW, *Growth Theory* (1970); and JOHN HICKS, *Capital and Growth* (1965). See also H. MYINT, "Economic Theory and Development Policy," *Economica*, n.s., 34:117–130 (1967), reprinted in his *Economic Theory and the Underdeveloped Countries* (1971). See R. PREBISCH, *Towards a New Trade Policy for Development* (1964), for an influential exposition of the role of foreign exchange; and MEIER, *op cit.*, for a documentation of the close relationship between economic growth and the expansion of agricultural and export production. Economic development in the broader perspective is discussed in the last three chapters of KUZNETS, *op. cit.*

(H.My.)

Economic Forecasting

To forecast is to foresee, to prophesy, to predict. Economic forecasting, therefore, is the prediction of any of the elements of economic activity. Such forecasts may be made in great detail or may be very general. In any case, they describe the expected future behaviour of all or part of the economy.

Obviously, everyone forecasts all the time; the decision to accept a new job, for example, involves the assumption (in effect, a forecast) that the employer will remain solvent and able to pay his employees. The decision to buy a new car or house involves an assumption (on the part of both buyer and seller) that future payments can and will be made according to the terms of the contract. The distinction between everyday forecasting and economic forecasting is simply that an economic forecast is more formal and explicit and usually involves the application of organized theory and carefully developed statistical techniques.

The importance of theory

Formal economic forecasting is usually based on a specific theory as to how the economy works. Some theories are complicated, and their application requires an elaborate tracing of cause and effect. Others are relatively simple, ascribing most developments in the economy to one or two basic factors. Many economists, for example, believe that changes in the supply of money determine the rate of growth of general business activity. Others assign a central role to investment in new facilities—housing, industrial plants, highways, and so forth. In the United States, where consumers account for such a large share of economic activity, some economists believe that consumer decisions to invest or save provide the principal clues to the future course of the entire economy. Obviously the theory that a forecaster applies is of critical importance to the forecasting process; it dictates his line of investigation, the statistics he will regard as most important, and many of the techniques he will apply.

Judgment in forecasting

Although economic theory may determine the general outline of a forecast, judgment also often plays an important role. A forecaster may decide that the circumstances of the moment are unique, and that a forecast produced by the usual statistical methods should be modified to take account of special current circumstances. This is particularly necessary when some event outside the usual run of economic activity inevitably has an economic effect. For example, forecasts of 1970 economic

activity in the United States were more accurate when the analyst correctly foresaw that Congress would reduce taxes, that gradual troop withdrawals from Vietnam would continue, and that the tight money policy of the Federal Reserve banks would be eased in early 1970. None of these conclusions followed from purely economic analysis; they all required judgment as to future decisions. Similarly, an economist may decide to adjust an economic forecast that was made by traditional methods to take account of other unique conditions; he may, for example, decide that consumers will alter their spending patterns because of special circumstances such as fear of future inflation or of threatened shortages.

Although judgment may be based on experience and understanding, it may also be no more than unconscious bias. Forecasts based on judgment cannot be subjected to the kind of rigorous checks applied to forecasts developed by the use of more objective techniques. Consequently, the most accurate and useful forecasts are likely to be those founded on essentially economic considerations and standard statistical techniques. Though they can then be modified by the application of judgment, the resulting changes should be stated explicitly enough so that anyone wishing to use a forecast will know where, and how, it has been affected by the forecaster's own judgment, or bias.

Economic forecasting is probably as old as organized economic activity, but modern forecasting got its impetus from the Great Depression of the 1930s. The effort to understand and correct the worldwide economic disaster led to the development of a vastly greater supply of statistics and also of the techniques needed to analyze them. After World War II, many governments committed themselves to maintaining a high level of employment. Most governments of the industrialized Western countries were prepared to intervene more often and more directly in economic affairs than previously. Business organizations manifested more concern with anticipating the future. Many trade associations now provide forecasts of future trends for their members, and a number of highly successful consulting firms have been formed to provide additional forecasting help for governments and businesses.

TYPES OF FORECASTING

Forecasting the GNP and its elements. Perhaps the forecasts most familiar to the public are those of gross national product and its elements. Gross national product, or GNP (see NATIONAL INCOME ACCOUNTING), is the total value of the goods and services produced in a nation. It is, therefore, a convenient and comprehensive measure for assessing changes in general economic welfare. A forecast of the GNP also provides a useful framework for more detailed forecasts of specific industries. Almost all developed nations maintain sets of national income accounts and make forecasts as well.

The gross national product can be regarded as being composed of three major components: spending by government, private investment spending, and spending by consumers. Net exports (that is, exports minus imports) are also counted in the GNP but their magnitude is usually small. (For the nations that depend more heavily on foreign trade, net exports are of course more important.)

Forecasting government spending

Government spending is usually the easiest part of the GNP to forecast. Government expenditures can be determined with a fair degree of accuracy for well over a year in advance by studying existing budgets and appropriations, modified to take account of new political or economic developments. Most such adjustments are relatively minor for any forecast that runs only a year or two into the future; new government programs usually have only a small effect on expenditures in the short run. An obvious exception to this is a major change in the military situation, which can drastically alter spending plans.

It is important to note that government spending, as counted in the GNP, is not the same as total budgeted expenditures. Spending gets into the GNP only when money is paid for goods—military equipment, buildings, and so on—or services, which principally means the wages and salaries of government employees. These kinds of expenditures account for only part of the government budget; the remainder represents money transferred to bondholders, other private citizens (particular people receiving pensions), and state and local governments. These funds affect the GNP only when they are finally spent by the recipients.

Forecasting private investment

Private investment poses far more difficult forecasting problems because it reflects many thousands of individual and corporate decisions that are not recorded publicly (as government budgets are) and that can be, and often are, changed very substantially. Private investment is the most erratic of the major categories of the GNP—the most subject to "boom and bust" cycles. A good forecast of investment spending is therefore essential to an accurate appraisal of the overall economic situation.

Capital investment by business (spending for new plants and equipment) is particularly important. The incomes generated in the process of manufacturing new equipment and building new plants play a major role in increasing consumer spending during periods of expansion. But when investment slumps, employment and incomes generally also suffer, slowing the entire economy. Business investment has thus been studied with great care, and a wide variety of methods to guide forecasters has been developed, including econometric models, surveys of business investment plans, regular reports on commitments for investment, and fundamental studies of the condition of the nation's stock of capital goods (see below *Forecasting Techniques*).

Business also invests in inventory—that is, goods in the process of production and finished goods not yet sold to the final consumer. Most of the time, inventories increase roughly in line with the trend of sales. If sales fall short of expectations, however, inventories tend to become excessive. Business then moves to reduce stocks by cutting back production. Such cutbacks can aggravate economic recessions; as production is reduced because of disappointing sales, incomes are thereby reduced and sales fall further, inventories must be cut even more, and so on in a downward spiral. Consequently, correct forecasts of inventory investment are both essential to good economic forecasting and also particularly difficult. Surveys of business plans to build or reduce inventory have been helpful; econometric methods have also been applied; but inventory investment remains one of the weak links in the forecasting process.

New home construction accounts for a relatively small share of the GNP, but it is important to the forecaster because home construction is a relatively volatile industry. Homebuilding activity responds quickly to changes in the availability of mortgage money from the principal mortgage lending institutions, and thus forecasters follow closely the flow of savings to these institutions and also the level of commitments that have been made to finance future construction. Information on the number of building permits issued is also helpful, as are statistics on the volume of new construction contracts.

Forecasting consumer spending

Economists used to believe that forecasting consumer spending was fairly simple; as a rule of thumb, consumers could be counted on to spend 94–95 percent of their current income and save the rest. Thus an analyst could calculate the amount of personal income generated by government spending, private investment, and past consumer spending, adjust for tax payments, and arrive at a good estimate of consumer spending. This method still works well in determining the average rate of spending over an extended period of time. But in rich countries consumers as a group are quite free to vary their spending patterns in the short run; they may at any particular time spend more than usual because they anticipate shortages or because they believe that their incomes will rise further; or they may cut back their spending if they fear that a recession is about to develop. Such variations from normal spending patterns have their main effect on durable goods such as automobiles and household appliances; spending is far more stable for nondurables (food, clothing, and the like) and for services.

Because consumers account for such a large proportion of all economic activity, a shift of just one or two percent between spending and saving can make the difference between rapid growth or recession for the entire economy. Economists now use surveys of consumer attitudes in attempting to read the mood of the public; surveys of intentions to buy durable goods such as automobiles and appliances have also been helpful.

Forecasting for an industry or firm. General economic conditions set the tone for all parts of the economy. Good forecasting for an industry or firm begins, therefore, with a good analysis of the overall economy. Within this framework, the analyst must then take account of the particular factors that are most important to his own industry. In some cases, the sales of an industry may correlate fairly directly with one or more of the elements of the national income and product accounts—lumber sales with home construction, for example, or sales of nondurable consumer goods with consumer income and total consumer spending. Forecasting for industries that produce basic materials usually requires a series of projections for specific markets. A steel forecast might be based on the outlook for such major steel markets as automobiles, construction, and metal containers. The basic forecast would then be adjusted for expected shifts in exports and imports of steel and for changes in inventories of steel or steel-using products. Such inventory changes are usually the result of actual or threatened strikes.

Forecasting is most difficult for companies that produce durable goods such as automobiles, industrial equipment, and appliances and for companies that supply the basic materials for these industries. This is because sales of such goods are subject to extreme variation. Annual sales of automobiles in the United States have increased by as much as 38 percent in one year and declined by as much as 26 percent in another. Consequently, the durable goods industries in general and automobile companies in particular have developed especially complex and sophisticated forecasting techniques. In addition to careful analysis of income trends (based on a general economic forecast) automobile companies support a number of studies of consumer attitudes and surveys of intentions to purchase automobiles.

Forecasting for an individual firm obviously begins with a forecast for the industry or industries in which it is involved. Beyond this, the analyst must determine the degree to which the company's share of each market may vary during the forecast period. Such variations can result from the introduction of a new product, the improvement of an existing product, the opening, closing, or expansion of plants, the activities of competitors, a change in sales effort, or a variety of other factors. Information required to make such assessments may come in part from the company's own investment and marketing plans. Information on the activity and sales prospects of competitors is frequently collected from the firm's own salesmen. An increasing number of companies now employ sophisticated market research techniques to determine the probable reaction of their customers to new products.

Long-term forecasting. In recent years, increasing effort has been devoted to long-range forecasting for periods extending five, ten, or more years past the normal "short-term" forecast period of one or two years. Business has come to recognize the usefulness of such forecasts in developing plans for future expansion and financing.

Labour force projections

Long-range forecasts usually are based on the assumption that activity toward the end of the period will reflect normal "full" employment. Given this assumption, the overall rate of growth depends on two principal factors: the number of people in the labour force and the rate at which productivity (output per worker) increases. The number of people of working age is known, barring some natural disaster, far into the future; they have already been born. Forecasters usually assume that productivity will continue to grow at the typical rates of recent decades. Expected technological developments, however, may alter the projected rate of change. The combination of changes in the labour force and productivity produces an estimate of the total growth rate for the economy.

A measure of total economic activity arrived at by such methods as these serves, in effect, as a control total for making long-range forecasts of the constituent elements of the economy. If estimates for spending by consumers, government, and business add up to more than the total of goods and services that can reasonably be expected, then the projection for one or more of these elements must be reduced. If the sum of the projected parts is less than the probable total, the analyst is likely to assume a shift in economic policy that will move the economy up to full employment by the end of the forecast period and adjust his various projections up to the appropriate total.

Long-range forecasts for individual parts of the economy depend on many of the same factors as do short-range forecasts, except that cyclical factors are usually ignored. Over the longer range, however, additional factors enter. Among the most obvious of these are growth in the population and shifts in its age composition. Changes in age composition—particularly the greater percentage of young people—have had a major effect on both consumer and government spending patterns in many countries since World War II. Young adults now constitute a huge market. They tend to buy large amounts of durable goods and also add to the need for home construction; on the average, they save less and borrow more in relation to their incomes than most older people do. This has major significance for financial institutions and financial markets.

In addition to population pressures, a number of other trends and assumptions influence long-range forecasts. Assumptions about war and peace are obviously critical. Assumptions must be made about government spending programs; expensive new programs may bring higher taxes and less consumer spending, whereas slower growth in government spending may lead to tax reductions. Over longer periods of time, technological discoveries or changes in financial institutions can affect the overall economy. When the forecast is made for an industry or a firm, the expected introduction of new products is also important.

FORECASTING TECHNIQUES

Economic forecasters have a vast array of information to work with and a growing variety of techniques. A few economists, believing that just one or two key factors determine the future course of the economy, limit their observations to these factors and develop forecasts based on them. A leading example of this is found in the school of thought that ascribes most importance to changes in the money supply. But most economists use a wider range of material.

Information on spending. Some elements of the future are known with reasonable accuracy. Government spending is reflected in existing budgets. These budgets indicate how much will be spent and how much money will be extracted from the stream of private spending by taxation. Similar information is available on some parts of the private economy. Periodic surveys conducted both by government and by private organizations measure business plans to invest in new plants and equipment. Increasingly, attempts are made to probe the mood and intentions of consumers concerning the possible purchase of automobiles, houses, appliances, and other durable goods. Regular surveys are also made to determine the general mood of the public—whether people are optimistic or pessimistic about their own economic future and thus whether their spending is apt to be relatively strong or relatively weak. In general, such information obtained from the various surveys of investment plans, spending plans, and attitudes has been highly useful to economic forecasters. Such information helps to limit the range of possibility. But plans and attitudes change, sometimes quite abruptly, and although the surveys are useful tools they are not clear and reliable guides to the future.

Surveys of business and consumer intentions

Selection of turning points. Probably the single most difficult economic forecasting problem is to pick the turning points in economic activity—the times at which the

economy turns from growth to recession or from recession to recovery. Because of the difficulty and importance of the problem, major efforts have been made to develop tools for this purpose. The National Bureau of Economic Research in the United States has identified a number of statistical series that normally turn up or down before the economy does. Common stock prices, business inventories, and changes in consumer installment debt are among these series, which are known as "leading indicators." Other statistical series normally move in line with overall activity ("coincident indicators"), and a third group changes direction after the economy does ("lagging indicators"). Although careful study of these groupings of statistics can be helpful, unfortunately there is no fixed relation between the movements of the economy and those of the various indicators. Although the "leading indicators" do ordinarily lead, the length of the lead varies. This reflects the dynamic nature of a complex economy that is constantly changing and in which strength or weakness may come from a variety of sources.

Some economists also use sets of statistics called diffusion indexes to calculate economic turning points. A diffusion index is a method of summarizing the common tendency of a group of statistical series. If a greater number of the series are rising than are declining, the index will be above 50; if fewer are rising than declining, it will be below 50. In effect, a diffusion index measures the degree to which either strength or weakness pervades the economy. If, for example, most of a group of industries are increasing their production rates, the economy as a whole is probably expanding; if the proportion of industries that are growing begins to decline and falls significantly below 50 percent for a period of time, the economy is probably in a recession, or at least moving in that direction.

Econometric methods

Economists frequently use mathematical equations to express the normal relations between various economic factors. As a simple example, a given increase in consumer income will ordinarily produce a certain increase in sales, saving, and tax revenue, and these developments can be expressed mathematically. With a sufficient number of equations, all the important interactions within the economy can be simulated in a mathematical model. With the advent of computers able to make millions of calculations in a few moments, economists have begun to construct more and more complex sets of equations, called econometric models (see ECONOMETRICS). These models, some of which include hundreds of equations, can be used to forecast overall economic activity (macroeconomic forecasting) or developments in particular parts of the economy (microeconomic forecasting). The success of econometric forecasting has so far been limited because the exact nature of economic relations is not fully known, and also because of the inadequacies of existing statistics. Nevertheless, the improvement of these techniques represents the greatest hope for more accurate economic forecasting in the future.

Input-output analysis

The computer has also stimulated development of another potential forecasting tool, input-output analysis. Input-output tables show the relations between the various industries and sectors of an economy. They show, for each industry, the amount of its output that goes to every other industry to be used as raw materials or semifinished products, as well as the amount that goes to the final markets of the economy. Input-output tables also show each industry's consumption of the products of other industries, as well as each industry's contribution to the production process. With such a table it is possible to trace the effects of changes in one industry or sector upon all the other industries and sectors.

The usefulness of input-output analysis for forecasting purposes has been limited by a number of factors. One problem is that it takes years to put such complex sets of statistics together; in a changing economy, relations may have shifted by the time the data for a base period have been assembled. Recently, however, progress has been made in developing methods to bring these relations (called technical coefficients) up to date, and input-output analysis shows increasing promise as a forecasting tool.

THE ACCURACY OF ECONOMIC FORECASTS

Major improvements have been made in the accuracy of economic forecasting. A competent economist can usually predict accurately enough to provide guidance to those who make policy decisions. Consensus forecasts—the average of a group of forecasts made by different individuals or organizations—have come closer and closer to the mark in recent years. Errors persist, nevertheless, and they occasionally lead to bad decisions.

Sources of error

The sources of error in economic forecasts are many. Some lie outside the realm of economic analysis; wars, agricultural disasters, or political upheavals are examples. Some forecasts go wrong for essentially ideological reasons: people who do not believe that an economic system will function tend to forecast its failure, which accounts for the many predictions of another great depression. Adherents of a political party in power have a notable tendency to optimism, whereas their opponents, including economists, tend to view the future with alarm. The student of forecasts must obviously consider their source; purity of motive is an important virtue for economists.

The most vexing sources of error, however, lie within the realm of economic knowledge. Many are statistical. Not only are some of the published data inaccurate but even the best statistics are available only after a period of time; the forecaster is forever predicting the future when he cannot be completely sure of the present. Statistics of inventories, among the most volatile economic elements, are noteworthy in this respect: for example, the published estimates of inventory growth in the United States for the last quarter of 1966 started with an annual rate of $14,400,000,000, were revised upward to $16,400,000,000, revised again to $18,500,000,000, and ended finally at $19,900,000,000. Forecasts based on the original figure could have gone seriously wrong.

The most persistent form of error in economic forecasting, however, is probably theoretical. Man's knowledge of his own economic institutions is limited. Good analysis is made more difficult by the fact that these institutions are constantly changing. This means that economic theory based on experiences of the 1950s may be of limited use in the 1970s. Some of the greatest contributions to the continued improvement of economic forecasting may come from economists who are not necessarily forecasters themselves but have the insight to understand the changing economy of today.

BIBLIOGRAPHY. WILLIAM F. BUTLER and ROBERT A. KAVESH (eds.), *How Business Economists Forecast* (1966), a collection of essays by specialists in various fields of forecasting; LEONARD S. SILK and M. LOUISE CURLEY, *Forecasting Business Trends* (1956, reprinted 1963), a general description of forecasting techniques; U.S. DEPARTMENT OF COMMERCE, OFFICE OF BUSINESS ECONOMICS, *Survey of Current Business* (monthly), one of the most useful sources of statistical data for U.S. forecasters.

(R.W.E.)

Economic Growth

The idea of economic growth is a modern one. Before the Industrial Revolution that began in the 18th century, most men had no notion that a whole country might grow richer within the span of a single lifetime. By the middle of the 20th century, the hope and expectation of this kind of progress had become important to public opinion in many countries. Much of this concern could be traced to political causes: the desire of statesmen for power and the desire to prove the superiority of one social system over another. Through rapid growth, it is argued, the economic base vital to defense can be kept large enough to deter any enemy. Another reason for the high importance given to growth is that it is easier to redistribute a rapidly growing national income than a slowly growing one, since in the former case nearly everyone may be made better off in some material way. Rapid growth also enables governments to spend more on public services.

Growth defined

Economic growth may be defined and measured in various ways. Most economists start with some indicator of overall economic activity, such as Gross National Prod-

uct (GNP), which measures a country's total production of goods and services. More often those concerned with measuring economic growth are interested in changes in the level of welfare of individuals; thus the rate of growth of consumption per capita is often cited as an index of the growth of welfare within a society. This measure can be objected to as misleading because it does not take account of expenditures by government on health, education, and other amenities, which in many countries are large and growing and should be considered in measuring human welfare. On the other hand, two countries may show the same rate of growth in per capita consumption; but the rate of growth of investment—*i.e.*, the growth of additions to the capital stock and therefore to productive capacity—may be higher in one than in the other. Thus the rate of growth of GNP per capita may be a better measure of an economy's long-run prospects. This indicator is closely related to another measure, the rate of growth of labour productivity, which is simply the rate of growth of output per unit of labour employed.

All these measures suffer from various shortcomings. Human welfare, for example, is a function of many things besides the consumption of goods and services. Estimates of consumption or GNP cannot take account of the importance of leisure or measure the quality of life. Another difficulty with statistical measurements of growth is the paucity of data for most countries before World War II.

HOW ECONOMIES GROW

Growth can best be described as a process of transformation. Whether one examines an economy that is already modern and industrialized or an economy at an earlier stage of development, one finds that the process of growth is uneven and unbalanced. Economic historians have attempted to develop a theory of stages through which each economy must pass as it grows. Early writers, given to metaphor, often stressed the resemblance between the evolutionary character of economic development and human life—*e.g.*, growth, maturity, and decadence. Later writers, such as the Australian economist Colin Clark, have stressed the dominance of different sectors of an economy at different stages of its development and modernization. For Clark, development is a process of successive domination by primary (agriculture), secondary (manufacturing), and tertiary (trade and service) production. For the American economist W.W. Rostow, growth proceeds from a traditional society to a transitional one (in which the foundations for growth are developed), to the "take-off" society (in which development accelerates), to the mature society. Various theories have been advanced to explain the movement from one stage to the next. Entrepreneurship and investment are the two factors most often singled out as critical.

Stages of growth

Economic growth is usually distinguished from economic development, the latter term being restricted to economies that are close to the subsistence level. The term economic growth is applied to economies already experiencing rising per capita incomes. In Rostow's phraseology economic growth begins somewhere between the stage of take-off and the stage of maturity; or in Clark's terms, between the stage dominated by primary and the stage dominated by secondary production. An aspect of the transformation of one economy, that of the United States, can be seen in Table 1, which describes changes in the percentage distribution of the civilian labour force over almost a century. The most striking fact is the enormous decrease in the proportion of the labour force employed in agriculture. Similar changes are apparent in the historical development of other industrialized economies. There are other aspects of growth. The decline in agriculture and the rise of industry and services has led to concentration of the population in cities, first in what has come to be described as the "core city" and later in the suburbs. In earlier years public utility investment (including investment in transportation) was more important than manufacturing investment, but in the course of growth this relationship was reversed. There has also been a rise in the importance of durable consumer goods in total output. In the U.S. experience, the rate of growth of capital goods production at first exceeded the rate of growth of total output, but later this too was reversed. Likewise, business construction or plant expenditures loomed large in the earlier period as an object of business investment compared to the recent era. Whether other countries will go through the same experience at similar stages in their growth remains to be seen.

Table 1: Occupational Distribution of the U.S. Labour Force
(percentages of total labour force)

year	agriculture	manufacturing	other (including services)
1870	50	17	33
1880	50	18	32
1890	42	20	38
1900	37	22	41
1910	31	22	47
1920	27	26	47
1930	22	23	55
1940	17	23	60
1950	12	26	62
1960	8	25	67
1968	5	24	71

Sources: *Historical Statistics of the United States; Economic Report of the President; Manpower Report of the President.*

Table 2 gives some comparative growth rates for a group of developed countries. It shows how uneven the process of growth can be. Partly this unevenness reflects the extraordinary events of the 1913–50 period, which included two major wars and a severe and prolonged depression. There are sizable differences, however, in the growth rates of the various countries as between the 1870–1913 period and the 1950–66 period. For the most part these differences indicate an acceleration in the rate of growth in more recent times. Many writers have attributed this to a relatively higher ratio of business investment in plant and equipment.

National comparisons

Table 2: Annual Average Growth of Output per Head of Population
(percent)

	1870–1913	1913–50	1950–66
France	1.4	0.7	3.9
Germany*	1.8	0.4	5.9
Italy	0.7	0.6	4.9
Sweden	2.3	1.6	3.3
U.K.	1.3	1.3	2.4
U.S.	2.2	1.7	2.5
Japan†	2.9	2.9	8.8

*Data for Germany after 1945 do not include the area incorporated in the German Democratic Republic. †Data for Japan for the two long periods are the average growth rate for 1880–1959.
Sources: Angus Maddison, *Economic Growth in the West;* D.C. Paige, "Economic Growth: The Last Hundred Years," *National Institute Economic Review,* July 1961; and United Nations and International Monetary Fund publications.

The relatively high rates of growth for West Germany, Japan, and Italy in 1950–66 have stimulated a good deal of discussion. It is often argued that "late starters" can grow faster because they can borrow advanced technology from the early starters. In this way they leapfrog some of the stages of development that the early starters were forced to move through. This argument is nothing more than the assertion that late starters will grow rapidly during the period when they are modernizing. As indicated by Table 2, Italy did not succeed in growing rapidly and thereby modernizing until after World War II. Together with Japan and Germany it also experienced a large

amount of war damage. This has an effect similar to starting late, since recovery from war entails building a stock of capital that will, other things being equal, embody the most advanced technology and therefore be more productive and allow faster growth. The other part of this argument is the assertion that early starters are actually deterred from introducing on a broad front the new technology they themselves have developed. For example, firms in a country that industrialized early may be inhibited from introducing a more modern and efficient means of transportation on a broad scale because there is no guarantee that other firms handling the ancillary loading and unloading tasks will also modernize to make the change profitable.

Related to this is the problem of whether or not growth rates of developed economies will eventually converge or diverge. For example, as per capita incomes of fast growers like the Italians and Japanese approach those of economies that developed earlier, such as the American and British, will the growth rates of the former slow down? Economists who answer in the affirmative stress the similarities in the changing patterns of demand as per capita income rises. This emphasis in turn implies that there is less and less chance to borrow technology from the industrial leaders as the income levels of the late starters approach those of the more affluent. Moreover, rising per capita incomes in an affluent society usually are accompanied by a shift in demand toward services. Therefore, so this argument goes, differences in growth rates between countries should eventually narrow because of the low growth in productivity in the service sector. The evidence is inconclusive. On the one hand, growth is a function of something more than the ability to borrow the latest technology; on the other hand, it is not clear that productivity must always grow at a slower rate in the service industries.

A rapidly increasing population is not clearly either an advantage or a disadvantage to economic growth. The American Simon Kuznets and other investigators have found little association between rates of population growth and rates of growth of GNP per capita. Some of the fastest growing economies have been those with stable populations. And in the United States, where the rate of growth of population has shown a downward historical trend, the rate of growth of GNP per capita has increased over the last century and a half. Another finding by Kuznets is that while GNP per capita in 1960 was substantially higher in the United States than in any European country, there was no significant difference in the per capita growth rates of all these countries over the period 1840 to 1960 as a whole. The conclusion is that the U.S. started from a higher per capita base; this may have been the result of its superior natural resources, especially its fertile agricultural land.

THE ANALYSIS OF GROWTH

Capital and labour

To explain why some countries grow more rapidly than others or why a country may grow more rapidly during one period of history than another, economists have found it convenient to think in terms of a "production function." This is a mathematical way of relating some measure of output, such as GNP, to the inputs required to produce it. For example, it is possible to relate GNP to the size of the labour force measured in man-hours, to capital stock measured in dollars, and to various other inputs that are considered important. An equation can be written that states that the rate of growth of GNP depends upon the rates of growth of the labour force, the capital stock, and other variables. A common procedure is to assume that the influence of the separate inputs is additive—*i.e.*, that the increase in the growth of output caused by increasing the rate of growth of, say, capital is independent of the rate of growth of the labour force. This is the starting point of a great deal of current empirical work that attempts to quantify the importance of different inputs.

Under certain assumptions, some reasonable and some patently false, it is possible to conclude that what labour and capital receive in the form of wages, profits, and interest is a fair measure of what they contribute to the productive process. Thus in the United States in the period following World War II the share of output going to labour was approximately 79 percent, while the share of output distributed as "profits" was 21 percent. If we assume that these proportions determine how much we should weight the rate of growth of the labour force and of capital respectively in determining their contribution to the rate of growth of output, we must conclude that the relative contribution of capital is slight. Alternatively, we may say that some given percentage increase in the rate of growth of the labour force will have a much larger influence on the rate of growth of output than the same percentage increase in the rate of growth of capital. This is a puzzling result and can be traced to the assumption that the influence of separate inputs is additive.

Quality improvements in the inputs. Much work has been done in an effort to measure the inputs in the productive process more accurately by taking account of improvements in the quality of both labour and capital over time. For example, it has been argued that the amount of a worker's time spent on his formal education is positively related to the income he receives and to his productive contribution. Measuring the number of man-hours worked from one period to the next will not give a true picture of the increase in labour input if the average amount of education received by workers is changing. Man-hour units must be converted to "efficiency" units. Thus if a labour force of 100 workers in the first year all had an eighth-grade education, while 20 years later each member had a tenth-grade education, then measured in efficiency units the labour force had grown. If the length of time spent on formal education increases over time, then the growth of the labour input will be larger if measured in efficiency units. There is, thus, an element of capital in the labour force.

Investment in human capital

Examples of investment in human capital are expenditures on health and on all types of education including on-the-job training. Expenditures of this sort increase the quality of the labour force and its ability to perform productive tasks. Many economists have argued that technological progress is really nothing but quality improvements in human beings. Some economists take an even broader view and speak of the "production of knowledge" as the clue to technological progress. The production of knowledge is a broad category including outlays on all forms of education, on basic research, and on the more applied type of research associated especially with industry. It is argued that fast-growing industries tend to be those having a high research and development component in their total costs. In addition, firms within an industry that have large research and development budgets tend to experience the most rapid technological progress. The argument is that technical change and improvements must originate in inventions that lead to innovations in the products produced or in the processes whereby existing products are manufactured.

A similar argument applies to the size of the capital stock. It can be maintained that design improvements increase the efficiency of capital goods so that a dollar's worth of machinery purchased today may be much more efficient than a dollar's worth of depreciated machinery purchased yesterday. The rate of growth of the capital stock measured so as to take account of quality improvements will be greater than the rate of growth of the capital stock measured in a way that neglects the differences between "vintages."

Some economists have stressed "economies of scale." For example, if an increase in the use of capital and labour leads to a greater than proportionate increase in output, this is said to result from economies of scale. Economies of scale may arise because an expansion of the market justifies a radical change in productive techniques. These new techniques may be so much more efficient that the returns in the way of increased output are much greater proportionately than the increase in inputs.

Another source of growth and of technical progress in

particular has been seen in shifts of demand from low productivity sectors to high productivity sectors, thus causing resources to be reallocated. The most notable movement has been the shift of resources, especially labour, out of agriculture—a traditionally low-productivity sector. Such shifts act to increase the rate of growth of output in ways that cannot be accounted for by simply measuring growth in total inputs. Historically, the allocation of both capital and labour have shifted during the growth process from low productivity sectors to high ones, causing the rate of growth of output to exceed the weighted average of the rates of growth of total inputs.

Entrepreneurship. This historical fact points to an element that has received little attention so far: the influence of entrepreneurship. If the allocation of resources changes during the course of growth and development, it does so under the leadership of an entrepreneurial class. The quality of entrepreneurship is seen by many economists as an important explanation of differences in the rate of technical progress between countries. Decisions must be made somewhere along the line as to whether a new product or process will be introduced. It has been argued that two countries undertaking similar amounts of investment leading to more or less identical rates of growth in the capital stock will not necessarily show the same rate of technical progress. In one country entrepreneurs may be undertaking enterprise investment that has as its aim the introduction of the most advanced types of production techniques, those that will lead to a rapid growth of labour productivity. In the other, because of hesitation or ignorance, the investment program may lead only to marginal changes in productive processes; the resulting growth in labour productivity and GNP will be small. For example, much has been said since World War II about the more aggressive nature of German businessmen as compared to their English counterparts. The emphasis on the role of the entrepreneur in economic growth stems from the theoretical work of the economist Joseph A. Schumpeter, but many others have echoed it.

Arguments about the significance of capital investment

The play of influences. Much thinking assumes, then, that contributions to output from growth of individual inputs are independent of one another. This assumption allows many growth theorists to conclude that capital investment is relatively unimportant as a growth factor. If there is interaction between the rates of growth of the different inputs, however, then it is possible to draw different conclusions. For example, over time there are likely to be improvements in the quality of capital goods. A machine that requires so much steel and so much labour to manufacture may be twice as productive as an older machine that required the same amount of raw materials and labour in its manufacture. Thus the rate of growth of technical progress and the rate of growth of the capital stock measured in natural units interact. Furthermore, the interaction between technical progress and capital formation is not necessarily in one direction. New knowledge opens up new production possibilities and gives rise to potential increases in technical progress and profits. Or the better educated the labour force, the more adaptable is it likely to be and therefore the better able to cope with new production techniques. At the same time, the higher the rate of growth of capital, the higher will be the growth of incomes and therefore the demand for education. The fact that much of the overall growth of technical progress stems from the transfer of resources and the positive association between the rate of transfer of resources and the rate of growth of the capital stock is another example of interdependence or complementarity between the growth of the inputs. But, again, capital investment undertaken to develop new lines of production will also be dependent on technological progress going on in those areas.

Conventional marginal productivity doctrine argues that as an input such as capital rises relative to labour, the additional output or marginal product that can be attributed to this extra amount of capital will be less than what a unit of capital on the average had been producing before. Marginal productivity doctrine also assumes that each unit of capital is identical with the next. This assumption is the basis for the argument that as more units of capital are utilized in production with a given amount of labour, it will push down the former's marginal product. There is the possibility, however, that additional units of capital may enhance the productivity of existing units: for example, an increase in the amount of capital resources devoted to the development of transportation and distribution may raise the productivity of capital employed, say, in manufacturing. The development of this kind of social overhead capital is certainly a prerequisite for a high return to capital in manufacturing, wholesaling, and retailing.

The analysis can be carried back one more step, to the basic determinants of growth. Economists ask why it is that capital, labour, or technical progress has grown more rapidly in one economy than in another or at one time than at another. Historically, the transition from a subsistence-level, underdeveloped state to a higher-level, developed one has been accompanied by a decline in the death rate followed by a decline in the birthrate. This naturally has the effect of first speeding up the rate of growth of the population and labour force and then reducing it as birthrates fall. Migration can greatly alter this picture, often in unpredictable ways. In the United States, for example, the rate of growth of the population and labour force during the 19th and early 20th centuries was higher than in most other developed countries, mainly because of high rates of immigration. From 1840 to 1930, the native-born U.S. population increased about 600 percent, while the number of those of foreign birth increased 1,300 percent.

The role of government. The differences in rates of growth are often attributed to two factors: government and entrepreneurship. The two are not mutually exclusive. In the early stages of sustained growth, government has often provided the incentives for entrepreneurship to take hold. In some economies the development of transportation, power, and other utilities has been carried out by the government. In others the government has offered financial inducements and subsidies. The land given U.S. railroad developers in the second half of the 19th century is a notable example of the latter. Another important role governments have played in the early stages is to help establish the sort of capital and money markets in which lenders could have confidence. Without financial intermediaries acting as brokers between lenders and business borrowers, it is difficult to envisage economic growth taking place on a sustained and rapid basis.

Government measures in the 19th century

In the 19th century, most liberal thinkers held that the main role for government in a developed capitalist system was that of a policeman: to preserve law and order, uphold the sanctity of private property, and give business as much freedom as possible. The depression of the 1930s persuaded many that a laissez-faire system did not automatically provide the necessary incentives to the innovation and risk bearing essential for economic growth. This led to a good deal of writing on the role that governments might play in stimulating growth. Economists have argued that, at the very least, governments can undertake to prevent serious and prolonged recessions. Only in this way can a general business psychology be developed that assumes growth to be the natural course of things, so that investment programs will pay off.

Growth theorists since World War II have gone further, arguing that it is not enough simply to achieve full employment periodically. Some maintain that it is necessary to maintain full employment over an extended period of time if high growth is to result. This argument relates to the earlier point that two economies may experience the same rate of growth of capital but that overall growth and technical progress will proceed at a much more rapid rate in one than in the other because of differences in the quality of new capital goods produced. The term enterprise investment has been used to describe the kind of capital formation that involves innovations and that by building ahead of demand generates rapid rates of growth of productivity or technical progress. But to get

such growth, it has been argued, an economy must be run "flat out," at full speed. While this has been subject to some dispute, there is a fairly general consensus that growth will be faster when unemployment fluctuates within a narrow range and at low levels.

Government intervention in expenditure

A variation on this argument is the question of how a government may intervene to determine the distribution of output between those types of expenditure that contribute to growth and those that lead to the immediate satisfaction of consumer demand. Here the choice lies between business investment, research, and education on the one hand and consumption on the other. The larger the first three, the more rapid will be the rate of growth. Governments giving a high priority to growth have various means at their disposal for influencing it. Consumption can and has been constrained through increases in income tax rates. The same is true of other tax rates such as the property tax—the chief revenue source for primary and secondary education in the United States. Tax credit for research and development expenditures is a common method for encouraging business outlays that may lead to innovations. The same method has been used to stimulate business investment outlays. "Easy money" policies on the part of the central bank, whereby the cost of borrowed funds and their availability are indirectly regulated in such a way as to encourage business borrowing, may lead to higher levels of real investment.

The true cost of stimulating growth will always be a temporary cut in current consumption. Only in the future can the economic benefits of the higher investment be realized. By the same token, current consumption can always be enlarged by a neglect of the future. It is even possible for current production to be so biassed toward the satisfaction of immediate needs that the productive capacity of an economy slowly declines as capital goods are not replaced. Between the extremes of total neglect of future generations and the paring down of current consumption to a bare subsistence minimum lie an infinite number of possibilities.

THE SOCIAL COST OF GROWTH

The belief that governments should have a large say in choosing the "right" rate of growth has also led some writers to challenge the social and economic value of economic growth in an advanced industrial society. They attribute to growth such undesirable side effects of industrialization as traffic congestion, the increasing pollution of air and water, the despoiling of the landscape, and a general decline in man's ability to enjoy the "real" amenities of life. As has been seen, growth is really a transformation whereby certain industries experience a rise in importance followed by an eventual decline as the market for their output becomes relatively saturated. Demand, relatively speaking, moves on to other types of industries and products. All of this naturally implies a reallocation of resources over time. The faster these resources move, other things being equal, the more rapidly can growth and transformation proceed. The argument can be recast in terms of this transformation. A slower rate of growth in per capita consumption will slow down the rate of transfer of resources, but it may also result in a more livable environment. The rate of growth of individual welfare may even be increased. Some argue that in a growth-oriented society wants are created faster than the industrial machine can satisfy them, so that people are more dissatisfied and insecure than they would be if growth were not given such a high value. It is held by some critics that, in modern industrial society, consumption exists for the sake of justifying production rather than production being carried out to satisfy consumer desires. These arguments are a powerful challenge to those who see growth as the most important economic goal of a modern society.

THEORIES OF GROWTH

In discussing theories of growth a distinction must be made between theories designed to explain growth (or the lack of growth) in countries that are already developed and those concerned with countries trapped in circumstances of poverty. Most of what follows will be confined to the former.

The savings–investment equation

As the British economist John Maynard Keynes pointed out in the 1930s, saving and investment are not usually done by the same persons. The desire to save does not necessarily generate investment. If savers attempt to save a larger share of their income than before (thereby consuming less) and if this is not matched by an equal increase in the desire of others to invest, total spending will decline. A natural reaction on the part of business will be to cut back on production, thereby reducing incomes earned in production. The final effect may be a cumulative movement downward as total demand becomes insufficient to employ all of the labour force. This break in the circular flow of income and expenditure suggests the possibility of a capitalist economy alternately experiencing periods of prolonged and severe unemployment (when desired savings at full employment exceed what the economy wishes to invest at full employment) and periods of serious inflation (when the inequality is reversed). This situation has not been the case historically. Since 1900 the unemployment rate (on an annual basis) has seldom remained outside the range of 3–6 percent for more than a year in the United States, if the war years and the Great Depression are omitted. While inflation has been serious at times, especially during wars or immediately after, it has not been of the prolonged, runaway type. For the most part growth has been a fairly regular process. The same can be said for other developed economies. In the following discussion, some attention will be paid to the ways in which the various theories of growth account for this important historical fact.

Modern growth theory can be said to have started with Schumpeter. Unlike most Keynesian or pre-Keynesian theorists, Schumpeter laid primary stress on the role of the entrepreneur, or businessman. It was the quality of his performance that determined whether capital would grow rapidly or slowly and whether this growth would involve innovation and change—*i.e.*, the development of new products and new productive techniques. Differences in growth rates between countries and between different periods in any one country could largely be traced to the quality of entrepreneurship. The latter in turn reflected certain historical and cultural values carried by the business class. Schumpeter also attributed much of the growth of technical progress and of the supply of labour to the entrepreneur. Thus, in more modern terminology, Schumpeter's explanation of why demand and supply have grown more or less at the same rate would be that supply adjusted to demand while demand in turn reflected the activities and investments of the entrepreneur.

Schumpeter believed that capitalism by its very success "sows the seeds of its own destruction." The American economist Alvin H. Hansen argued in the late 1930s that capitalism was in trouble in the U.S. for other reasons. According to Hansen, the closing of the geographic frontier, the decline in the rate of population growth, and the capital-saving character of recent innovations had all worked to increase the likelihood of stagnation by reducing the need for investment. The savings available in a mature economy would tend to exceed the amount that the economy would want to invest (at levels of full employment) and by increasingly larger amounts as time went on. This condition naturally would lead to increasing rates of unemployment as the discrepancy between demand and potential output widened. Hansen's views were very much coloured by the 1930s. The record since World War II has done much to overcome the pessimism generated by the great depression. As has already been observed, with the exception of the 1930s, most Western economies have not experienced prolonged and severe recessions. Demand and supply have grown more or less together in most capitalist countries, whether due to government policy or not.

The role of investment. In Keynes's *General Theory* investment played a key role in that it was the most important factor governing the level of spending in an econ-

omy, despite the fact that it typically was only one-fifth to one-sixth of total spending. This paradox can be understood in terms of a concept also developed in the 1930s, the multiplier. The multiplier was the amount by which a change in investment would be multiplied in achieving its final effect on incomes or expenditures. If, for example, investment increases by $10, the extra $10 of expenditures will generate, assuming unemployed resources, an extra $10 of production and subsequently incomes in the form of wages and profit. This increase, however, is hardly the end of the matter since most of the additional incomes earned will be respent on consumer goods. If nine-tenths of any change in income are spent on consumer goods and one-tenth is saved, consumption will increase by $9. But again, one person's expenditures are another person's income, so that incomes now rise by $9 of which $8.1 is respent on consumer goods. The process continues until expenditures, incomes, and production have increased by $100, of which $90 is consumption and $10 the original change in investment. In this case the multiplier is ten.

The negative aspects of investment

But investment may be a source of instability if it is not maintained at a rate sufficient to stimulate demand for the production it is creating. Is there any guarantee that supply or productive capacity will grow at the same rate as demand so that neither excess capacity nor excess demand results? The British economist R.F. Harrod and the American economist E.D. Domar put this question in a very simple mathematical form. In their equations, the rate of growth of supply (*i.e.*, the production function as defined above) is equal to the rate of growth of capital stock. Through investment this capital stock is augmented. The rate of growth of demand depends upon the rate of growth of investment or, more correctly, upon the rate of growth of nonconsumption expenditures. Thus investment affects both demand and supply. But the Harrod–Domar analysis still did not answer the question of what kept the system from becoming increasingly unstable.

Demand and supply. Much contemporary growth theory can be viewed as an attempt to develop a theoretical model that would bring the rate of growth of demand and the rate of growth of supply into line, since a model implying that capitalist systems are inherently unstable would not correspond to the historical facts. Models of growth may be classified according to whether they emphasize adjustments in demand or adjustments in supply, supply-determined models or demand-determined models. One of the better known examples of the supply determined model was developed by the British economist J.R. Hicks. Hicks assumed that the spending propensities of consumers and investors were such as to cause demand to grow at a rate in excess of the rate of growth of maximum output. This assumption meant that during any "boom" the economy would eventually run into a "ceiling" that, while also moving upward, was moving less rapidly than demand. The long-run rate of growth of the economy would be determined by the rate of ascent of the ceiling, which in turn would depend upon supply factors such as the rate of growth of the labour force and the rate of growth of technical progress or productivity. If for some reason these were to grow more rapidly, then output would also grow more rapidly as demand adjusted upward to the more rapid growth of supply.

An example of a demand-determined model of growth is one developed by the American economist J.S. Duesenberry. In the Duesenberry model, spending propensities of consumers and investors are such as to generate steady growth in demand. Assume that instead of spending nine-tenths of any change in income on consumer goods, as in the multiplier example above, they choose to spend 0.95. This increase will cause the rate of growth of demand to increase. The question is whether it will also cause the rate of growth of production to increase or whether it will merely result in price increases. If productivity or technical progress responds to a higher rate of growth of demand, as Duesenberry assumes, then production can grow more rapidly. Although in both the Hicks and Duesenberry models demand and supply grow at the same rate, the adjustment mechanisms are entirely different. In the Duesenberry model supply adjusts to demand; in the Hicks model demand adjusts to supply.

Other models of growth also illustrate this distinction between demand-determined and supply-determined growth. The British economist N. Kaldor assumes that there is a mechanism at work generating full employment. Simply stated, in his model an inadequate rate of investment will be offset by shifts in the distribution of income between profits and wages, which will cause consumption to change in a compensating manner so that overall demand is unchanged. While there are important differences between the Hicks and Kaldor models, both can be described as models of supply-determined growth.

Another model of supply-determined growth is that implicit in the traditional neoclassical analysis. The mechanism that adjusts demand to growing supply is the price mechanism, or Adam Smith's "invisible hand" of the market. This type of model assumes a world devoid of monopoly and uncertainty, in which the markets for capital goods and labour are free to adjust quickly so that "markets are always cleared" in the very short run.

Government intervention in demand and supply

A final example of a model of growth that illustrates the problem of adjustment between supply and demand is to be found in the work of the Dutch economist Jan Tinbergen and his followers. In contrast to neoclassical growth models where the market brings about an adjustment of demand to supply, the "target-instrument" models of Tinbergen assume that the government (as in the Netherlands and other European countries) undertakes to regulate demand and supply in an effort to achieve certain targets such as full employment or a predetermined rate of growth. For example, economists are expected to provide the fiscal authorities with a model that approximates the working of the economy and that indicates what will happen if the government, say, does not change its tax and spending programs in the coming period. These forecasts are appraised in terms of what the authorities consider desirable as a matter of social and economic policy. If it appears that unemployment will be too high and the rate of growth too low, the authorities take countermeasures. The government may, for example, cut taxes on corporate profits in order to stimulate investment. If investment is excessive and there is danger of inflation, the government may take other measures to reduce aggregate demand such as cutting its expenditures. This type of planning procedure has been tried with varying degrees of success. Sweden and the Netherlands are outstanding examples of attempts to offset fluctuations in private spending so as to realize full employment and growth. It should be noted that these models do not fit neatly into the demand-determined or supply-determined classification. In the example just given, both the rate of growth of demand and the rate of growth of supply are effectively determined by the fiscal authorities.

Foreign trade. Little has been said about foreign trade. Yet growth in most economies is very much dependent upon imports and the ability to export in order to pay for imports. The fact that some economies recovered relatively quickly from World War II and grew much more rapidly in the postwar period than others has stimulated a great deal of comparative analysis in growth theory. The exceptionally high growth rates in Japan and Germany compared to the general sluggishness of the British economy are related to foreign trade. Economists have pointed to the periodic balance of payments crises experienced by the British and the lack of such crises in Germany. During a boom, as incomes rise the demand for imports will rise also as a natural feature of prosperity. But if exports do not also rise at the same time, the authorities may be forced to take fiscal or monetary countermeasures and slow down the economy in an effort to bring imports and exports back into balance. Or exports may fail to grow sufficiently because labour costs are rising very rapidly and pushing up prices of exports faster than in competing countries.

A policy of encouraging growth has the effect of keeping the demand for imports high and making labour markets tight, thereby tending to push up money wage rates. At the same time, such a policy also tends to encourage innovations and investment projects that are very productive, particularly if the demand pressures are sustained. A "stop" policy naturally has just the opposite effects, both good and bad from the point of view of a country's balance of payments. The question is which policy will in the long run result in less rapidly rising costs and prices. Many writers have argued that if demand pressures are maintained the response or adjustment of productivity and therefore of supply to these pressures will be such that the country will soon find itself in a more competitive position. Running an economy "flat out," however, is likely to cause a short-run balance of payments crisis and lead to devaluation of currency.

Mathematical growth theories. In addition to the theories discussed above, a large body of literature has developed involving mathematical models. Because this field of analysis is so technical, only a general picture of the kinds of problems and questions discussed can be given. First, a set of equations is drawn up describing what the model builder feels are the important relations between economic variables such as output, capital, investment, and consumption. These equations must relate economic variables to one another at different points in time: for example, output last year determines consumption this year, which in turn helps to determine output this year and therefore consumption and output next year. It is possible to work out the movements of the variables over as long a period as desired. At the centre of much of this analysis is the concept of a steady-state rate of growth: one in which all the economic variables contained in the set of equations grow at the same constant rate equal to the growth of the labour force.

A related class of studies attempts to take account of the welfare of workers and consumers in the maximization of growth. These "optimal growth" models seek to maximize consumer satisfaction over time. In a model such as this the solution will not be the highest possible growth rate but one that will maximize the welfare of consumers. The importance of such models for planners would seem to depend on the realism of their assumptions as to consumer desires and technology.

The role of money in growth

Some mention of the role of money in growth is in order, although the field of monetary theory is too complex to be summarized here. Some theorists, such as the American economist Milton Friedman, advocate that in the interests of stability—and, by implication, growth—the monetary authorities should follow a simple rule in managing the nation's money supply. The supply of money has an important influence on the behaviour of an economy, it is argued, but this influence is felt only after some time. Consequently if the authorities try to control the economy through variations in the money supply, they may actually make matters worse, since the conditions they are responding to may have changed by the time their actions take effect. Friedman's prescription is to minimize the actions of the authorities; he would have the central bank confine its monetary intervention to assuring a steadily growing money supply.

At a much more abstract level, money has been introduced into mathematical growth models. These models treat money as an asset that has a certain return for one who holds it; for example, the return is positive if the price level falls and negative when the price level rises. An alternative to holding money in these models is to hold physical capital that has a return from its use in production. Changes in the money supply, through their effect on the price level, affect people's decisions to hold their wealth as money or as physical capital. The result will be changes in the rate of capital formation (investment), which in turn will affect the growth rate.

Model building and theorizing about growth has been proceeding on various levels of abstraction. Some of the work has to be considered of little practical value, in the sense that its explanatory value is negligible. Such studies, however, may stimulate other work that is helpful in an understanding of the growth process. Some models, while realistic, are not applicable to all economies. Thus a model that neglects international trade is of little use to a European economist trying to understand the more basic causes of differences in growth rates between countries. But model building will continue to be one of the more active fields in the study of economic growth.

BIBLIOGRAPHY. S. KUZNETS, *Capital in the American Economy* (1962), a descriptive study of trends in capital formation in different sectors of the American economy over the previous 100 years; C. CLARK, *The Conditions of Economic Progress* (1951), an attempt to develop a very simple theory of stages of growth that economies will pass through as they modernize; W.W. ROSTOW, *The Stages of Economic Growth* (1960), another attempt, somewhat more elaborate than that by Clark, to develop a series of stages of development and growth; E.F. DENISON and J.P. POULLIER, *Why Growth Rates Differ: Postwar Experience in Nine Western Countries* (1967), a "production function" approach to growth with no attempt to go behind the immediate causes of growth; A. LAMFALLUSY, *The United Kingdom and the Six* (1963), a study of comparative growth performances of seven European economies in the post-World War II period; E.J. MISHAN, *The Cost of Economic Growth* (1967), a critique of rapid economic growth as a policy goal; JOHN KENNETH GALBRAITH, *The Affluent Society* (1958), a critique of modern capitalism written for the intelligent layman; J.A. SCHUMPETER, *Business Cycles* (1939), a pioneering work in the field of growth and cycles that stresses the importance of entrepreneurship; J. DUESENBERRY, *Business Cycles and Economic Growth* (1958), a theoretical study of a developed economy of the American type with emphasis on the importance of demand in growth; J. TINBERGEN, *Economic Policy: Principles and Design* (1956), one of several studies by the author describing the "target-instrument" approach to policy by means of many simple mathematical models; A. MADDISON, *Economic Growth in the West* (1964), a comparative study of growth performance of Western economies, with emphasis on the importance of demand and investment in stimulating growth; C.P. KINDLEBERGER, *Europe's Postwar Growth: The Role of Labor Supply* (1967), on the importance of rapid population growth for rapid economic growth; I. SVENNILSON, *Growth and Stagnation in the European Economy* (1954), a pioneering study of growth in western Europe from World War I until the early post World War II period; F.H. HAHN and R.C.O. MATHEWS, "The Theory of Economic Growth: A Survey," *Economic Journal*, 74:779–902 (Dec. 1964), a survey of most of the types of models that fall under the heading of "mathematical growth models."

(J.L.C.)

Economic History Since 1500

This article covers the economic history of the industrialized world. It is divided into the following sections:

I. Economic development of the modern world before the Industrial Revolution
- The decline of the feudal system and the growth of commerce
- The rise of the entrepreneur and the labour market
- Capitalism and the Protestant ethic
- The voyages of discovery and the price revolution
- Nationalism and mercantilism
- Empire and the world economy
- The significance of early economic developments

II. The Industrial Revolution
- The Industrial Revolution in Britain
- Continental Europe
- Social consequences of the Industrial Revolution
- The course of industrialization, 1870–1914

III. Economic developments since World War I
- The aftermath of World War I
- The Great Depression
- The industrialization of the Soviet Union
- World War II and after in the industrialized nations
- The developing nations

I. Economic development of the modern world before the Industrial Revolution

THE DECLINE OF THE FEUDAL SYSTEM AND THE GROWTH OF COMMERCE

No sharp divide separates the economic life of the Middle Ages from that of the early modern world. But, in

retrospect, much of the 15th century in Europe displayed the symptoms of the decay of an old order. Population in many places was smaller than a century and a half earlier, and prices had generally declined. These changes were reflected in both the rural and the urban economies.

In those areas of Europe where a manorial type of agrarian economy existed, the links between lord and peasant were becoming either looser or tighter. The customary tenant, who held his land in return for services performed on the lord's demesne (estate), was in some places becoming a rent-paying tenant, even (in England, for example) turning into a tenant-farmer exploiting the disintegrating demesne lands. In eastern Europe, beyond the Elbe River, however, the reverse was happening. There the great lords were tightening their hold on the peasants' lands. Peasant was becoming serf and serf becoming slave. Thus, the old balance within feudalism was being tipped one way or the other.

Trade and industry were similarly changing in character. The traditional organization of the city crafts into guilds was breaking up, and companies of merchants, on the one hand, and of journeymen, on the other, began to assume responsibility for the specialized functions of their members. Germany in the 15th century suffered continual urban conflicts as the guilds struggled among themselves for power and privilege, and everywhere in Europe guild control over industry receded in the face of an almost universal tendency for individual urban capitalists, or groups of capitalists, to assume control, through mercantile capital, over the entire process of manufacture in textiles, leatherworking, mining, and the metallurgical industry.

The Hanseatic League

The structure and direction of international trade was also changing. The cities of the Hanseatic League (a league of towns in northern Germany that banded together in the latter part of the Middle Ages to protect and develop their foreign trade), which had long been dominant economically and politically in northeast Europe, had fallen into fratricidal war. Their conflict with the Scandinavian powers in the 16th century virtually brought about their final collapse, and this, in turn, had a depressing effect on those important sectors of economic life in such cities as Bruges, in Flanders, and London, which had been closely dependent upon the League. North Italy, which had long been the most advanced economic area, was also under severe pressure. The four great cities of Venice, Milan, Genoa, and Florence, which had dominated the life of the Po Valley, the Tuscan Plain, and large areas of the Mediterranean, were now faced with new competition and problems. The great voyages of discovery had opened up by 1500 an alternative route to the Orient around the southern tip of Africa, thus threatening the overland spice trade of Venice, although the effects of this were slow to be felt and their incidence was only intermittent until well into the 17th century. More serious was a slow but steady shift of the centre of European economic activity away from the Mediterranean toward the Low Countries. North Italy still remained an advanced area of wool and silk production and glass and porcelain manufacture and a centre for the great merchants in spices, grain, alum, wine, sugar, and other exotic products. Venice and Genoa also showed a sinuous flexibility in meeting the competition of the Portuguese spice traders. But the 16th century was to be the age of Antwerp, in the Netherlands, rather than of Venice, and its successor in the 17th century was a north Netherlands economy dominating the world from the great entrepôt of Amsterdam.

The influence of population on economic growth. Modern research has established almost beyond doubt that behind the new impetus stimulating the developments of the 16th century lies the reversal of the population decline of the 15th century. In its last decades there began the steady growth that may be detected everywhere in Europe throughout the 16th century and seems to level off only by the middle of the 17th century. By 1600 Europe had a population approaching about 100,000,000. Of these, perhaps one-half lived in countries bordering the Mediterranean. This probably represented between 20 and 30 percent more than in 1450 and was to increase by another 30 or 40 percent by the middle of the 18th century. Most countries shared in the 16th-century increase, especially Italy, England, the Low Countries, and France. By 1700 it was apparent that the increase was slowing down and even declining in some countries. The population of Germany was severely cut by the disease and poverty that followed the Thirty Years' War, and the population of Spain may have dropped from 8,000,000 to 6,000,000 in the 17th century. But, in general, the European economy reflected the buoyant increases in areas such as England and the Low Countries, especially in the area that today constitutes The Netherlands, where the population virtually doubled between 1500 and 1650.

The reasons for this increase have been much debated. Population movement must depend upon the balance between the birth rate and the death rate. Mortality, especially through the bubonic plague, which attacked everywhere sporadically, was extraordinarily high, but the incidence of plague seems to have slackened in the 16th century, and it finally disappeared from western Europe by the turn of the 17th century. The declining virulence of the plague, coupled with the removal or relaxation of restraints on early marriage, thereby facilitated a rise in the birth rate. This, in turn, was followed by urban growth, the decline of guild and apprenticeship laws, and the increase of individual farm holdings.

Bubonic plague

If the origins of this demographic revolution are obscure, its consequences, on the contrary, are quite clear. Everywhere it generated an increased demand for food and the necessities of life. This increase, in turn, not only became a major factor in the great inflation of the 16th century but also brought into being a new class of entrepreneurs, or middlemen, who seized the opportunities for trade and profit that it offered.

The rise of Antwerp and Amsterdam. The rise of Antwerp in the first half of the 16th century to the position of the leading entrepôt in the world vividly reflected this new movement. From about 15,000 in the mid-15th century, its population had passed the 100,000 mark by the middle of the 16th century. Antwerp was supreme during the half century from 1520 to 1576 as the warehouse and market of world trade. It not only absorbed the industrial production of its own textile centres, such as Ghent, Lille, and Ypres, but it also formed a sophisticated centre for merchant adventurers, bankers, and speculators from all over Europe. The merchant adventurers from England concentrated the largest part of the export trade in English cloth in Antwerp. The south German bankers (Fuggers, Welsers, Paumgartners) traded in cloth, spices, and metals with Germany and Italy. Italian merchants sold the produce of north Italy into northern Europe through the Antwerp market, while agents for the kings of Portugal sold great cargoes of Indian pepper, nutmeg, cinnamon, and cloves. Ships from Cádiz brought in cargoes of Spanish wool and wine and, later, of silver, to pay for purchases of cloth, iron, coal, and glass or to repay the financiers who fed, armed, and clothed the Spanish and Imperial armies.

Antwerp was not only a centre of trade but also of finance, where the European governments borrowed on a vast scale, especially for war purposes. Queen Elizabeth I's agent, Sir Thomas Gresham, acted as royal factor from 1553 to 1574, with the duties of raising long- and short-term loans for the Queen on the Antwerp money market. Most other kings and princes did the same. Among the great banking houses were those originating from south Germany, of whom the Fuggers were the greatest. Starting as weavers and manufacturers of cheap textiles in Augsburg in the late 14th century, the Fuggers had risen to be a world trading and banking house. They were now among the leading exponents of the arts of public finance, farming the revenues of the emperor and advancing capital against concessions of silver, copper, and iron mines in the Habsburg territories. They were thus the greatest subsidizers of European war and dynasticism, lending 500,000 florins to secure the election of Charles V as Holy Roman Emperor in 1519.

The Fugger banking houses

The sack of Antwerp in 1576 by Spanish troops added severely to the economic problems that beset Antwerp increasingly after the 1550s. Its capture by Alessandro Farnese, duke of Parma, in 1585 sealed its fate, for the Dutch blocked the approaches of the Scheldt River and virtually ended Antwerp's career as a world market for two and a half centuries.

The stream of refugees from the southern Netherlands began in the 1550s and continued for half a century but was at its peak in the 1570s and 1580s. Up to this time the northern Netherlands were much poorer than their southern neighbour, but, with the aid of the migrants, much wealth, enterprise, and skill was transferred across the great rivers from the vulnerable south to the more defensible north. Amsterdam, seceding from Spanish loyalty to the rebels in 1578, swiftly succeeded to the position lost by Antwerp. At Leiden immigrant merchants and textile workers from the area around Ypres launched a variety of woollen industries making the "new draperies." At Haarlem others set up bleaching centres that attracted great quantities of coarse linens from Germany. Other southern industries, such as silk weaving, glass blowing, paper making, sugar boiling, distilling, and printing, also moved to Amsterdam.

The growth of international trade. The great migration out of the southern Netherlands exported much talent and wealth also to London, Norwich, Colchester, and other English towns and cities, bringing the new textile skills and much mercantile and financial technique to England. But at this stage the Dutch republic was the greatest gainer, and between 1585 and 1620 Amsterdam became the centre of a world network of trade. This network was extended until it reached from the Baltic to the Mediterranean, from the British Isles deep into Germany, and throughout the 17th century it was increasingly linked to India and the East by the operations of the new Dutch East India Company (1602) and to America, the West Indies, and Brazil through the Dutch West India Company (1621). As a victualling station, the southern tip of Africa was settled with Dutch farmers, while Dutch slave traders built castles on the coast of West Africa. As merchants, financiers, shippers, and drainage engineers, the Dutch were as ubiquitous in the 17th century as were the Scots in the 19th.

Dutch East India Company

The Bank of Amsterdam (1609), the Loan Bank (1614), the Bourse (1609), the Marine Assurance Chamber (1598), and the Grain Bourse (1616) collectively formed the central machinery of an economy through which a large proportion of world trade passed. Except for the great herring fishery and the cloth industry of Leiden (the largest in the world), the Dutch were middlemen, carriers, and brokers rather than industrial producers, but their commercial skill, advanced financial techniques (including a fully developed stock exchange and company-share system), together with a low rate of interest, abundant capital, and a willingness to put it at risk overseas as well as at home, gave them a commanding position in the 17th-century world economy. The silver that represented the profit won in their trade with Spain was not hoarded but used to purchase goods from areas such as the East Indies and the Baltic, and these, in turn, were resold to the European nations, including Spain itself. Above all, the Dutch were the first to develop a purpose-built, oceangoing boat (the "flyboat"), which was unarmed and cheap to build and operate. With the cheap freight rates provided by the flyboat, the Dutch captured such great bulk trades as corn, timber, salt, and sugar.

Repression in the agricultural sector. The demand for Baltic grain, arising from the general increase in European population, which had everywhere turned food surpluses into deficits, was the greatest commercial opportunity seized by the Dutch. Throughout the 17th century a procession of Dutch flyboats passed through the Danish Sound carrying cargoes of grain from the southern shores of the Baltic Sea. To supply this demand, the large-scale farming system, based everywhere on the process that came to be known as *Bauernlegen* ("repressing the peasants"), spread throughout central and eastern Europe. Throughout eastern Germany, Austria, Hungary, Bohemia, Poland, and Russia, a process of increasing serfdom set in, culminating by the mid-18th century in Russia, where peasants were brought and sold like cattle. The rural population repressed thus had actually enjoyed more freedom than their counterparts in medieval western Europe, but now the great demesnes grew as prices rose and corn exports increased. Germans and Slavs were lumped together in virtual slavery in Mecklenburg and Pomerania. In Bohemia the German Catholic conquerors imposed brutal terms on the Czech heretic population. Conditions did not improve until the early 18th century in central Europe, and in Russia serfdom survived until the 19th century. Everywhere, the system produced economic stagnation, depriving trade and industry of capital and labour.

Increase in serfdom

Agriculture in western Europe. The agrarian system around the Mediterranean was less harsh, but, even there, the scarcity of fertile land, such as the Po Valley, the Roman Campagna, and the Tuscan Plain, gave rise to severe social problems. In Italy the limited areas of cultivable land demanded heavy capital investment, either for irrigation or drainage. The nobility or the municipality that provided the capital and carried out such works naturally demanded a lion's share of the profit. Hence the bitter comment of a 16th-century writer: "The plains belong to the lord, the mountains belong to the peasant."

Yet, sun, bread, and wine were not lacking. Rice and corn were among the crops that the enterprise of Venetian and Milanese landowners added to the Mediterranean scene. Italy therefore stood halfway between the stagnation of the East and the modest progress of the West, where some increase in productivity was discernible. England, northern France, and western Germany formed an area quite different from that in the East. In the West, lords tended to become *rentiers*. Peasants redeemed their quit rents, achieved the status of something like owner-occupiers in France, or became tenant farmers in England or declined into landless labourers. Providence offered a variety of fates; but all were free men, and the society of which they were a part could move on to a commercial or industrial future, exchanging feudalism for a reformed monarchy or constitutionalism.

Agricultural productivity. The consequences of the different agrarian economic systems and social structures adopted east and west of the Elbe were very marked. The social organization and capital investment in Western countries, together with modest improvements in plows, implements, drainage, manure, and enclosure, resulted in higher yields and a perceptible increase in productivity, especially by the second half of the 18th century. In east Europe, serfdom and stagnation produced an actual decline in productivity from 1650 onward.

THE RISE OF THE ENTREPRENEUR AND THE LABOUR MARKET

The 16th century was an age of industrial expansion that in some places, such as Britain, continued into the second half of the 17th century. Elsewhere, it levelled off by the middle of the 17th century, as, for example, in Holland. In Germany, progress was halted by the Thirty Years' War. This age of expansion resulted from many related causes: price inflation, population growth, the demands of governments in the new centralized states for war and court luxury (France, Spain, Sweden, and Britain).

Increased demand led to greater production. Since industry in this period most often meant production of consumer necessities, it relied for its raw materials everywhere on the countryside. Agriculture produced the wool, flax, and leather that formed the basis of the clothing industries supplying rich and poor alike. It also provided the corn for the baker and barley for the brewer. Brewing was one of the few great capitalist industries employing intensive capital equipment in the centuries before the Industrial Revolution, and the great brewers of London and the large cities of Holland and Germany were among the largest contemporary capitalists. Vineyards supplied the wine industries of France, Italy, Spain, and Germany. In addition, princes and noble landowners

Brewing

owned the rights over minerals that lay under the soil of their estates. They, therefore, exploited their mines and the deposits or (more commonly) sold concessions to urban capitalists to do it for them.

Role of the entrepreneurs. Increasing demand raised the incomes not only of urban entrepreneurs, who seized the opportunities offered by new demands, but also of those noblemen and gentry who, by luck or shrewdness, extracted opportunity from their estates. Throughout the enormously varied mining and metallurgical enterprises of western and eastern Europe (including coal, iron, tin, silver, copper, lead, and alum) and the glass and ceramics, chemical, textile, paper-making, and many other industries, there is clear evidence of the increasing role played by the capitalist organizers. These consisted mainly of city merchants whose capital was increasingly required to finance growing output and new technology. Such changes are most clearly reflected in the growth of these merchants in cities like London; Antwerp; Amsterdam; Cologne, Augsburg, and Nürnberg (Germany); Lyons (France); and, to a lesser extent, in smaller cities such as Norwich (England), Leiden, Haarlem, Hamburg, Barcelona, and Milan. In the Dutch cities they formed after a generation or two a merchant elite influencing the policies of the state at home and abroad. The philosophy of the Dutch republic is most vividly and accurately represented in *Het Interest van Holland* (1662; *Political Maxims of the State of Holland*), allegedly by Johan de Witt. In reality, this was the work of a great Leiden textile entrepreneur, Pieter de la Court. Its theme was the need for a context of peace, neutrality, tolerance, and economic freedom in which trade and industry could flourish. In London, a representative figure earlier in the century was Alderman Sir William Cokayne, confidant of the King and the contriver of 12 great commercial and industrial schemes, including a project for modernizing and expanding English textile manufacture. This came hopelessly to grief, however, and precipitated one of the greatest crises of the age by 1618. Another entrepreneur was Sir Hugh Myddleton, brother of a lord mayor of London, a mining contractor of Welsh origin with great enterprises in Wales and founder of the New River Company, which between 1609 and 1620 dug a freshwater channel 40 miles (60 kilometres) long to supply London with drinking water. In Europe the south German bankers performed similar functions as financiers and organizers for German industry. Dutch capitalists originating from Liège similarly exploited the iron and copper resources of Sweden. The Trips and de Geers were the most prominent of a large group of Netherlanders who settled in Scandinavia as agents of the monarchy to develop mining and metallurgy.

Early industrial schemes in London

Use of capital. The accumulated wealth of these new merchant princes was either saved and loaned as capital to borrowers, including the state, traders, and industrialists, or spent on personal consumption. In England a good deal of their wealth was spent on social advancement. Cokayne, for example, succeeded in getting all his daughters married to peers. Great merchants commonly moved out of the city and bought country estates for themselves in the counties surrounding London. Dutch merchants, though tending to remain in trade, also built themselves country palaces on the Vecht River near Amsterdam, just as Venetian merchants built great mansions along the Brenta River. Both invested capital in the countryside, purchasing farms and undertaking drainage and general improvement schemes. In the middle of the 17th century, Amsterdam capitalists financed and implemented the drainage of great areas of the English Fenland under the eye of the great Brabant engineer Cornelius Vermuyden.

Rise of the landowner-entrepreneur. Such enterprises were not entirely limited to the urban merchant. Everywhere in Europe, landowners joined in exploiting their own mineral deposits, such as the Lowther family in northwest England and the Dudleys in the English Midlands. At Stolberg in Germany, the counts encouraged mining and iron working on their estates, while Duke Julius did much to organize and modernize mining in the Ober Harz. One of the most famous woollen mills in central Europe was that of Count Waldstein at Oberleutensdorf. Many other nobles used the forced labour of peasants to develop textile production in Bohemia and Moravia. Others, in Silesia and Bohemia, exploited clay deposits on their estates to develop famous porcelain manufactures. Many a princely palace had a *Porzellan-zimmer* ("porcelain room") exhibiting products from the estate. Especially in England, but in many other countries also, there was intermarriage between members of these commercially oriented noble families and their aspiring bourgeois partners.

Technological innovations and improvements. The early modern period was not an age of revolutionary invention, but industrial expansion in some cases led to a number of improvements, mostly of an empirical kind. In mining, galleries and shafts were improved. Hoists brought up the ore and water. As early as 1550, in Bohemia, pumps worked by water power made possible shafts sunk to a depth of 1,300 feet (400 metres). Waterwheels were used to drive hammer forges and rolling mills. New methods of making brass and copper and lead alloy were discovered. Blast furnaces replaced open hearths. Coke was substituted for charcoal by 1700 to smelt lead, tin, and copper ores. Later, Shropshire ironmasters smelted iron similarly. In 1712, in England, Thomas Newcomen's newly invented steam engines were installed in Staffordshire mines to pump out surplus water. In the Netherlands windmills were improved and used for sawing timber. Everywhere textile industries developed improved equipment: fulling mills were driven by water; the Dutch at Leiden developed a multiple ribbon loom that was borrowed for English use; Haarlem improved techniques of bleaching linen in the early 18th century, substituting sulfuric acid for lactic acid; William Lee, an Englishman, had invented his frame knitting machine in 1589. In spite of opposition, mechanization slowly established itself in the succeeding centuries.

Use of steam engines and windmills

Need for commercial capital. These technological improvements that accompanied and made possible economic expansion had one thing in common: they cost capital; and these higher costs, mainly for working capital but also for fixed plants, equipment, tools, and buildings, explain the need for a new organization in industry and for a new type of entrepreneur to carry it on. Probably the most highly capitalized industry of all was brewing. By the end of the 16th century, the great London brewers, many of them of Dutch or German extraction, were producing beer in breweries that already bore a resemblance to a modern plant. To a lesser extent, this was also true of the great glassworks of London and the English Midlands. A late 16th-century Leipzig merchant of great wealth, Henry Cramer, had a textile factory, representing a large capital outlay, that included fulling mills, weaving sheds, a dye plant, and a complete residential area for the workers. In the Swedish iron and copper industries, capital was provided by the so-called *Brukspatroner* (proprietor), who provided the necessary credit. In England, the Netherlands, and in a rather less sophisticated manner in many centres of France, Germany, Switzerland, and Bohemia, the complicated processes needed to produce a piece of cloth—wool sorting, carding or combing, spinning, weaving, fulling, dyeing, and finishing—were organized under the control of an urban capitalist. His function was to provide credit, pay wages, set the standards and designs to be followed, and, finally, to seek markets for the product.

The spread of industrial skills. Much transfer of skill was involved in the rise of the pre-Industrial Revolution's manufacturing system. The nurseries of technology can be identified as north Italy, the Low Countries, and south Germany. The 16th century had no blueprints, so both capital and technology were transmitted across national boundaries in the most highly personalized way. Colonies of Italian merchant bankers showed Antwerp and Bruges the secrets of double-entry bookkeeping, and Italians dominated the London money market until well into the 17th century. One-third of the population of Norwich in 1600 was of Dutch origin. There, as in London, Colches-

Italian merchant bankers

ter, and many manufacturing centres in the east of England and the Midlands, Dutch, Flemish, and Walloon experts were to be found in textiles, brewing, glassblowing, printing and bookbinding, horticulture, and many other occupations. The Höchstetters and other south Germans bought concessions to develop the Cumberland mines of England, as they did in Spain and throughout the territories of the Holy Roman emperor. The process continued after the expulsion of the Huguenots from France in 1685, when another wave of highly skilled merchants, artisans, and professional men flooded into the Dutch republic, England, and Prussia, where they reinforced strongly (and in some cases originated) the processes of invention and development in such industries as linen and silk manufacture, clock making, paper making, and hat making.

The guild system and the labour market. Thus, nobles, merchants, and, in some cases, middlemen rising from the ranks of industry itself, inaugurated major changes in the early modern economy throughout Europe. Rising demands caused industry to outrun the supply of skilled labour (journeymen) controlled by the urban guilds. Therefore, the unskilled and unorganized rural workers were recruited into a new industrial system known as the "putting-out" system, or domestic system. Under this system, the capital, material, and, sometimes, tools were supplied by a merchant capitalist, who coordinated the different processes necessary to manufacture and disposed of the final product on the market. The processes (*e.g.*, combing, spinning, weaving, etc.) were carried on by a widely dispersed system of cottage labour.

This did not mean the guilds immediately disappeared. In some places, for example London, they changed their nature as the great merchant companies gained control of the craft guilds through their control of capital and power over the market. In other English towns, after a long series of amalgamations, the guilds weakened and finally disappeared. There were few left by the end of the 17th century, and industry increasingly was left to its own devices. The Industrial Revolution of the 18th century was able to spread all the more rapidly because only the vestigial remains of the old guild system were left.

Survival of the guilds

In France the economic reorganizations of Jean-Baptiste Colbert revived the guild structure largely to create revenue for the crown, but even in France guild power was weakened by the rise of the putting-out system in rural areas and the grant of licenses and exemptions by the crown to individual entrepreneurs in return for a fee. In Spain guilds actually increased—a characteristic symptom of Spanish backwardness. Likewise, in Germany and central Europe, the roots of the guild system lay deep and withered away only slowly as the enlightened despots of the 18th century introduced liberty of craftsmanship. In Italy the situation varied. In Rome guilds grew in number, whereas in Tuscany local rulers subordinated guilds to themselves. In Venice their power seems to have constituted a major obstacle to technological progress, as in textiles, for example. In Poland, as in Germany and Spain, the guilds remained strong. In Russia, oddly enough, there was far less regimentation until Peter I the Great reinforced the guild system as part of his westernization program.

Broadly, guild power was undermined most easily in the great entrepôt ports of northwest Europe, where capital and enterprise were relatively plentiful and where sea and water transport enabled economic change to come about most easily. The putting-out system, which had existed here and there in the medieval economy, now spread to many branches of production—such as mining, metallurgy, textiles, printing, and paper making. These were all industries demanding capital for equipment, raw materials, the payment of a large labour force, and a knowledge of marketing. Under the new system, guild and free labour might be combined. Spinning and weaving could be done in the countryside; combing and dyeing, in the towns. Thus, hierarchies of workmen grew up appropriate to each individual industry. In the metal industry of Solingen (Germany) the swordsmith, the temperer, and the sword furbisher all took part in the manufacture of a sword, but only the last (the *Reider*) was allowed to market the product. The same tendency for the man who was nearest to the market to assume control of the industry can be seen throughout the Birmingham and Sheffield industries making small ironware.

Emergence of large-scale enterprises. Large-scale production was not typical of the 16th century. Here and there an entrepreneur might concentrate production under one roof, as the famous Tudor manufacturer Jack of Newbury (John Winchcombe) is said to have done, employing some 500 or 600 adults and another 400 or 500 children. Less spectacular instances were to be found in late medieval Florence and Flanders. In the late 17th and 18th centuries such examples multiplied, however, not only in western Europe but also in Bohemia and Russia. A sailcloth factory in Moscow employed over 1,000 workers in 1729. The van Robias textile factory at Abbeville, France, employed many thousands of workers living and working in a large compound. A Bavarian copper mine had 1,000 workers, and a Tirolean silver mine employed 4,000. Some English coal mines had between 500 and 1,000 workers in 1700. The Swedish ironworks at Stiernsund made a variety of ironware, using machinery and a large labour force, as did the London and Zaandam (Holland) shipyards and the Amsterdam, Hamburg, Liverpool, and Bristol (England) sugar refineries. The French Royal Plate Glass Company was employing 1,000 men in its workshops by the mid-18th century.

Such examples are exceptions to the general rule of a small unit of production, but they show how industry was everywhere trying to burst out of its medieval straightjacket and assume the forms of organization most appropriate to the relatively economic and efficient use of capital, labour, and technology. Transport problems tended to restrict industrial location to the places where raw materials, labour, fuel, and power were available, but, where goods could be carried by sea or canal or where workers could be attracted by better pay or conditions of life, industries grew up despite such locational restrictions. The desire of governments for greater income or more power likewise reinforced the tendency to set up industries in places not naturally suitable for them. Such examples were to be found in countries like Spain and Prussia, the 18th-century rulers of which emulated the models of Holland, France, and England.

Factors influencing industrial locations

Unemployment and social problems. This new economy, larger but more flexible than its predecessors, provided greater opportunities for employment but had greater social consequences. The estimates of Gregory King, the late 17th-century English social statistician, suggest that at least one-third of the English population was unemployed or underemployed. This phenomenon was remarked on increasingly by many observers in England, Spain, Germany, and Italy in the 17th and 18th centuries. It is a central theme of Sébastien Le Prestre de Vauban, the great French soldier turned social commentator, in his *Projet d'une dixme royale* (1707). It accounted for the attempt to gather the poor together in workhouses that were supposed to combine the functions of poor relief and larger national production. The movement was strongest in 18th-century England. Much criticized and only fitfully effective, it served as a model for the more efficient deployment of resources by private enterprise in factories that was to characterize the Industrial Revolution itself.

CAPITALISM AND THE PROTESTANT ETHIC

Historians and sociologists who have looked for the deeper motivations of 16th- and 17th-century capitalism have sometimes suggested that they may be found in Protestantism, especially Puritanism. In 1904 the German sociologist Max Weber put forward the proposition that the linked phenomena of capitalism, laissez-faire, and the justification of business as a "calling" all derived their warrant from these sources. In 1922, Richard Henry Tawney, an English economic historian, tried to substantiate this thesis. Calvinism, in particular, was designed for a rising bourgeois class that was "conscious of the contrast between its own standards and those of a

The Weberian thesis

laxer world, proud of its vocation as the standard-bearer of the economic virtues" (*Religion and the Rise of Capitalism* [1926]). In spite of counter-arguments, such theories have had wide influence. Yet, they now need to be re-examined in the light of the knowledge that has emerged since they were first put forward. This includes the realization that the growth of capitalism in the early modern world received many stimuli unknown to Weber and Tawney, including the growth of population. Some of the changes that were supposed to be specifically linked with north-European Protestantism also involved no real innovations, being rather a shift to northern Europe of economic practices well established in north Italy, the cradle of the modern European economic system.

Calvinism. Further examination of the basic theology of Martin Luther and John Calvin suggests a medieval rather than a modern economic outlook, containing little to appeal to a rising bourgeoisie. Even Calvin gave his views on problems of economic morality with caution and reluctance. They occupy only a few pages in his 20 volumes of printed works. What Calvin did was to remove the question of usury from the complex of theological pedantry and place it on a basis of simple morality. Reasonable interest on a loan was no more unjust than a reasonable rent for land, but extortionate interest squeezed from the poor was "evil and foul." His disciples took a similar line. (The doyen of Dutch Calvinist theologians, Gisbertus Voetius, excommunicated a God-fearing woman in his congregation because her husband was a pawnbroker.) Generally, Protestant countries adopted attitudes to these problems very similar to the Catholic tradition, fixing maximum rates of interest, controlling prices, and setting up public pawnshops.

The Puritan societies of such areas as the Netherlands, East Anglia in England, New England in the United States, and Huguenot France likewise contradict the assumption that Calvinism was particularly associated with merchant capitalism. The Protestants who fled from the southern Netherlands before Spanish persecution into Holland and eastern England seemed to have been mainly skilled artisans. The capitalists among them, however, constituted only a small minority of the whole. Calvinist influence is clearly discernible among the personnel and policies of the Dutch West India Company, but, in general, dogmatic Calvinism was not characteristic of the rich merchant elite of the Netherlands. They were, for the most part, liberal and anticlerical in outlook, no more inclined to trust dynasticism or theocracy in Protestant form than in Catholic form. Indeed, in the 17th century, Dutch politics turned largely on the conflict between these "libertine" merchant princes of the Dutch republic and the party known as Orangists, led by the Calvinist ministers but composed largely of the humbler social orders in support of the House of Orange. The Orangists wanted to continue the religious crusade against Spain, which the great merchants realized would be disastrous to their business.

Protestant refugees from the Netherlands

The settlement in New England in North America was theologically Puritan and socially medieval. It was not until the middle of the 17th century that newer immigrants from the City of London began to challenge the older ideals of social stability, order, and discipline inherited from the Pilgrim Fathers. Like their London or Amsterdam contemporaries, these new capitalist elements wanted mobility, consumption, even luxury, all thoroughly deplored by men of the old dispensation, such as Gov. John Winthrop.

Similar conflicts may be found among the Huguenots of the areas around Bordeaux and Marseilles in France. Probably the majority of Huguenot support came, here as elsewhere, from the lower orders together with the lesser gentry, both hard hit by rising prices. Rich merchants may certainly be found, but they were probably no more important as an element than professional lawyers or doctors.

Origins of the Weber thesis. Such observations throw considerable doubt on the Weber thesis, which may have received some of its plausibility from later evidence of the predominance of Dissenters in 18th- and 19th-century British economic life. But much of this was Quaker or Methodist and based on a theology totally different from that of Calvin. It may well have been the result of the exclusion of Dissenters from many spheres of public life after 1660. In general, it looks as if the majority of the adherents to radical forms of Protestantism came from the ranks of displaced persons from the lower orders of society. The new economic developments tended to take place beyond the control of the old medieval towns and guilds. In what is now Belgium and, later, in Holland, there was a floating population from the ports and the countryside, much at the mercy of the economic cycles of alternating prosperity and depression. Religious persecution was another force turning thousands of artisans adrift in search of new places of settlement. The enforced mobility and restlessness of the times seemed to have given a special attraction to class, status, or wealth. Radical religions that emphasized the importance of character and conduct and that even preached socially egalitarian doctrines seemed to give purpose even to humble occupations. The doctrines of election and predestination had a special appeal to those who could believe that economic success based on thrift and sobriety was clear evidence of God's favour. It seems probable that in the casual, easy-going society of the early modern world, men thus motivated toward steady and methodical work would inevitably tend to rise in the social scale. On this interpretation, while the initial attraction of the new religions to capitalists may have been exaggerated, the practice of their beliefs and social habits may have contributed to the rapid growth of a class of capitalists later on. At the upper end of the social scale also, it has been argued that the appeal of Calvinism was not so much to the prosperous gentry as to those who were in difficulties through price inflation.

The Dissenters

THE VOYAGES OF DISCOVERY AND THE PRICE REVOLUTION

The age of discovery. Between the mid-15th and the mid-16th century, Portuguese, Spanish, Italian, and English navigators provided signposts to vast areas of the world previously unknown. In 1490 the world familiar to Europeans was limited to western Europe, the Middle East, Scandinavia, and North Africa. Hazier ideas based on myth and tradition intimated the existence of a continent across the Atlantic and of Arabia, India, and China (where there had been Muslim penetration and trade). A century later, this known world had undergone a vast expansion. Spain had discovered and claimed as its own much of the coastal regions of South America, Central America, and the West Indies. French and English voyages had reached the eastern coast of North America. The Portuguese had seized many places on the west coast of Africa, discovered Brazil, rounded the Cape of Good Hope, and reached past India into the Moluccas (Spice Islands). East Africa, Arabia, Cathay, the Philippines, and Japan had thus been made known incidentally to the explorers. Here was a vast new world open to trade and enterprise and offering a treasure house of spices, exotic goods, and precious metals to those with capital, stamina, and the will to trade.

Discovery of the Americas

The development of chartered companies. The late 16th century saw the nations of western Europe—Spain, Portugal, England, and the emergent Dutch republic—locked in the first scramble for trade and empire. Some of the contestants were individual merchants who attempted to conduct their business in the new areas in the traditional way, either as individual adventurers or at most loosely associated with other merchants in a fellowship or society. This was the way the merchant adventurers had conducted a large part of England's export trade in cloth to northwest Europe. It was an association of individuals in which every member traded alone with his own stock and employees, subject only to the regulations imposed by the company in order to secure benefits for all members. This was what was called the "regulated company," and it did not represent any accumulation of common capital. Such an institution had serious defects when it came to organizing and financing the much longer voyages, employing larger armed ships with larger

crews and with larger cargoes at stake. Voyages to the new areas of discovery often necessitated heavy outlays on defensive forts and equipment to protect the adventurers against hostile attacks. Such ventures were more than an individual investor could finance, and they required the collection of capital from many investors and its professional management. Such methods had occasionally been employed in the Middle Ages. These joint-stock methods now became far more common. In the second half of the 16th century, London capitalists, aristocratic and mercantile, raised money jointly to finance Sir John Hawkins' slave journeys, Sir Martin Frobisher's search for the Northwest Passage, Sir Francis Drake's raids on Spanish America, Sir Walter Raleigh's colonization of North America, and the trade enterprises to Russia and the Levant. Similar methods were followed by the adventurers of France and Holland, but very few of them succeeded.

The enterprises most significant for the immediate future were those to the East Indies. In Holland 12 groups of investors in the 1590s joined to finance such voyages by a separate parcel of capital, but the profits were sadly reduced by the competitive buying and selling of the valuable spices by the rival entrepreneurs. In 1602, therefore, they were amalgamated into the United East India Company. Like most Dutch institutions, this was a federation—six "chambers" situated in six different towns or provinces, each with its own organization and capital and subject only to general control from the central body. The need for permanent fixed capital for forts, ships, and arms suggested that the capital should also be permanent. A stockholder who wanted to get his money back could only do so by selling his claim on a share of the capital to somebody else at the price it would fetch on the market at the time of selling. Thus, the East India Company shares became one of the first and largest blocks of share capital to be traded on the new Amsterdam Bourse.

The United East India Company

Anglo-Dutch rivalry. The English East India Company obtained its charter on the last day of 1600. A company of 220 adventurers was established, which sounded very like a regulated company, but they recognized that India was too far away to be exploited by individual enterprise. Therefore, there was to be "a joint and united stock," though each voyage remained a separate capital investment. It was more than 50 years before the English adopted the Dutch method of a permanent capital fund.

The Dutch were the pacemakers under the vigorous, ruthless leadership of Jan Pieterszoon Coen, governor in the East Indies until 1629. By 1680 they had a monopoly of the spice-producing regions of the Malay Peninsula and Archipelago. They held Ceylon and many points on the mainland of India, the Persian Gulf, the Red Sea, China, and Japan. They employed every device to maximize profits on the cargoes of Japanese copper, silk, and porcelain; Chinese tea, pottery, and wallpaper; Indian cottons and spices; Persian carpets; Arabian coffee; Malayan tin; but, above all, the precious spices of the Malayan Islands, which left Asia at Christmas each year.

The English company was less successful. It managed to make high profits on the early voyages and set up many trading posts in the East, but it was evicted by Japan and driven out of the Spice Islands by the Dutch. On the Indian mainland the costs of defense were high. Lacking the expertise and the resources that enabled the Dutch to organize a self-balancing trade in many Eastern areas, the English company was always under pressure to export much larger quantities of silver to pay for its Eastern purchases than the Dutch. Throughout the 17th and 18th centuries it faced constant criticism for draining England of silver. Its reply was that it resold a large proportion of the spices, textiles, tea, coffee, and porcelain that it brought back from India in other European markets, thereby winning a net surplus of treasure.

North American trade

The Dutch were also active in North America, where colonists had settled in the New Netherlands, along the Hudson River, trading with Newfoundland, Virginia, and the West Indies. Their hold was not strong, and it was easily broken by the English in 1664. Nor were their enterprises in the West Indies and South America completely successful. The Dutch West India Company was formed as much to fight Spain as to develop trade. Large areas of Brazil were annexed but lost again to Portugal. The sugar market was a dangerously fluctuating one. Dutch Guiana, with its coffee, cacao, and cotton, produced better results, as did the slaving ports on the coast of West Africa, which were a main source of supply to America and the West Indies. But the West India Company was not, in the main, successful.

Success and failure of the companies. Indeed, many of the European chartered companies suffered results varying from barely successful to disastrous. The Virginia Company was a costly failure. The 30 or more French companies set up in the 17th century, including an East India, West India, and Senegal, did little but consume French capital. The English Royal African Company had a checkered career, as did competing companies formed in Denmark and Prussia to exploit Africa and the East Indies.

Somewhat late in the field, France managed to retrieve an otherwise dismal string of failures in the Caribbean. In spite of the collapse of Colbert's West India Company, individual French traders did well in what was known as the Sugar Bowl, developing sugar plantations on Martinique, Guadeloupe, Haiti, and other smaller Caribbean islands.

By 1700 a century of company activity by the Western nations in America, India, Asia, and Africa had left indelible marks. South America and the West Indies had been designated forever as "Latin" America. With the Dutch trading stations eliminated, North America's future was poised between England and France. England was firmly entrenched on the Indian subcontinent, while to the East the Dutch were equally the masters of the Malay Peninsula and Archipelago, with regular trade connections to China and Japan. It is likely that the Dutch East India Company had the best record of profit. None of the other company enterprises had anything but a patchy record, and some had lost enormous sums of capital and seen little, if any, profit. Some individual merchants in Amsterdam, Middelburg (Holland), Rouen and Nantes (France), London, and Bristol (England) had made personal fortunes. Others had lost them.

Success of the Dutch East India Company

But, if the colonial trades were by no means the easy gold rush they have sometimes been thought, it must be remembered that the capital investment they represented was still only a small part of the total trade of the Western world. The bulk of trade and shipping was still inter-European. The spices, sugar, cotton, and other goods were commodities that leavened the lump, but they were far from being the dominant influence.

The price revolution. *The influx of silver and price inflation.* One commodity produced in the New World nevertheless had a direct bearing upon an economic movement that, in turn, had incalculable effects upon the economy and society of the 16th and 17th centuries. The relationship between the flow of Spanish-American silver and the price inflation of the 16th century is still a matter of debate, but many historians still maintain that the major cause of the price inflation lies in this addition to the monetary supply of the world; and most others would agree that it played at least an important role. The precise dimensions of the inflation are not easy to establish, but in Spain, for example, prices at the end of the century were between three and four times greater than those at the beginning; elsewhere in France, England, Holland, Alsace, Italy, Germany, and Poland, they ranged from two to three times their 1500 level at least. The movement was not uniform or smooth. It began about 1480 in Germany, England, and Poland, about 1500 in Spain, and about 1530 in Italy (which generally experienced a smaller rise than most other places). Everywhere, the sharpest rise came between about 1540 and the 1570s. The inflation ended, as it began, at different times in different places: in Germany as early as the 1620s and in England and Poland by the 1640s. After the mid-17th century the price level almost everywhere remained steady or even declined for more than a century.

Imports of silver from Spanish America

Basing their arguments on the theory that prices move broadly in relation to the volume of money in circulation and the velocity of that circulation, some observers from the 16th to the 20th century have seen a major cause of inflation in the swelling stream of silver from Spanish America. They include Francisco López de Gomara, a Spanish historiographer, writing in 1558, Jean Bodin in his dispute with Malestroit in 1568–78, the author of the famous English *Dialogue of the Common Weal* (1581), and the followers of Earl Jefferson Hamilton, a U.S. scholar who has done most to reveal the dimensions of the problem in the 20th century.

The sequence of events was that the conquistadores, finding silver and gold plates and statues relatively common in Spanish America, began very soon after their arrival to mine precious metals. Boom levels were reached by 1550, after the discovery of rich silver fields in Bolivia and Mexico coincided with a new method of extracting silver through the use of mercury. Up to the 1630s large supplies of silver were mined and sent to Europe. The peak years, about the turn of the century, saw an annual flow of about 36,000,000 pesos. Thereafter, the flow began to dwindle until it fell to a relative trickle by 1660. During this whole period some 18,000 tons of silver (roughly the equivalent of one year's output from present-day South African mines) was added to Europe's monetary stocks. Of the total, about one-fifth was taken by the Spanish crown, some circulated in Spanish America, while the rest was convoyed twice a year to Seville, from where it was distributed throughout Spain and ultimately throughout Europe and to the East.

Opinions on the causes of inflation. The relationship of this flow to the inflation nevertheless raises certain difficulties. It has been noted, for example, that in some countries the price rise began at least 75 years before 1570, yet Hamilton himself is inclined to minimize the export of bullion from Spain before the latter years of the 16th century. Again, the dimensions of the price rise varied not only between countries but also between commodities. Thus, in general, every price investigation has confirmed that grain and other food prices rose a good deal more than metals, textiles, industrial goods, and colonial commodities, such as spices.

Clearly the influx of silver from the New World does not answer all the questions about the inflation. The growing output of the mines of Germany between 1480 and 1540, which reached its peak in the early 1540s, may help to explain why German and Polish prices rose earlier than Spanish prices, but it is still necessary to explain the impetus behind the continental mining of silver itself; and the demand for silver detracts from the purely monetary and supply side of the problem. Turning to the demand side of the equation, the important increases in population and changes in income distribution obviously played a role. Everywhere in Spain, Germany, Poland, Holland, and England, the larger cities show a rapid growth. Recent research has tended to suggest that the most important nonmonetary change that made it possible for increased supplies of bullion to infiltrate the economy and drive up prices was a general increase of population, associated with growing demand, larger capital investment, and larger national incomes.

Effects of the price inflation. All over Europe landlords' profits grew and with them rural investment and consumption. Customary tenants and freeholders also frequently enjoyed the benefits of rising prices. Entrepreneurs, businessmen, and speculators reaped some harvest from the gap that often opened between prices and costs, though wages sooner or later caught up with prices.

Debasement of the coinage

Everywhere governments had to face increased expenditure, and, since their incomes were slow to grow, they often added to the inflationary pressures by debasing their coinage, as did the English Tudors in the 1540s and 1550s. Such debasements reached their climax in the notorious monetary manipulations known as the *Kipper und Wipper Zeit* (see-saw time) of the early 17th century. Larger rural incomes together with larger urban incomes promoted the production and consumption of more luxury goods. Thus, any satisfactory explanation of the inflation must include these demand factors as well as the changes in monetary supply.

Few contemporaries grasped the underlying causes of the inflation. Naturally, they tended to blame it all on human wickedness and greed, passing legislation to hold down prices, rents, and wages, all to little effect. The inflation and its social consequences continued. Generally, those who could adjust their incomes to the changing price level benefitted most, such as landlords who could force or persuade their tenants to accept flexible commercial leases or tenancies; tenants with security of tenure, who could enjoy higher profits if they had surplus produce to sell, while still paying fixed rents; and tradesmen, who could put up prices and keep costs down. The sufferers were landowners, whose *rentier* incomes were limited by custom, and wage earners whose wages stuck at customary levels while the price of necessities rose. It must be remembered, however, that the wage earner completely reliant on a money wage was still the exception rather than the rule in most societies. Much payment in kind was still in force. Many workers were still rural peasants with a small holding to provide them with at least a subsistence living.

The decline of Spain. It has been argued that the decline of Spain in the 17th century from its 16th-century peak of power and prosperity was fundamentally due to economic causes. The silver influx is claimed to have brought about a high degree of inflation, and, since costs rose with prices, Spain suffered a loss of economic strength that brought about political decline. This theory is not universally accepted, and other historians have argued that Spain's problems arose from the misguided determination of Philip II and his successors to burden Spain with such heavy costs of war and extravagance that the consequential high taxation and poverty drove out all enterprise and large numbers of valuable skilled workers until Spain was to become a proud but arid desert, inhabited by bankrupts, monks, and paupers. This picture, though in some respects exaggerated, gains some credibility from comparisons with other countries, such as the Dutch republic; for there is a fair measure of agreement that the same price inflation that is alleged to have ruined Spain was, at the very same time, exerting a powerful natural influence that added daily to the wealth of the Dutch.

NATIONALISM AND MERCANTILISM

The growth of nationalism. The early modern age was a time of conscious state building. One after another, governments attempted to assert control over larger areas, and the new absolutisms tried to combine economic policies designed to achieve increasing national wealth with political programs aimed at increased power at home and abroad.

England. England was quick to achieve a high degree of unification, and this was confirmed as the Tudors strengthened their hold on Wales, Ireland, and the northern counties. Dynastic union with Scotland came with the Stuarts; and full legislative union in 1707. It cannot be claimed that the economic policies that were promulgated for this increasing area were always satisfactorily implemented, but a national policy was, at any rate, outlined in matters of tariffs, employment and poor relief, price and wage regulations, and the subordination of local authorities, municipalities, guilds, and companies to the will of the national government.

Union with Scotland

France. Elsewhere, progress was slower. During the reign of Henry IV in France, a significant practical contribution was made (in the shape of road and bridge building) to a concrete national economy. Under Colbert, principal economic minister to Louis XIV, something was done to unify weights and measures, to continue Maximilien de Béthune, duc de Sully's work of transport improvements, and to reduce divisive local tolls on roads and rivers. Even so, and in spite of French absolutism, France remained, in many respects, a large collection of provinces representing a great and varied wealth of natural resources but much divided by internal economic obstructions.

Other European countries. Even the Dutch republic, the model of economic modernity of the period, deferred to much local option in the shape of powerful cities and provinces. Real economic unity was achieved only by the predominance of the powerful and wealthy province of Holland. Both Spain and Italy faced insuperable geographical obstacles to any real economic unification, both countries being divided by mountain ranges. In general, the farther east the countries were, the more backward were their economies and societies. Germany in the 17th and 18th centuries still consisted of some 2,000 independent states of widely varying size, each with its duke, count, landgrave, bishop, or abbot among others. Russia, in spite of some rapid but exceptional projects of industrialization by Peter I the Great, was not modernized until after the Crimean War. Central and eastern Europe remained a society of lord and peasant (almost lord and slave), much of its business still conducted on the basis of barter or payment in kind.

Backwardness, however, was itself the stimulus to economic change. The examples of north Italy and the Netherlands—especially the Dutch republic—were constantly before the eyes of princes, bureaucrats, and even merchants struggling to cope with natural obstacles to transport, compounded by the remains of man-made obstacles surviving from years of fiscal impositions. In France and Germany, in particular, trade was constantly obstructed by internal tolls on roads, rivers, and canals. In 1685 it was calculated that out of a cargo of 60 logs, floated down from Saxony to Hamburg, only 6 remained after the customs officers at innumerable points had taken off their share as a tax contribution. Everywhere petty fiscalism obstructed the free flow of goods, and only at sea did cargoes move relatively freely (though even these were frequently subject to the attentions of privateers, pirates, and enemy warships).

Customs and tariffs

The development of mercantilist theories. The 17th and 18th centuries nevertheless saw the rulers of these imperfectly integrated territories evolving a set of ideas with a remarkably high degree of common content. The theories were those that have come to be called mercantilist, though in practice they were as much the creation of ministers and administrators as of merchants, and their objectives were political and strategic as much as economic. In an age when war was almost continuous (the 17th century is said to have enjoyed only four years of peace), economic principles that rested on the assumption that wealth could be pursued in the abstract, without regard to the means to procure it, maintain it, and protect it by force, had little appeal to rulers. The mixture of politics and economics in mercantilism varied from state to state according to the resources, situation, and mode of government in force. In the Dutch republic the power of the state was most subordinated to the needs of trade. In England, France and, later, in Prussia and Russia, the economic initiative of the state was seen essentially as the indispensable other arm of the military and, often, aggressive intentions of the state toward its neighbours.

The balance of trade. In spite of such differences of emphasis, however, mercantilist programs in most countries came to turn upon the concept known as the balance of trade. This concept was most clearly analyzed in the large volume of economic literature that appeared in England between the mid-16th and the mid-18th century. It may be discerned in the *Dialogue of the Common Weal* (1549–81) and is stated clearly by Thomas Mun in his *England's Treasure by Forraign Trade* (1664). The central theme is that, if the nation's exports exceed its imports, the nation will grow rich by reason of the influx of bullion that will result. But if the nation's imports exceed its exports, the nation will grow poor because of the loss of bullion that will necessarily follow. The next 150 years saw many refinements in the presentation of this simple proposition, as Mun's essay was discussed by scores of English writers and translated into French, German, Spanish, and Italian. But the basic concept of the "balance," supported by new bookkeeping practices and even by the notions of the new physics and mathematics, remained central to economic thought and policy down to the emergence of the classical economists and the ideas of the self-regulating economic mechanism introduced by David Hume and Adam Smith.

Mercantilist policies. From this concept emerged, in turn, a series of economic measures designed to expand and adjust the national economy to meet the needs of the state and society. In its fully fashioned version, the mercantilist program aimed at encouraging the maximum influx of raw materials for industry by the removal of import duties and the imposition of export duties on local raw materials, wool being a frequent object of attention. Local manufactures, on the other hand, were to be encouraged by the imposition of protective tariffs against foreign competition and the grant of subsidies and export bounties to assist local manufacturers. An abundant supply of skilled labour should be provided by a large population, and, as the difficulties of calculating an exact monetary balance became fully appreciated, the main practical objective of mercantilists everywhere tended to become the employment of a larger and better-trained labour force.

Export duties

This central core of ideas is common to the English mercantilists and to French writers such as Barthélemy de Laffemas and Antoine de Monchrétien in the time of Henry IV, the two most notable forerunners of the greatest French mercantilist, Colbert. They are repeated in the writings of the Austrian and German economists Johann Joachim Becher, Wilhelm Freiherr von Schröder, Philippe Wilhelm von Hörnigk, and the German cameralists, whose ideas powerfully influenced the ambitious economic programs of Frederick II the Great of Prussia. They appear constantly in the writings of the Spanish neomercantilists, Jerónimo Uztáriz and Bernardo de Ulloa, toward the middle of the 18th century.

In all these states successive attempts to turn mercantilist ideas into legislative form may be found between 1600 and the French Revolution in 1789. In France and in England such programs were most concentrated in the century between 1660 and 1760. The years from the Restoration to 1714 in England have been called the age of commercial and financial revolution. The emphasis there was on the regulation of trade (especially foreign), while industry was increasingly left to its own devices and largely freed from guild or government interference. The most important group of laws were those passed between 1660 and 1663, which developed in practical, enforceable form the ideas behind the Navigation Act of 1651—basically encouraging the growth of the mercantile marine by the compulsory channelling of trade between England and the outside world (especially British colonies) into English ships manned by English officers and crews. The use of foreign-built (mostly Dutch) ships was prohibited in order to boost English shipbuilding. Specified commodities of economic or strategic importance (sugar, cotton, timber, dyestuffs, naval stores) were enumerated so that their illegal carriage in foreign ships could be easily checked. At the same time, shipping was encouraged and agriculture boosted by the grant of bounties on exported corn (1673), and subsidies and protection were granted to a large range of industries including woollen textiles, linen, sailcloth, paper, metal manufactures, and many other products. Thus, a whole program of import substitution was launched and developed well before the Industrial Revolution. Between 1660 and 1700 exports and re-exports rose from rather more than £6,000,000 annually, while imports rose only from about £4,000,000 to £5,000,000 annually. This increase in export values continued, though at varying rates, until 1760, when they reached a figure of nearly £15,000,000.

English mercantilist legislation

Colbert's economic policies. The French program of Colbert was very different. Colbert was essentially the royal servant, ministering to royal needs, and his major task was to reform the tax system to improve the royal revenues. This he aimed to do by promoting manufactures through tariffs, privileges, and the import of skilled artisans, so that industry should be a source of national wealth and social contentment. It would be regulated

through a universal guild system that would bring not only discipline to industry but also revenue to the crown and a reputation for high quality to French producers of luxury goods. Colbert was by no means the pacifist he has sometimes been thought, and he placed the whole of his regulated economy behind his master's aggressions in Europe, believing, as he did, that the amount of world trade was fixed and that France would only enlarge its share at the expense of the Dutch and English by force.

Financing of French wars

Effects of mercantilism. The results of mercantilist policies remain debatable. England certainly enjoyed a century of economic development. French industries likewise spread and, in some cases, seem to have benefitted, as did their English competitors, by the vigorous application of intelligence to their promotion. In an age when invention was limited and the price level stagnant or declining, it is impossible to deny that the mercantilists seem to have brought great stimulus to economic development. In Russia, the imitative mercantilism of Peter the Great has been seen by recent scholars as a remarkable example of state-promoted economic advance. In the Prussia of Frederick the Great, many attempts at unsuitable industrial development, such as silk manufacture, copied slavishly from French models, proved hopelessly unsuccessful, but the works of improvement to harbours, roads, and canals, the clearance of forests, the drainage of marshes, and the settlement of colonies of thousands of industrious Huguenots laid the basis for the industrial future of a united Germany.

It is no longer possible to write off mercantilism as an ignorant contradiction of self-evident economic truth as revealed by orthodox economic theory. Its political, military, and fiscal preoccupations were nevertheless dangerous, and, in an age when public opinion had only limited means to challenge the vested interests of bureaucracy, authority, and private monopoly, mercantilist institutions could easily become a hindrance to further economic progress. The French economy, in particular, was to suffer from the growing burden of ossified institutions, obsolete laws, and harmful privileges corruptly bought and sold.

EMPIRE AND THE WORLD ECONOMY

Decline of Italy and the Dutch republic. The first half of the 18th century saw the completion of certain momentous developments. Internationally, trade moved away from north Italy and the Low Countries to centre more positively around the two foremost expanding nation-states, France and Britain. By 1700 there was no doubt that, commercially and industrially, Italy was in a state of relative decline. The great cloth-producing centres at Florence, Venice, and Como had all virtually ceased production, though silk manufacturing continued, often outside the old cities. Venice had finally ceased to be an entrepôt of world importance. In north Europe, the Dutch republic, heir to the Italian legacy, had lost its former dynamism. Essentially a middleman economy, it was now feeling the draft of competition from Britain and France. Merchants from these and other countries, who had formerly relied on the intermediary services of Amsterdam, were now trading directly with one another. Dutch foreign trade, while not positively collapsing, remained at about its mid-17th-century level. Agriculture even prospered, but the once-great industries of Leiden and Haarlem suffered badly. Although no longer the leaders of trade and manufactures, the Dutch continued to finance an important part of the needs of other states, especially for war. Dutch investors held an important stake in the rapidly growing English national debt, and there were few smaller governments that did not at one time or another turn to Amsterdam for loans.

Anglo-French competition

Italy and Holland may be regarded as examples of relative decline. Only Spain seems to present a picture of absolute decline from which there was to be no significant recovery even in the 19th century.

British economic and industrial growth. True expansion was evident only in Britain and France, and even there the early 18th century was punctuated by hesitations and losses. Everywhere population growth had slowed down. Cities frequently seemed to have reached the limits of expansion for the time being. London, for example, after a period of unprecedented population growth to over 500,000 people, had reached a point of overcrowding that threatened health and a point of general congestion on the Thames River that threatened profits. The new century brought about little expansion of its activities and with its stagnation went a levelling off of London demand for food and drink, necessities and luxuries. Breakneck expansion gave way to consolidation and more intensive development. These changes had their effects on agriculture, industry, and population throughout the Home Counties and East Anglia, and until 1750 landlords and tenants alike faced more years of depression than of prosperity.

Population growth

The total population of England and Wales may have grown by perhaps 1,000,000 to its 6,700,000 figure of 1760. It now seems probable that this continuing, though still modest, increase was due to a fall in deaths (33 per 1,000 in 1730 to 27 per 1,000 by the 1760s), whereas the birth rate rose from a low figure of 28 per 1,000 in 1710 to over 33 per 1,000 in the 1760s. In contrast to the deceleration of London, provincial cities had begun to grow. Norwich flourished on agricultural progress and its "new draperies," but the most significant vigour was to be found on the west coast, where Bristol, Liverpool, Whitehaven, and Glasgow in Scotland were all growing rapidly on the profits of the Atlantic trade—American and West Indian sugar, cotton, and tobacco and the provision of African slaves for the plantation economies of those areas. There was also a marked shift of manufactures, especially cheaper and lighter textiles such as linens, cottons, and fustians into the hinterlands of Lancashire and Scotland to meet the demand for exports to the new colonies. Ironware, hardware, nails, tin, glass, and small arms also filled the holds of Bristol ships sailing for West Africa, America, and the West Indies.

The western ports therefore accounted for a good proportion of the total increase in British shipping, which rose from about 320,000 tons in 1700 to just under 500,000 tons by the 1760s. Another important contributor to shipping demand continued to be the coal trade. Colliers taking northeastern coal from Newcastle to London had provided a major natural stimulus to English shipping and shipbuilding from 1550 to 1700. This cheap transport had, in turn, stimulated British coal production to a 14-fold increase by 1700. This expansion continued in the 18th century and was unparalleled in Europe; however, it took a new direction. Northeastern coal production, based on the London demand, grew less rapidly, while new coalfields, especially in the Midlands, ministered to the needs of an expanding pattern of output from a range of industries of widely varying type: sugar boiling, salt refining, brewing, and many small manufactures based on iron, such as tool making and lock, chain, nail, gun, and sword making. Abraham Darby discovered how to smelt iron ore with coke in 1709, but it took between another 50 and 100 years for the invention to penetrate widely throughout the iron industry.

Nevertheless, by the midcentury there were already signs that the growth of home and overseas demand were beginning to interact. As machines and engines called for more metal and fuel, coal output, in turn, responded, and this drew forth more demand for transport and labour. There was a notable increase in labour-saving mechanization everywhere, but, in spite of a few large-scale enterprises utilizing true factories, such as the Lombe's silk mill at Derby (1724), Wedgwood at Burslem (1759), and Boulton at Soho (1762), the new industries still grew by multiplying the number of relatively small units of production, employing new labour, and exploiting established technologies, such as dyeing, that were capable of improvement and expansion. All this new growth tended to provoke manufacturers and traders to challenge the older ideas of state paternalism and company monopoly, which now seemed to be a brake on output and profit rather than a help to their efforts.

Mechanization

Expansion of French trade and industry. In spite of the pattern of technical conservatism that state and guild

control imposed on the French economy during the post-Colbert period, France also shared in the general expansion of trade and industry in the 18th century. From Labrador down to the Caribbean, the French participated in the trans-Atlantic trades, importing fish, furs, tobacco, sugar, coffee, indigo, and cotton. The French ports of Saint-Malo, Nantes, and, above all, Bordeaux rose on the new trades. Even Marseilles, though centred on the Levant and Mediterranean trade, benefitted from the re-export trade in West Indian sugar and coffee.

Certainly in the first half of the 18th century, the output of French cotton and iron was also greater than that of the rival industries of Britain. Some French industries were hindered by the rigid, monopolistic, secretive, and legalistic guild system that cramped enterprise and made it impossible to follow fashion. But elsewhere, French industry, like its English rival, sought expansion through rural development and the putting-out system. It enjoyed the great advantage of a large domestic market (three or four times the population of Britain and more than ten times that of the Dutch republic) and rich natural resources. Probably the expansion of the French export trade was even greater than that of the English, before the 1789 Revolution. French capacity for invention, especially in industrial chemistry and such ancillary processes as textile printing, was truly remarkable. Yet, in certain respects, French development may be seen in retrospect to have had less potential than its English rival. Shipping and shipbuilding were less closely geared to industry, and the same was true of the relationship between coal output and coal-using industries. The French pattern of manufacture was therefore less adaptable to the growth of mass demand. The custom known as *dérogation*, which penalized aristocrats who participated in trade or industry below the level of the great state colonial trading companies, may well have deprived the French economy of valuable financial aid and initiative of the kind that came from the shrewder members of the British nobility, such as that of Francis Egerton, 3rd duke of Bridgwater, the great canal promoter, or of the coal-owning Lowthers.

Anglo-French rivalry. Two other related weaknesses may be discerned in the French economic structure. England between 1688 and 1750 had developed a surprisingly well-balanced system of public finance, based on a well-managed public debt, an evolving system of taxation (both direct and indirect), the Bank of England, and the great trading companies, together with a widespread system of private banking. French financial institutions, on the other hand, did not succeed to anything like the same extent. Indeed, some contemporary observers saw in the weakness of French public finance and its inability to raise the great masses of capital represented by the growth in the English national debt (£12,000,000 to £132,000,000 in 1763) a major reason for British military and economic superiority that was to emerge decisively at the Treaty of Paris, which brought an end to the Seven Years' War in 1763. Militarily, France suffered in its overseas possessions—in the West Indies, North America, and India—by reason of this financial weakness, and military defeat was certainly reflected in overseas trade, not least in India, where sea power, especially, gave England the whiphand by 1761.

The national debt

The observer of these events in historic retrospect must nevertheless be careful. By 1750 the future of world power and of European control of trade both within Europe and in the colonial world beyond was still in suspense between the two great powers, France and Britain. It is not easy to determine how much of Britain's success in overtaking its rivals was inherent in the situation in 1750 and how much was enveloped in the events and developments that took place subsequently.

THE SIGNIFICANCE OF EARLY ECONOMIC DEVELOPMENTS

Modest though the advance of these two and a half centuries seems compared with what was to follow, it was by contemporary standards truly momentous. Europe's population very nearly doubled, to reach 140,000,000 by 1750. The growth in the area and volume of international trade and shipping was without precedent. The roles of capitalists and capital in the changing industrial system were as revolutionary as those of entrepreneurs and factories in the next phase of development. Finally, the men of the 17th century had begun to grasp how scientific method could be applied not only to the problems of agriculture and industry but also to those of society itself, a realization pregnant with possibilities for the future. (C.H.Wi.)

II. The Industrial Revolution

From the mid-18th century until World War I, economic history centres around a group of changes known as the industrial revolution. The term industrial revolution is commonly used to denote those changes in the processes and organization of production that mark the passage from an agrarian, handicraft economy to one dominated by industry and machine manufacture, from what is sometimes called a premodern or traditional economy to a modern one. The term, when capitalized, refers specifically to the first historical instance of this transformation, which began in Britain in the 18th century, spread from there to the Continent and to offshoots of Europe overseas (the United States, in particular), and transformed in the span of a century the life of Western man, the nature of his society, and his relationship to the other peoples of the world.

At the heart of the Industrial Revolution was a cluster of innovations in the technique and mode of industrial production: (1) the substitution of inanimate for animal sources of power (in particular, the introduction of steam power fuelled by coal), (2) the substitution of machines for human skills and strength, (3) the invention of new methods for transforming matter (in particular, new ways of making iron and steel and industrial chemicals), and (4) the organization of work in large, centrally powered units (factories, forges, mills) that made possible the immediate supervision of the production process and a more efficient division of labour. These innovations in industry promoted and were, in turn, supported by major changes in the technology of agriculture and transport.

Underlying all of this was the systematic application of knowledge to the devising of more efficient production. In the first part of the Industrial Revolution, the knowledge was mostly the result of practical experience and informed empiricism. With time, however, beginning in such industries as chemical manufacture, applications were derived from pure science, and, by the end of the Industrial Revolution, the relationship was reversed. Theory had moved ahead of practice in most domains, and inventors were tapping the pool of scientific knowledge for usable ideas and information.

The Industrial Revolution was essentially completed in the first industrializing countries by the late 19th century. By then they had made their passage to the new technology and were pursuing new paths of change. Again, there was a clustering of innovations that gave the impetus to this further growth: electrical power, the internal-combustion engine, petroleum fuel, the automobile, and science-based chemical manufacture (the first synthetics). Some have called this array of changes a second Industrial Revolution. In the second half of the 20th century, a similar clustering of innovations (*e.g.*, atomic power, electronics, and computers), is effecting perhaps a third Industrial Revolution. Each of these stages has been marked by important gains in the production and delivery of energy; in the speed, precision, and convenience of tools and machines; and in the effectiveness and usability of final products (standardization of parts, refinements of tolerances, miniaturization of components).

All of these technological advances translated into economic growth—that is, increased productivity and income per head. Productivity is measured by the ratio of output to one or more factors of input (labour, capital, materials). As a result of the Industrial Revolution, labour productivity increased thousands of times in some branches of industry (cotton spinning, for example), hundreds of times in others (weaving, steelmaking, chemicals), with corresponding decreases in the cost of manu-

Changes in productivity

facture. By contrast, those activities little affected by the new techniques—skilled arts and crafts, for example, or barbering—showed almost no gain in productivity and little reduction in costs. The effect was a massive shift of demand in the direction of those products and services the costs of which fell the most and an increase in the numbers of persons and amount of capital employed in the branches furnishing these products and services. This shift was reinforced by the gains in income per head, which made it possible for the populations of advanced industrial countries to devote a smaller portion of their income to necessities, and to spend more of it on luxuries or "extras." This expenditure over and above the requirements of subsistence is especially responsive to the changes in relative prices entailed by changes in relative costs of production. The effect was a substantial growth of those branches most affected by the new technology and, in general, a shift of resources (including labour) out of agriculture into industry and services and a movement of population from the countryside into the cities and towns.

THE INDUSTRIAL REVOLUTION IN BRITAIN

Advantages of Britain. The first country to make the transition to modern industry and economic growth was Britain. The reasons for this precocity were partly material, partly social and institutional. Britain was favoured by its oceanic position, which gave the country easy access to markets and suppliers overseas; by its modest size and highly indented coastline, which placed most of the country within easy reach of water transport; and by the abundance of those natural resources that proved to be particularly important for the new industrial technology. These included water (for waterpower, chemical processes, textile finishing, and, in the form of high humidity, for the spinning of yarn); salt, for the manufacture of industrial alkalis and acids; china clay; iron and nonferrous ores; and, above all, huge, accessible deposits of coal, for use as fuel and as a source of carbon in the smelting of iron. The one resource Britain lacked was timber, but this very handicap was turned to an advantage by the precocious recourse to fossil fuel.

The internal free trade of Britain

At the beginning of the 18th century, Britain was a medium-sized nation. Its population of nearly 6,000,000 people was less than one-third that of France; yet, Britain constituted the biggest single market in the world, not only because of the island's compactness but also because the movement of goods through the country was unimpeded by man-made obstacles. France, Britain's chief political and economic rival, was divided into three major customs areas, and traffic was further trammelled by a multiplicity of special tolls and by customary restraints on the shipment of critical commodities such as grain. Most of central Europe was even more fragmented, and the burden of feudal tolls was such as to make traffic shun the rivers (though they were more navigable than those of France and Britain) and opt for the costly and uncomfortable alternative of land transport.

This large market was also the richest in purchasing power per head. Insofar as one can estimate the incomes of mid-18th-century Europe, the average Briton would seem to have been receiving some £12 to £13 per year (perhaps $300 in today's money), half again as much as a Frenchman and more, in proportion, as one moves eastward or southward to the poorer nations of the Continent. The greater purchasing power of the Englishman was shown in his diet—he ate white wheaten bread and much meat, where the Frenchman ate dark bread and soup; and in his consumption of manufactures—he wore leather shoes rather than clogs, wool rather than linen, and used some 13 kilograms of iron per head (as against 2.7 kilograms in France) in the late 1780s. This was far more than even the rapidly growing domestic industry could supply, and the difference was made up by imports from Russia and Sweden.

The purchasing power of the British population was, in its mass, an important stimulus to agricultural and industrial production. The effect was reinforced by the relatively wide distribution of wealth, both socially and geographically. Unlike most European nations, Britain was not polarized between a small class of wealthy consumers of highly individualistic, labour-intensive luxury products and a large mass of poor peasants with little to spare for manufactures and those only the highly differentiated products of local industry. British society was more complex in its structure, with a large and heterogeneous middle class, and the peasantry was, in many parts, fully integrated into the national market. This, in turn, reflected the precocious commercialization of agriculture, which learned early to specialize in cash crops, above all for the great London market. Specialization means dependency on the outside, and one of the striking features of the English countryside in the 18th century was the extent to which not only itinerant peddlers but even fixed shops (which, of their nature, are evidence of a substantial and continuing trade) were catering to the rural population. All of this conduced to the breakdown of regional peculiarities and the homogenization of taste, which promoted a demand for the kind of standardized product that lent itself to machine production.

Social homogeneity

Coal and iron. This domestic demand, reinforced by a large and growing demand for British products abroad, placed heavy strains on the existing techniques and organization of production. One of the first industries to feel the pressure was mining, where the deeper one dug, the more difficult it became to drain the infiltrating underground water from the pits. Coal mines especially were in trouble. In one colliery, 500 horses were required to raise the water, bucket by bucket. The solution was to replace the animals with tireless machines. In 1698 Thomas Savery introduced his "fire engine," a steam pump rather than a true prime mover; and then, in 1705, Thomas Newcomen invented his atmospheric steam engine, which worked a piston up and down a cylinder, alternately raising and lowering another piston that did the actual pumping. By 1767 there were 57 of these in the Newcastle coal basin alone, supplying some 1,200 horsepower, and there were probably hundreds of them in operation throughout the kingdom. The Newcomen engine was simple and sturdy. Some of them remained in use for half a century or more. But it was wasteful of energy, and as thus largely confined to the coal pits, where fuel was almost a free good. James Watt's invention of the separate condenser (patent 1769; first commercial application, 1776), cut fuel consumption 75 percent, freed steam power from this constraint, and made possible the industrial city; and it was Watt again who first succeeded (patents of 1782 and 1784) in converting the reciprocating (back-and-forth) movement of the piston into rotary motion, so that steam could be used not only to pump mines or blow furnaces but also to drive the wheels of industry.

Another industry that felt the strain of increased demand was ironmaking. Here, the most acute bottleneck was the shortage of cheap fuel and carbon. Since England had long since cut down most of its timber, repeated efforts were made to use coke in smelting, but the necessity of mixing the fuel with the ore in the furnace made it difficult to cope with the impurities in the coke. It was not until 1709, when Abraham Darby, a Quaker ironmaster at Coalbrookdale, Shropshire, had the good fortune to work with a fairly clean mix that coke smelting became a commercial possibility. Even so, other ironworks found it hard to follow suit, and it was not until more powerful bellows became available in the 1760s that the technique spread widely. By the end of the century, the old charcoal smelting had almost entirely ceased, and British production of pig iron, once also stagnant, was growing rapidly: approximately 25,000 to 68,300 tons from 1720 to 1788 (1.5 percent per year) to about 244,000 in 1806 (7.3 percent per year).

Effect of the use of coke on ironmaking

Pig iron, the product of the blast furnace, has serious limitations for industrial use. It is brittle, cannot be worked, and must be cast. It is suitable for objects such as pots or pipes that need not take pounding or stress, but it cannot serve to build machines, frame buildings, or carry vehicles. For these purposes, before the age of cheap steel, only wrought iron would do—that is, iron

refined to eliminate the carbon imparted by the smelting process. The Industrial Revolution in iron manufacture, therefore, was not possible until a way was found to use coal here too. The breakthrough came in 1783–84, when Henry Cort patented the combined technique of puddling followed by rolling. The first decarburized the pig; the second squeezed out the remaining dross and forced the hot metal into the desired shapes.

The putting-out system. The scale of enterprise in coal mining and iron manufacture was already substantial even before the Industrial Revolution. The dependence on bulky raw materials required these branches to concentrate in those areas dictated by nature, and the technology called for a heavy investment in fixed plant and equipment. By contrast, light industries, such as textiles, were widely dispersed and relied extensively on the distribution (putting out) of raw materials to cottage workers, who turned them into semifinished and finished products. The so-called putting-out system offered the merchant-manufacturers two important advantages. First, it shifted the burden of fixed costs to the workers. All the merchant needed was a room to stock materials and finished goods. When the market was bad, the putter out simply held off, with few overhead charges to pay. Second, it gave the merchant-manufacturer access to a large pool of cheap labour. Country dwellers would work for far less than urban, not only because the cost of living was lower but also because the rhythm of agricultural labour was uneven and left long periods of relative underemployment that could be profitably used to eke out the income from the land.

The putting-out system marked a major advance over the artisan workshops of medieval towns. By freeing enterprise from the constraints of town guilds, it opened the way to growth and technological innovation. Britain's precocity in this regard (by 1400, over half the wool cloth made was woven in the countryside) changed what today would be called an underdeveloped country, exporting its raw wool to manufacturing centres abroad, into Europe's biggest producer of cheap textiles for export.

Textiles and machine technology. The growth of demand in the 18th century pushed rural putting out to its limits, especially in spinning, where it took four or five persons to keep one weaver supplied with yarn. Given the high cost of land transport, the extension of the radius of labour recruitment increased disproportionately the cost of distribution of materials and collection of finished work. Nor was it possible to get more work from those already engaged by increasing wages; on the contrary, the worker tended to set his task by his need, and increased payment only meant that he could earn what he needed with less work. At the same time, the temptation to embezzle the raw materials increased, at the expense of the quality of the final product.

It was in this context of frustrated entrepreneurial opportunity that manufacturers sought for some way to bring the labour force under control. One way was to concentrate work under one roof, so that the employer could supervise performance. This solution, however, was not feasible in the context of the old technology. So long as enlarged workshops had to rely on human power and on the same machines used in rural cottages, they could not compete, for they were burdened with higher fixed costs and had to pay their workers more to lure them into a disciplined environment. The answer was found in the invention of machines that grew to be too large and costly for cottage work—so big, indeed, that they had to be driven by animal or inanimate power: the spinning jenny, by James Hargreaves in 1765; the water frame, by Richard Arkwright in 1769 (the prototype of all modern ring spinning devices); and the mule (so-called because it was a hybrid of the two), by Samuel Crompton in 1779.

Machines for spinning

This sequence of inventions shifted the bottleneck from spinning to weaving, and, for a short while, weavers lived in unaccustomed affluence trying to keep up with the output of cheap yarn. But in 1787 Edmund Cartwright invented a machine loom, which, though it took some two decades to perfect, inexorably displaced the handloom weaver. The power loom was applied first to the coarser fabrics, for only the tougher yarns could stand up to the irregular motions of the shuttle. The machine builders, however, learned to solve this problem; and by the 1820s handloom weaving was a dying craft, and its practitioners were among the most pitiful victims of technological obsolescence.

The innovations in textiles were applied first in cotton, primarily because cotton fibres lend themselves well to mechanical handling. Wool, an animal fibre, is less uniform and tractable. It also was more costly than cotton and constituted a much larger share of total costs; so that the potential cost reductions offered by mechanization were smaller. Even so, the innovations in spinning offered gains in labour productivity of the order of over 100 to 1; those in weaving, up to 10 to 1; so that mechanization came to wool also, lagging about two decades behind cotton, and eventually extended to all the textiles. The effect of prior mechanization in cotton is shown by the change in the relative position of the two industries. Between 1769 and 1829 the value of wool exports increased from £3,897,000 to £5,362,000, while that of cotton grew from £212,000 to £37,269,000.

Similar changes took place in other industries: paper manufacture, metalworking and woodworking, chemicals, machine building itself. The principles of mechanization were susceptible to wide ramification, and the factory context tended to promote change by restoring the link between enterprise and actual production that had been broken by putting out. In chemicals, more than in other branches, the lessons of science found application in industry. Here, too, the initial motivation came from a bottleneck: the high cost of washing, dyeing, and finishing the flood of fabrics engendered by mechanized spinning and weaving. The great innovations here consisted of the substitution of abundant mineral for scarce vegetable or animal resources: the French chemist Nicolas Leblanc's substitution of coal and salt for seaweed and wood in the manufacture of soda; bleaching by chlorine compounds instead of milk and sunshine; washing of wool with soap rather than urine. The manufacture of these new industrial compounds gave rise to considerable waste; the effort to dispose of these noisome by-products without damage to litigious neighbours, plus the desire to recover some of the potentially valuable components of the waste, proved one of the most powerful incentives to innovation in the chemical industry. Seen together, these technical changes made possible substantial gains in industrial output, overall product, and per capita income (see Table 1).

Table 1: Indicators of Economic Growth in Great Britain (1801–51)

	national income (£000,000; current prices)	population (000)	per capita income (£)	income from manufacture, mining, and building (£000,000)	share of manufacture, mining, and building (percentage)
1801	232.0	10,686	21.7	54.3	23.4
1811	301.1	12,146	24.8	62.5	20.8
1821	291.0	14,206	20.5	93.0	32.0
1831	340.0	16,368	20.8	117.1	34.4
1841	452.3	18,551	24.4	155.5	34.4
1851	523.3	20,879	25.1	179.5	34.3

Source: Phyllis Deane and W.A. Cole, *British Economic Growth, 1688–1959* (1962); B.R. Mitchell and Phyllis Deane, *Abstract of British Historical Statistics* (1962).

Changes in transport. The growth of a modern industry in Britain generated a rapid increase in goods traffic. For all of Britain's natural advantages, the Industrial Revolution could not have taken place had there not been a concomitant improvement in facilities of transportation. Canals can be traced in Britain back to the 16th century, but the first truly industrial canal was the Sankey Canal (1757) from the Mersey to St. Helens, Lancashire, cut to bring Cheshire salt to Lancashire coal. This was followed in 1761 by the waterway that usually marks the beginning of the so-called canal age. This was the canal of Francis Egerton, the 3rd duke of Bridgwater, from

Canals and land transport

his coal mines at Worsley to Manchester. By 1815 a network of some 2,200 miles crisscrossed the country, along with 2,000 miles of improved river navigation. The most prosperous links in the network were those that carried the heavy, bulk commodities of industry—coal, above all.

For many commodities, however, as well as for persons, only land transport was fast enough. The road network inherited from the pre-industrial economy was manifestly unable to handle the increase in traffic, and restrictions on vehicles (the imposition of wide wheels, for example, that would cut less into the road surface) were at best a palliative. Here, the answer was found, as in canals, in private enterprise. In the first half of the 18th century, Parliament averaged eight road acts a year, most of them authorizing new turnpikes; from 1750 to 1790, almost 40 acts a year were passed; and, from 1791 to 1810, the figure jumped to 55. Even so, the wear and tear played havoc with traditional road surfaces, and it was not until John Loudon McAdam and Thomas Telford introduced their new hard-packed surfaces in the 1790s that roads began to be built that could stand industrial traffic. By 1815 there may have been 1,000 miles of highway built to this standard.

All this time, the pressure built up to speed communication and travel. Coaches were built to make better times; horses were driven harder; connections were improved so that the trip from London to Birmingham, which took two days in the 1740s, was cut to 19 hours in the 1780s. By the early decades of the 19th century, these efforts had reached a limit, and the newspapers of the period are sprinkled with gruesome stories of horses snapping their legs or dying of exhaustion on the fast routes.

Railways

It was this apparently insatiable demand for speed that made the railway an instant success for passengers as well as goods. The use of rails to facilitate the movement of heavy vehicles went back centuries in mining districts; rails diminish considerably the force required to haul a given load. What made the modern railway, however, was the invention of locomotive traction introduced in 1825 on the Stockton and Darlington Railway line (primarily for coal haulage) and then, triumphantly, in 1829 on the new Liverpool and Manchester Railway. This was the beginning of the railway era. The new line had been intended for the transport of goods; but to everyone's astonishment passenger revenues exceeded freight receipts. By 1845, 1,000,000 persons were using the London and Birmingham Railway line alone; by 1850, 6,559 miles of track were in service.

The railway was probably the most impressive and significant technical innovation of its time. Not only did it enable the economy to move the unprecedented volume of goods required and produced by the new techniques, but also its demand for iron (for rails, engines, rolling stock), coal, wood, brick, and other materials stimulated other industries. The special requirements of a railway roadbed, in particular, the need for level track, inspired some of the most spectacular engineering achievements of the age. The construction of the roadbed, including tunnels, bridges, viaducts, and similar works, called for large gangs of labourers and proved to be a laboratory for new techniques of industrial management. Similarly, the unprecedented costs of railway construction called for new methods of mobilizing capital and pushed Britain in the direction of corporate business enterprise with limited liability.

All of these connections of the railway with the rest of the economy made it the most important single determinant of the level of business activity. At the peak of the railway boom of the 1840s, the industry employed some 300,000 men in construction alone—more than in the entire cotton industry—and absorbed about half the country's total investment; that is, between 5 and 7 percent of the national income. It was also a potent force for geographical mobility and contributed thereby to the efficiency of the labour market and the democratization of society—this, in spite of segregation of passengers into different price classes.

Improvements in agriculture. The revolution in industrial technology was accompanied by important changes in the agricultural sector. Indeed, these were indispensable to the expansion of manufacturing; the new industrial work force was cut off from the land and depended for existence on the production of a food surplus and on such victuals as could be purchased abroad. The problem was aggravated by a rapid growth of population, beginning in the middle of the 18th century, and by the constriction of overseas supplies during the period of the French Revolutionary and Napoleonic Wars (1792–1815).

At the start of the 18th century, Britain, in years of good harvest, was still able to send food abroad. Agriculture was highly commercialized and specialized, and the techniques of cultivation were among the most advanced in Europe. Some scholars have even argued that the innovations of the previous period—water meadows plus mixed farming, new crop rotations, marling—were more important than the better-publicized advances of the so-called agricultural revolution of the 18th century.

New farming techniques of the 18th century

Over the course of the 18th century, the techniques of agriculture continued to advance. Certain names stand out: Jethro Tull, who advocated drilling seed rather than sowing it broadcast and who also advocated light iron horse-drawn implements instead of cumbersome wooden tools drawn by slow oxen; Charles ("Turnip") Townshend, 2nd viscount Townshend, who learned about turnips and clover in Holland and helped popularize them in England; Robert Bakewell (1725–95), the most prominent livestock breeder of his day, the creator of four new breeds of sheep selected for meat instead of wool; Thomas William Coke of Holkham, the personification of the gentleman farmer and chief exponent of the so-called Norfolk system of rotation. But what really made an agricultural revolution possible was not these few public figures but the large number of anonymous improvers, adopting and adapting the new methods with an alacrity and vigour that reflected both the enhanced market incentives and a veritable agronomania, which found expression in and was nourished by a flowering of agricultural societies and publications. In all of this, the emphasis was on rational accounting procedures that could provide the basis for an exact comparison between old ways and new: the reports of a contemporary observer such as Arthur Young consisted in large part of records of outlays and receipts, county by county, farm by farm.

The introduction of these new techniques encountered a serious obstacle in much of England in the prevailing pattern of dispersed, fragmented land tenure and open fields subject to common rights of usufruct. These arrangements went back to the Middle Ages or even beyond. They reflected, no doubt, an effort to equalize the quality of soils allocated to the different members of the village community: each received a piece of each of the different kinds of land cleared for cultivation. The common rights of usufruct followed: this was the only feasible way to permit each man to pasture his livestock on the stubble after harvest. Common rights of pasture, in turn, implied concurrent sowing and harvesting, hence a common pattern of crop rotation, which usually entailed one year of three in fallow. This wasteful system locked the farmer into a vicious circle of low yields—insufficient food for wintering livestock—low output of dung—inadequate fertilizer—low yields. By contrast, introduction of clover or root crops not only did away with fallow but provided additional food to keep the animals over the winter, to build up the herds and flocks, and to accumulate stocks of dung for the growing season.

The enclosure movement

The attempt to free land from these communal servitudes and regroup holdings gave rise to what was known as the enclosure movement. In village after village, in an area covering perhaps one-third of the arable land in England, the major landholders agreed to regroup holdings into contiguous parcels that could then be "enclosed" by hedges or ditches and farmed as the owner saw fit; at the same time, meadowland and wastes held in common were divided among the landholders in proportion to their holdings. This was not the first such wave of enclosure in English history: perhaps half of the arable land had already been converted by 1750. These earlier

enclosures—in the 16th century especially—had aimed at converting arable land into large sheepwalks, to meet the growing demand for wool. The enclosures of the 18th century, however, were intended to increase cultivation.

Although enclosures were perhaps the greatest single factor in permitting and encouraging the diffusion of the new agricultural techniques, they entailed a heavy charge on the smaller landholders, who could not easily afford the legal costs and outlays for surveying, hedging, ditching, and the like. Even worse was the situation of the landless cottar or bordar, who lived by a kind of tolerance on common land, was permitted to pasture a few animals, and worked as a day labourer in the fields of his propertied neighbours. Enclosure settlements were scrupulous in their protection of the rights of those who owned land but were equally rigorous in their exclusion of those who did not. As a result, many small holders were forced to sell their parcels to those who could afford the costs of enclosure; and many more landless labourers had to leave their homes and take service elsewhere. For a long time, it was thought that these uprooted victims of enclosure provided the basis of the factory labour force. This has been disproved by the census data, which show the population of areas of enclosure growing throughout this period. This is exactly what one would expect of an institutional change that increased the demand for farm labour. Even so, agricultural employment could not keep up with increasing manpower, and it was the surplus population of the countryside that flowed to the cities. Enclosure selected many of those who chose to go; but, on balance, the agricultural revolution it helped make possible slowed the course of rural emigration.

There are no official figures on British farm output in this period, but it is estimated that wheat production, for example, almost doubled from 1700 to 1820—13 million to 25 million quarters. Even so, the agricultural sector could not supply the growing food needs of the country, and, from the 1770s on, Britain was on balance an importer of grain. To be sure, imports represented a small fraction of overall wheat consumption—less than 5 percent of home demand in the period from the end of the French Revolutionary Wars to the abolition of the Corn Laws (tariffs on imported grain) in 1846. Yet, the proportion would have been significantly larger had it not been for these protective duties, which constituted a subsidy to British farming and kept British prices higher than they would have been in a free market.

Abolition of the Corn Laws

The abolition of the Corn Laws and the introduction of almost complete free trade in manufactures in 1845–49 marked the substantial completion of Britain's Industrial Revolution and its economic pre-eminence at midcentury. The first signalled the victory of the manufacturing interest over the agricultural: as early as the second decade of the century, employment in industry exceeded that in agriculture; by 1851 it was twice as large, as was the share of manufacturing in the national product. If one added trade, transport, and other predominantly urban activities to the balance sheet, the disparity would be much greater. Understandably, the modern sector refused to go on paying tribute to the land.

The second was testimony to Britain's confidence and superiority in the world market; no one could undersell its wares (some 93 percent of British exports were manufactures). At this point, this small country, with less than 10 percent of the population of Europe, accounted for two-thirds of the coal mined in the world, half the make of iron, half the market of cotton cloth, and over one-quarter of the goods in international trade. It was a moment of dominance without parallel before or since.

CONTINENTAL EUROPE

The Continent until 1815. The other countries of Europe actually took a long time effecting their industrial revolutions. The rapidity with which some of them moved once the new technology took hold—Germany is the usual example—is deceptive. The hard job was to start moving. By comparison with Britain, the Continent was severely handicapped: geography, political boundaries, and a legacy of medieval tolls were barriers to the movement of persons and goods; forests were relatively abundant but coal scarce; incomes were lower; labour was cheaper and labour-saving devices were thus less advantageous; society was more rigidly divided by status and occupation; it was more rural, and country dwellers made poor or tight-fisted consumers; social attitudes were less favourable to business enterprise. In the lands east of the Elbe, the persistence of serfdom locked the greater part of the population into agriculture, so that the potential supply of labour to industry was drastically curtailed; moreover, since unfree men do not work so well as free, productivity was low, and the people were impoverished. Even in the freer lands to the west, however, the rural economy could not be compared with that of England. Subsistence farming predominated; there was considerably less regional specialization; techniques were less productive, and there was nothing like England's fascination with improvement. Commerce and industry had scarcely begun to penetrate the countryside.

Government sponsorship of manufacture

By the mid-18th century, even earlier in some countries, the governments of Europe were seeking systematically to promote industrial development. The reasons were primarily political: industrial growth and economic prosperity generated revenue and employed men, and money and men were the sinews of war, hence of power. The measures varied with and are a good reflection of the needs and shortcomings of these economies: subsidies in the form of money, gifts, low-cost loans, exemptions from taxes, use of public buildings, or free machines; technical assistance and the imposition of compulsory standards and quality control (particularly important in France); monopoly privileges to innovators of new products or techniques; an assured supply of labour, by assignment of serfs, prisoners, vagabonds, or such public charges as orphans and the residents of poorhouses (especially in central and western Europe); exemption of the work force from military service and the plant itself from the burdens of quartering troops; special privileges to immigrant artisans and entrepreneurs (Huguenots in Berlin, Germans in Russia, Jews throughout central Europe); an assured market via government or army orders or compulsory purchase by such unfree subjects as Jews (Germany especially); the award of places and honours to industrialists and of patents of distinction to their products. In some fields, particularly those linked to armament (powder works, arsenals) and to the supply of luxuries to the court (glass and mirror manufacture, tapestries and carpets, porcelain), the state itself became entrepreneur. In other branches, the state relied on private initiative and, when this failed to respond to the incentives offered, might order aristocrats or court purveyors or Jews or such others as were in no position to refuse to start putting their money and efforts into manufacture.

Much of this development from above was misdirected. Technical knowledge and managerial talent were in short supply, and the effort to force progress inevitably led to misallocation of resources and the manufacture of goods of poor quality at high cost. Subsidies only encouraged waste and inefficiency, while monopoly privileges eliminated the competition that might have compelled better performance. The difficulty was compounded by the preference of most rulers and their officials for large, centralized establishments—visible symbols of industrial achievement. The "manufactories," so called to distinguish them from true factories, were not centrally powered and were ordinarily equipped with the same devices used in conventional workshops. They had, therefore, no technical advantage over dispersed cottage manufacture yet were burdened with higher capital costs and were far less attractive to workers, who resented supervision and discipline—hence the need to assign labour, which was invariably sullen and unproductive. Most of these hothouse enterprises collapsed as soon as the state withdrew its support, and the setback was proportional to the amount of support. In Prussia, which was the greatest exponent of forced development, the death of Frederick II the Great, in 1786, was followed by a hecatomb of bankruptcies and liquidations.

This is not to say that the whole effort of state sponsor-

ship of industry was pointless. The privileged manufactories left behind a legacy of knowledge and skills. Many of the men who were trained in them constituted a second generation of entrepreneurs, less spoiled, more hardheaded, and far more successful.

On the Continent, as in Britain, the real progenitor of modern industry was the putting-out system. It had always been inhibited by the privileges of urban manufacturing centres and the efforts of the state to control the standards of production; but, beginning in the late 17th century, industry gained a foothold in the French, Walloon, Rhenish, and Swiss countrysides. In France, the prosperity and influence of this partially illicit industry was recognized by a law of 1762, which legalized the fait accompli. By this time, the efforts of merchant-manufacturers to reach farther afield for their labour (Norman clothiers were drawing some of their yarn from as far away as Picardy) were producing the same kinds of bottlenecks that had moved British manufacturers to turn to machinery and the factory, and there is no doubt that the same changes would have occurred on the Continent had the economy simply continued to develop along these lines. But this process was interrupted, first, by the British innovations and, then, by the French wars of 1792–1815. The first changed completely the conditions of international competition. It was at once apparent that any nation that wanted to maintain itself as a power, political as well as economic, would have to learn the new technology. The second, however, worked against this by cutting much of the Continent off from Britain for a generation and diverting resources into familiar but profitable techniques rather than more hazardous innovations.

The spread of British techniques

The extension of the Industrial Revolution to the Continent may be viewed as a process of cultural diffusion. The continental countries already had a pool of skilled artisans who might learn and apply the new techniques—clockmakers, instrument makers, millwrights, joiners, metalworkers. Even so, they needed considerable help to bring them to the British level. This was accomplished by missions of inquiry to Britain or by placing workers in British plants, by a certain amount of industrial espionage, by the purchase of British machines (usually illegally), and by engaging Britons to work in continental enterprises.

The first fruits of this diffusion were already seen before the French Revolution: a few steam engines, including some built in Europe; some thousands of mechanized spindles for spinning cotton; the first, abortive attempt at smelting iron ore with coal at Le Creusot in France (1785). During the French wars, the interruption of British imports permitted certain branches, textiles especially, to develop in a kind of splendid isolation—that is, to modernize with old-fashioned equipment. Such crucial branches as engineering and iron manufacture, however, made little headway. The one area in which isolation paid was chemicals, where the French responded to the shortage of alkalis by developing the Leblanc soda process.

The continental revolution, 1815–50. The return of peace brought both opportunity and peril: on the one hand, communications with Britain were open once again, and, on the other, British manufactures were so much cheaper and often better than continental wares that the infant industries of the Napoleonic era were menaced with liquidation. Most of the continental countries responded by setting up prohibitive tariff barriers; at the same time, both government and private enterprise bent every effort to import British skills and experience or to spy out the secrets of British technology. To combat this potential loss of superiority, the British had long since passed laws against the migration of skilled artisans or the export of machinery; but these could do no more than impede the process. By 1825 Britain's attempt to maintain a monopoly of the new technology was clearly a failure, and the ban on the emigration of artisans was lifted. Machines took longer, but in 1842 pressure from a machine-building industry whose capacity had outstripped home demand persuaded the Parliament to annul the prohibition on exports.

Determinants of continental changes

The rate of assimilation and adoption of the new technology in continental Europe varied widely. Among the more important determinants were (1) the degree of economic and technological backwardness; (2) the degree of exposure to British competition (too much could be fatal, but a certain amount could be stimulating, as in Belgium and Switzerland); (3) relative factor costs (if labour was cheap enough or capital too scarce and dear, it did not pay to install machines); (4) the political pressures for modernization (the rulers of smaller states, such as Belgium, Switzerland, and Sweden, not caught up in the competition for power, were willing to let economic development take its course); (5) the political pressures for conservatism (old-fashioned landed magnates in such countries as Austria and Russia correctly perceived industrialization as socially and politically subversive); (6) the resource base (other things equal, a goodly supply of coal, as in Belgium and Westphalia, was a potent incentive to the adoption of the new technology); (7) the size of the market (the unification of most of Germany into a customs union [Zollverein] by 1833 gave a powerful impetus to trade and regional specialization of industry); (8) the character of demand (peasants made poor customers for manufactures); and (9) the ability of the society to generate entrepreneurs (weakest in eastern Europe but also constrained in countries such as France by hostility to unbridled competition and by the conservatism of family enterprise).

The years from 1815 to 1850 saw several of the continental countries initiate their own industrial revolutions. Belgium was first, in the 1820s and '30s, with a marked emphasis on coal mining and metallurgy. France followed and, being poor in coal, concentrated on textiles and other light industries. From the 1830s on, the Swiss moved along the same path. Finally, Prussia, with large coal resources in Silesia and Westphalia, developed from the 1840s on a pattern similar to Belgium's. In all of these, railway construction was an important stimulus to development. Belgium, compact and densely populated, built its trunk network by the early 1840s. Germany was almost as quick, so that by midcentury one could cross the entire country by rail, east–west or north–south. France, by contrast, had only stubs of lines radiating from Paris and isolated stretches in the provinces. In the rest of Europe, the new technology was still exceptional, stunted by material and institutional obstacles.

The middle decades, the 1850s and 1860s. The decades of the 1850s and '60s were a period of catching up for continental Europe—not that Britain was standing still. Easy money favoured expansion: the effect of the gold discoveries in California and Australia (1848–49) was multiplied by the readiness of credit institutions to lend in a highly liquid market. The contrast was especially marked with the crisis years of the 1840s, when bad harvests, the collapse of a speculative boom in railway shares, and, finally, an epidemic of political revolutions in 1848 had first contracted the economy and then just about brought it to a standstill. Now confidence reigned everywhere, and promoters hastened to bring new commercial and industrial enterprises to the attention of investors. A striking feature of this renewed expansion was the role played by joint-stock, limited-liability corporations. These were not new; they went back to the chartered trading and colonial companies of the 17th century. But they had always been exceptional, subversive of a commercial morality that held a businessman personally liable for his losses "to the last shilling and acre." They had required the express authorization of the state and had been reserved for those few activities that could not be handled even by the collective resources of a partnership. Railway construction was one of these, and the railway did more than anything else to naturalize the limited-liability corporation. The principle was susceptible, however, of much wider application, the more so as technological advance and economies of scale increased the threshold investment required to found competitive enterprises. The contribution of limited liability was far more important on the Continent than in Britain, which had built its industrial establishment from

The role of limited-liability corporations

the ground up, plowing back profits into firms that began small and grew with the changing technology. Britain was also richer and possessed the most efficient money and capital markets in the world, so that industrial enterprises could count on generous bank credit, nominally short-term, but, in fact, renewed to cover medium- and long-term investments. By contrast, the follower countries across the Channel, which had to bridge the technological gap, were confronted with the task of paying for costly installations, not only with the industrial profits of yesterday but also with such resources as could be mustered from outside the industrial sector. In such circumstances, limited liability and joint stock were indispensable, and institutions had to be developed to take up the task of mobilizing capital and channelling it to potentially profitable ventures—hence, the importance of the corporate investment bank, already successful in Belgium but taking a new departure in the 1850s with the establishment of the Crédit Mobilier in Paris (1852) and the Bank für Industrie und Handel in Darmstadt (the Darmstäder Bank, 1853). These were followed by (and, in numerous instances, helped found) similar development banks throughout Europe. Some of them concentrated more on speculative promotions than on industrial development; the Crédit Mobilier itself, eponymous model for the new species, foundered in 1867 because of its vain effort to keep up profits and appearances by stock-market coups. Others worked better, especially in Germany, where the rapid growth of capital-intensive heavy industry, above all in metallurgy and mining, provided a large field of opportunity.

Free trade

The rapid economic expansion of these years set the stage for a major departure from the conventional wisdom of high protection. The key move was not Britain's unilateral step toward free trade in 1845–49 but rather the Anglo-French Treaty of 1860, which was swiftly followed by a series of similar agreements between the leading European economies, including the Zollverein. These were accompanied by reciprocal legislation to facilitate capital movements across national borders and by international monetary agreements providing for standardization and simplified exchange of currencies. This was not yet complete free trade, but it did bring closer the kind of international market that would, in turn, promote technological rationalization and diffusion. The Cassandras who predicted that the protected industries of France and the Zollverein would simply be crushed under the weight of British imports were proved wrong; the most efficient and strongest continental enterprises invested heavily in new plants and did better than ever. There was, however, a purge of smaller, marginally inefficient firms unable or unwilling to meet the new competition.

Improved transport

This purge of the old technology was further hastened by the vast improvement in transport technology that marked these years, in particular, by the competition of the trunk rail network of western and central Europe. The track mileage rose sharply (see Table 2), but more important than this was the completion of key connections and the consequent growth of passenger and goods traffic: 65,000,000 to 169,000,000 tons in Britain from 1855 to 1870; 4,000,000 to 44,000,000 in France from 1850 to 1870. Nothing did more to break down the autonomy of regional markets and promote the mobility of persons; internal migration, up to then confined largely to short intraregional movements, began now to develop along interregional routes. The railway, in combination with the steamship, also moved a rising current of transoceanic migrants—108,000 per year from Europe to the United States, 1830–49; 218,000 per year, 1850–69. This exodus, as much as modern market competition, hastened the demise of old trades in low-wage rural areas.

Table 2: Growth of Railway Networks in the 19th Century, Selected Countries
(in kilometres)

	1840	1850	1860	1870	1880	1890	1900
Great Britain	2,410*	9,791	14,595	21,826	25,046	27,811	30,063
France	499†	2,915	9,167	15,544	23,089	33,280	38,109
Belgium	325	625	747	869	2,724	3,249	4,060
Germany	549	6,044	11,633	19,575	33,838	42,869	51,391
Russia‡	26	601	1,589	11,243	23,857	30,957	48,107
United States	4,534	14,515	49,292	85,139	150,717	268,409	311,094

*United Kingdom. †1841. ‡European Russia only.
Source: B.R. Mitchell and Phyllis Deane, *Abstract of British Historical Statistics* (1962); France and Belgium, Statistical Annuals; W. Woytinsky, *Die Welt in Zahlen*, V. (1927).

Although the middle decades of the 19th century were dominated by the extension of known techniques to new enterprises and areas and within existing enterprises, technical advance did not cease. For one thing, considerable improvement was built into the very process of expansion; it is hard to build a new edition of an old model without incorporating improvements. For another, no technique or machine is usable without some adjustment to the specific conditions in which it is to be used, and this adjustment often entails improvements. In the meantime, the pressures of demand continued to stimulate invention. An excellent example is steelmaking, where the traditional cementation and crucible techniques could produce only small pieces at high cost, used in shears, razors, swords, watch springs, and the like. The development of high-speed machinery and vehicles, however, created a need for steel in far greater sizes and amounts than customary. From the 1830s on, then, metallurgists in several countries bent their efforts to the problem. The definitive solution was found at about the same time by William Kelly of the United States (*c.* 1851) and Henry Bessemer of England (*c.* 1854), who was looking for a way to make a better cannon: by blowing air through a furnace containing molten pig iron, the oxidation of the carbon would proceed without the application of any external heat; hence, Bessemer's claim to make steel "without fuel." Though taken up eagerly, the new technique needed some additional touches before it was susceptible of general application: first, Robert Mushet's clever trick of burning off all the carbon and then adding it as desired made possible a standardized product, and, much later, the Thomas–Gilchrist process (1878–79) enabled steelmakers to work with phosphoric iron ores and opened the greater part of the world's iron resources to the new technique. In the next decade, a second way was found to produce cheap steel, this time with even closer quality control: first, a new "regenerative" oven invented by William Siemens (mid-1850s) yielded much higher temperatures, so that iron could be refined in huge open hearths; and here again, it was a second inventor, Pierre Émile Martin of France, who worked out the technique of adjusting the mix by the admixture of scrap so as to yield a standard product (1864). The whole story of cheap steel is almost an ideal case study in the dynamics of invention: (1) the importance of serendipity, (2) the role of multiple invention, and (3) the incompleteness of initial solutions and the indispensability of subsequent improvement.

Cheap steel

The chemical industry also moved ahead; in the 1850s and 1860s artificial, synthesized dyes were developed, first in England, then in Germany. Much more than in steel-making, however, technical advance in chemistry depended on scientific knowledge and systematic experiment: artificial dyes were only the first application of the new science of organic chemistry. These first dyes, then, were a promise not only of what would rapidly become one of the most creative and flourishing branches of industry but also of a new approach to invention.

The chemical industry

By 1870 the more advanced industrial countries of continental Europe had reached a point at which they could compete with Britain on reasonably equal terms. The scale of their enterprises was still smaller, as was the place of industry in the larger economy. Yet, they had assimilated the essential techniques of machine manufacture and adapted them to their circumstances; and, where before they had been almost exclusively learners, they were now generating innovations of their own.

SOCIAL CONSEQUENCES OF THE INDUSTRIAL REVOLUTION

The question of the social consequences of the Industrial Revolution has been hotly debated since its effects were

first noted in the early 19th century, and the issue is still alive. The split has been political, with the right usually giving a favourable judgment, the left a negative one. The question is beset by difficult problems of interpretation, value judgments, and criteria, but the following general observations may be made concerning the impact of the Industrial Revolution in Britain.

Standards of living

First, wage data indicate that, on balance, those workers who shifted out of agriculture or pre-mechanized industry into the modern branches of industry improved their earnings significantly. The gap was largest for those who came from the most backward rural areas in western and southern Ireland or the Highlands of Scotland. Those who stayed behind in agriculture benefitted to the extent that labour could easily take advantage of job opportunities in manufacturing: rural wages were highest in those parts of England adjacent to the great industrial centres.

Second, the worst losers were those workers who hung on in bypassed trades, such as handloom weaving. By 1830, tens of thousands of these were living on the very edge of survival.

Third, real earnings fluctuated with prices, particularly food prices, and the level of employment, as they always had. Real wages were thus at their lowest during the period of the French wars, when wheat prices reached record heights, rose over the next 15 years, fell again in the depression of the late 1830s (despite falling prices), then moved back up from the trough of 1842. There is no agreement on what the composite trend was from, say, 1790 to 1850, though no one would argue for a marked shift either up or down; the consensus, however, is that real wages rose substantially from the middle of the century on.

Fourth, consumption of the major food staples seems to have increased, partly owing to new sources of nourishment (potatoes and fish, in particular). Consumption of textiles grew even faster, and the generalization of the use of washable undergarments thanks to cheap cotton fabrics constituted a major advance in comfort and health.

Fifth, on the other hand, housing deteriorated as thousands of migrants poured into swollen urban centres that could not grow fast enough to accommodate the influx. The population of 103 cities of over 20,000 (including London) rose from 3.5 million in 1821 to 6.8 in 1851 to 9.5 in 1871. (This, it should be noted, is a general characteristic of all industrial revolutions: because construction technology does not advance as rapidly as that of other branches, investment in housing is less profitable than other investments or can be made profitable only at the sacrifice of quality and space—this, both for systems of private enterprise and state planning.) The poor either jammed into old structures, filling every vacant space by building sheds in courtyards or converting cellars into dwellings, or moved into little jerry-built row houses that were slums before they were finished. Water supply and sewage disposal were grossly inadequate: the life of the urban poor was a compound of damp, filth, and nasty smells. All of these inconveniences were an old story; rural housing was often even more primitive; but 100 hovels scattered over the hillsides were a much less serious problem than back-to-back tenements. Density bred discomfort, disease, and neurosis. Urban death rates actually rose in Britain in the 1820s and 1830s.

Working conditions

Sixth, the conditions of work in the modern industrial sector varied widely. Generally, the larger and more modern the enterprise, the roomier, clearer, and safer the workrooms, the shorter the hours, and the higher the wages. The weaker, less-efficient firms had to save on their wage bill in order to compete. Over time, therefore, the construction of new plants and increasing scales of production yielded as a by-product a substantial improvement in working conditions. The worst feature of the new mode of production was the extensive reliance on women and children. In itself, this was not new: women and children had always worked hard and long in cottage industry. Nevertheless, the discipline of the factory (a discipline often imposed by strangers), the remorseless monotony of many of the tasks, and the physical hazard and discomfort of some of the new processes took a heavy toll.

By the first decade of the 19th century, the abuses of factory labour were already a cause of public concern in Britain. The government was at first reluctant to intervene, for fear of breaching the sanctity of private contracts, but, beginning with parish apprentices in 1802, then continuing with other children in 1819 and 1833 and with women (in mines) in 1842 and (in textile factories) in 1847, Parliament gradually increased its purview. The most important step was the passage of the Factory Act of 1833, which set a minimum age of nine years, limited the hours of children to eight per day, required employers to provide schooling for their young workers, and, most decisive, established the principle of official inspection to ensure compliance. Unfortunately, the law aimed at first to protect only workers in centrally powered textile factories, so that older modes of production were enabled to persist in abusive working conditions until the passage of more comprehensive legislation later in the century.

In general, the social consequences of the Industrial Revolution were not so uniformly bad as they have been made out but were far worse than they had to be. They reflected the unpreparedness of British society, the ideological commitment to laissez-faire, and the natural indifference of most beneficiaries of industrialization to the sufferings of those who paid the bill.

Across the Channel, observers of the British scene were inclined to congratulate themselves on being spared the "white slavery" of the cotton proletariat. Yet, their poverty problem was even more severe—though more widely distributed, hence, less visible. Their industrial labour force was paid less and worked harder, and more of their children died younger. On the other hand, data on the consumption of necessities, such as wheat, potatoes, and textiles, show substantial increases per head over time; and, since these gains could not have been confined to the well-to-do, one must conclude that people were eating more and dressing better. Housing, as always, was a dark spot, with death rates varying widely between rich and poor neighbourhoods. As in Britain, much depended on the level of economic activity: the late 1830s were hard years, and the mid-1840s, with their conjuncture of bad harvests and business contraction, produced in some areas (Flanders, Silesia) a Malthusian peak in mortality. Things picked up thereafter.

Another test of the effects of industrialization is the hypothetical question: what would have been the condition of the population had there been no Industrial Revolution? It is nearly impossible to conjecture an answer. One approach is to consider the fate of those parts of Europe, or even of the United Kingdom, that did not keep pace with the new technology. These proved invariably to be the poorest areas—sloughs of overpopulation, endemic underemployment, and underdevelopment. Like Silesia in 1844 or Flanders and Ireland in 1846, they had no reserves in time of want and fell easy prey to famine and disease. There was an autonomy to the population growth that began in the mid-18th century that made some kind of revolution in productivity indispensable if Europe was not to lose the slow gains of a millennium and sink back to the poverty of the Middle Ages.

THE COURSE OF INDUSTRIALIZATION, 1870–1914

Industrial proliferation, 1870–1914. The last part of the 19th century saw the further ramification of the Industrial Revolution into the peripheral European lands, the offshoots of Europe overseas, and—for the first time—a non-Western society, Japan. The result of this geographical extension, combined with differing national rates of productivity change and industrial growth, was a drastic revision of the international balance of economic and political power.

The United States

A new industrial giant appeared across the Atlantic. The United States was especially favoured for economic growth. It was a rich land, with varied climate and abundant natural resources. Population density was quite low, so that land could often be had for the settling, and the

return of labour was high and increasing. The existence of an open frontier was in the long run a stimulus to industry, because land could compete for labour at high wages and costly labour was a spur to technological innovation. U.S. industry early won a reputation for ingenuity, for example, in the precocious recourse to interchangeable parts—what came to be known as the U.S. system of manufacture.

The course of the U.S. Industrial Revolution may be summed up by a few key variables. From 1839 to 1899, value added in the industrial sector (manufacturing, mining, and construction) rose from $310,000,000 to $7,830,000,000 (at fixed 1879 prices)—that is, over 25 times or 5.53 percent per year. Deflating for the size of the labour force and the change in average man-hours per year, the increase was from about $385 per capita to $1,022, a gain of 2.65 times or 1.64 percent per year. Consumption of mineral fuels, a good indicator of both industrial activity and capital stock, grew 65 times from 1850 to 1910, or 7.2 percent per year; deflating for population change, the gain was still 16.4 times or 4.77 percent per year. The effect was to make the United States the greatest industrial power in the world: the U.S. passed Britain in steel output in 1890, in coal in 1898, and in raw cotton consumption at the turn of the century.

Russia

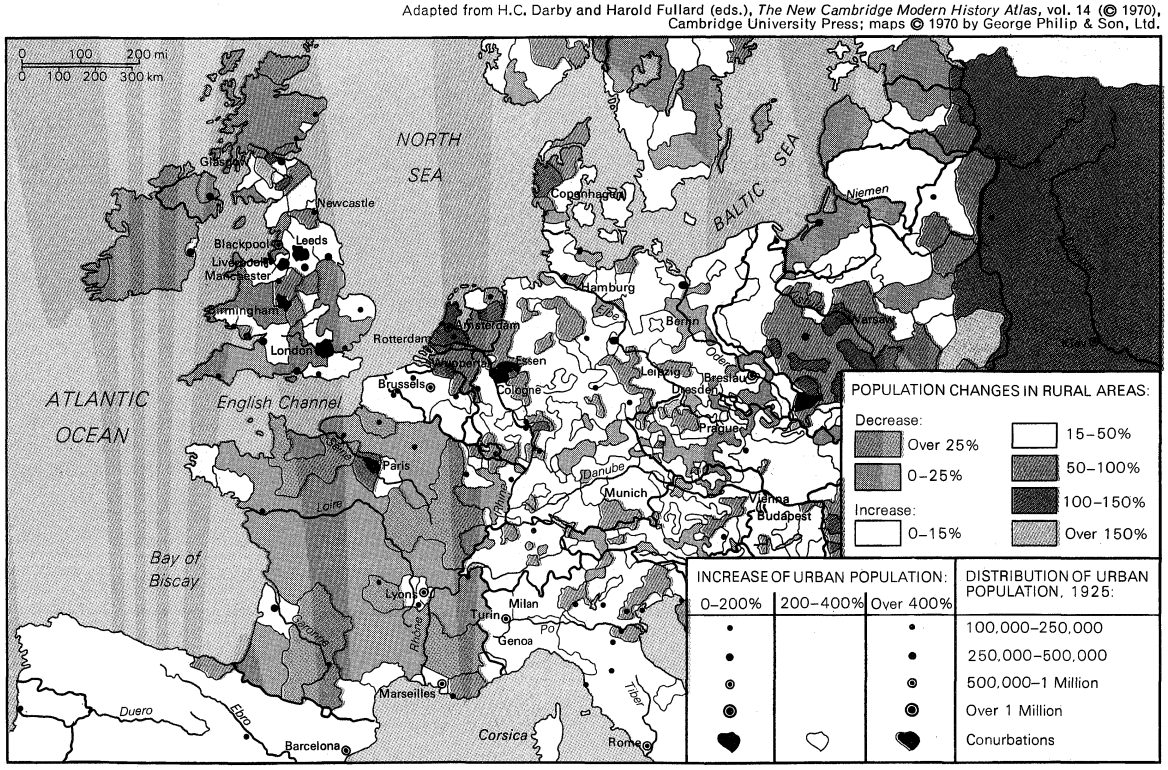
Political seers of the late 19th century were wont to pair the Russian Empire with the United States as the two giants of the future. Yet, Russia, big and populous though it was, lagged at that time far behind the United States in industrial development: in 1913 Russia (including Poland and Finland) produced some 34,000,000 metric tons of coal; the United States had reached that level in 1869 and by 1913 was extracting over 500,000,000 tons. Russian economic power was still potential—latent in the high land–man ratio, the rich natural resources, the large and rapidly growing population, and the existence of a small native cadre of highly educated scientists and technicians, reinforced by immigrant entrepreneurs and workmen. The similarity with U.S. conditions is striking. But, where the United States was enabled by its institutions and attitudes to make the most of its means and opportunities, the political and social constraints of tsarist Russia—a parasitical aristocracy, nationalist objections to Westernization, serfdom to 1861, and communal obligations and passport controls after 1861—all of these were serious impediments to innovation and to the efficient allocation of human resources. Moreover, the country was as poor as any country in Europe, and modern techniques cost far more at the end of the century than at the beginning. Government intervention and assistance were therefore indispensable to the development of social overhead capital (railroads, above all) and heavy industry. The key figure was Count Sergey Yulyevich Witte (minister of finance, 1892–1903); this was the period of Russia's most rapid growth, with a second spurt just before the war. The greatest gains were in metallurgy: once the railway opened the Donets Basin to exploitation, output of pig iron rose from 69,000 tons in 1887 to 1,507,000 in 1901—an increase of 22 times (25 percent per year). Even so, this was still a pre-industrial society; the work force in "large-scale" manufacture numbered less than 3,000,000 in a population of 170,000,000.

Japan

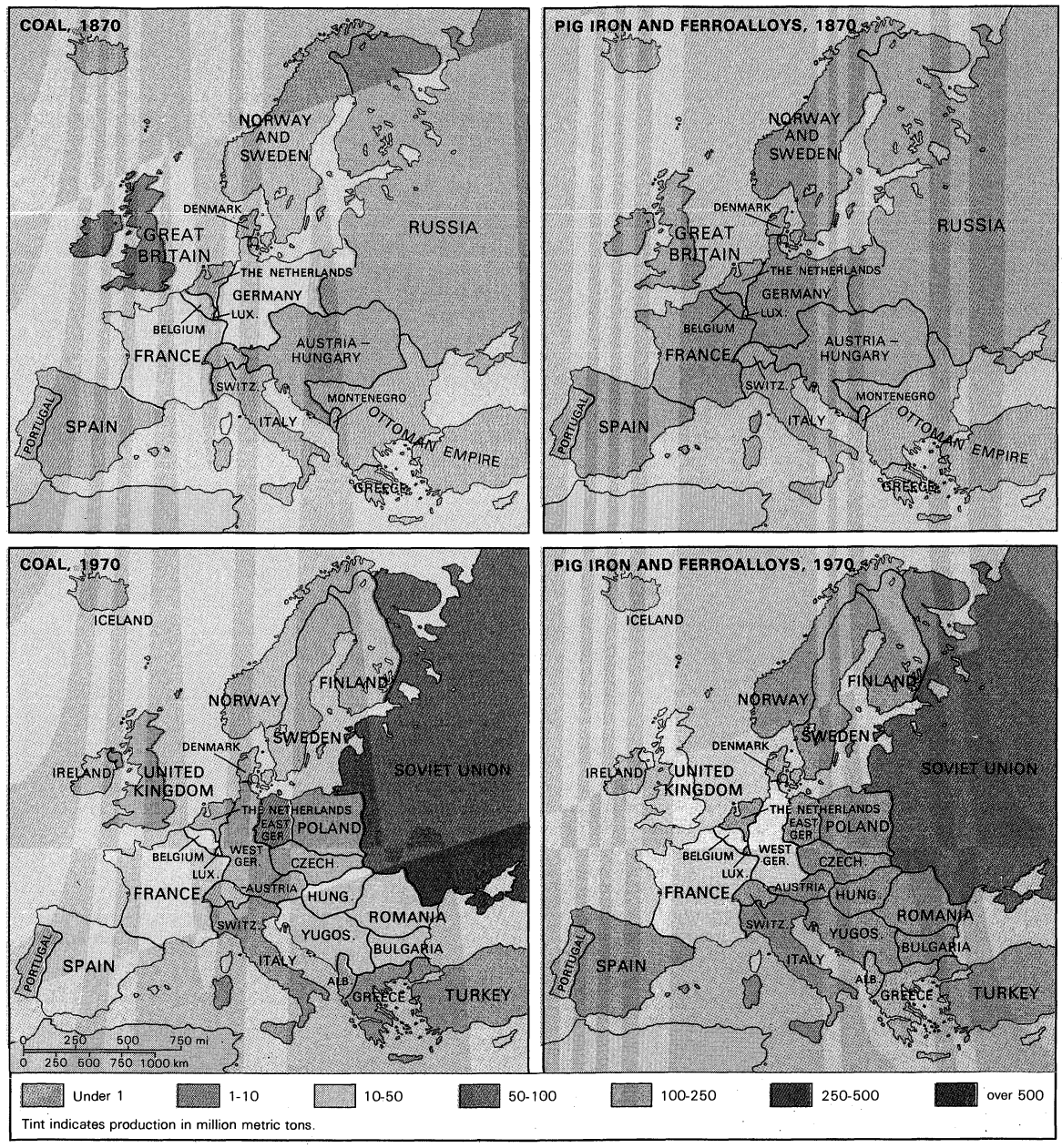
In many ways Japan was less prepared than Russia to undertake a technological revolution. It was much poorer in natural resources, at least as poor in income and accumulated wealth, and far more ignorant of modern science and technology. On the other hand, this was a much freer society, with a peasantry responsive to economic opportunity and a high degree of social and political cohesion that was to make possible the enlistment of loyalty and effort on the scale of both nation and enterprise. It was also an exceptionally literate society that was capable of learning quickly. The Japanese campaign to industrialize—and it should be seen as that—reflected the conviction of the leaders of the society that only by adopting Western technology could the nation maintain its independence vis-à-vis Western encroachment.

The success of modern technology in Japan owed much to government expenditure on social overhead capital (especially transport and education), to selective and moderate tariff protection, and to official support for those branches of industry that called for heavy investment and were slow to generate profit. This program, in turn, depended on gains in productivity and a growing surplus that could be converted into tax revenues. Advances in agriculture were crucial, with output rising 1.93 percent per year from 1885 to 1917. The first major industrial gains were characteristically in light industry—

Adapted from H.C. Darby and Harold Fullard (eds.), *The New Cambridge Modern History Atlas*, vol. 14 (© 1970), Cambridge University Press; maps © 1970 by George Philip & Son, Ltd.



Population changes in Europe, 1870–1925.



Production of coal and pig iron and ferroalloys in Europe, 1870–1970.
1870 data from H. Darby and H. Fullard (eds.), *The New Cambridge Modern History Atlas*, vol. 14 (1970); Cambridge University Press

silk textiles to begin with but, very quickly, cotton. From 1880 to 1913, cotton spindles increased from 13,000 to 2,287,000, or 176 times. By the latter date, the Japanese export of cotton yarn and thread was worth twice that of Germany and about 40 percent of that of the United Kingdom. Heavy industry, by contrast, was still in a state of tutelage, dependent on military orders and artificial stimuli; the payoff in technical training, managerial experience, and labour skills would come later.

The changing balance of economic power. Within Europe, these decades brought an important realignment of economic power as the first follower nations (Germany, France, Belgium) completed their industrial revolutions and new entrants (Sweden, Italy, Holland, Hungary) made their appearance. Britain, the first industrial nation, continued to develop and grow, though more slowly than its rivals. By the 1890s Germany had passed Britain in iron and steel manufacture, and, from then on, the gap between the two countries widened steadily. The result was a shift in the balance of political power that generated international anxiety and conflict and contributed to the outbreak of World War I.

The relative decline of Britain

The reasons for the relative decline of Britain have been a subject of extensive controversy. Some have tried to minimize the weakness of British performance, but this is unpersuasive, since most of the major economic variables give evidence either of a slowing down or of smaller gains than elsewhere, especially in the newer branches, such as electric power, electrical engineering, and organic chemicals. What is more, British technique was now lagging, even in industries such as steel, in which the key innovations were of British origin.

It has been argued that British technology, for all its shortcomings, represented a rational, optimal response to the relative costs of the factors of production; that is, that, if British productivity in certain industries was lower than that of competitors, this was unavoidable, implicit, among other things, in Britain's deteriorating natural endowment for industry. Britain, on this interpretation, was no longer able to compete. If true, this is a discouraging judgment and one that would no doubt have deterred such raw-materials-poor countries as Japan from ever entering the international market. It is far more plausible to see Britain's relative decline as the result of

economic reverses and human shortcomings than as inexorable destiny.

There were, first, the consequences of industrial proliferation: where once almost the whole world had been open to British goods, now the home markets of manufacturing rivals were closing through tariff restrictions (Germany, 1879; France, 1892; United States, 1890 and 1897). Competition in third markets was intense. The cotton industry furnishes the best example: from 1885 to 1913, British sales of yarn to China fell from 20,000,000 to 2,000,000 pounds, while India and Japan were selling hundreds of millions of pounds between them.

Second, Britain found itself handicapped by its priority in industrialization. To go first is both an advantage (entrepreneurial profits, accumulation of capital) and a disadvantage (costly mistakes, expenses of experimentation, related costs). The British, who had built much of their industrial plant and transportation system to the requirements of a smaller economy, now found that younger countries, building to new specifications, were better able to adopt more efficient techniques and operate on a larger scale.

Even when the British built a new plant, however, it was usually not so efficient as in Germany; and, in such areas as electrical engineering and organic chemicals, the British were not even building plants to the extent that market demand and natural-resource endowment warranted. In this case the explanation seems to lie in social values and institutions. The newer branches of industry, especially, and the older branches, to an increasing degree, demanded a high level of scientific training and a willingness to invest in research and development with a delayed payoff. The day of the inspired tinkerer was past. In this circumstance, the Germans were at a significant advantage, for their educational system was far ahead of the British in the production of graduates in science and applied science; and the society conferred a value on these activities that was wanting in Britain. To some extent, Britain compensated for this knowledge gap by importing continental technicians and entrepreneurs—a significant reversal of the earlier pattern; yet, this was, at best, a palliative. The difficulty was compounded by the great emphasis on experience and seniority in British works, by practical men's suspicions of "book learning," and by the scarcity of ambitious, tradition-free management. Too many British entrepreneurs were heirs of earlier achievers, industrialists by inheritance and duty and more concerned with spending than getting.

The "great depression" of the 19th century. The pressure of international competition was felt most keenly in Great Britain, which, as the dominant industrial power, had most to lose. The sense of loss was the stronger because these were decades of slow growth of manufactures (1.8 percent a year), declining prices (the wholesale index for manufactures fell from 127 in 1873 to 71 in 1894), protracted slumps, and small profits—hence, the traditional designation "the great depression."

Business cycles

Some scholars have viewed the great depression as a European phenomenon, but the course of economic activity in the continental countries was quite different from what it was in Britain. German industry did suffer a severe setback after the crash of 1873, a typical catharsis after some three years of enthusiastic company promotion (the *Gründerjahre*) fuelled by the large French war indemnity. Beginning in the 1880s, however, rapid growth resumed, and profit levels picked up, with no further serious interruptions until the crisis of 1900–02. Over the period 1873–1913, mining output rose 4.3 percent per year; manufacturing output, 3.4 percent. France, by way of contrast, had much less trouble in the 1870s, was hard hit by a financial crisis in 1882, had another, smaller shock in the mid-1890s, and then another, more protracted, in the first years of the new century. On the whole, these were years of below-average growth for manufactures (2.3 percent a year between 1873 and 1895). In the United States, the troughs of industrial activity came about every ten years, with especially severe conditions in the 1870s and 1890s. For the U.S., however, as for Germany, these were decades of sustained growth of manufacturing output (4.9 percent per year 1869–1913) and could in no way be described as depressed.

Even so, it would be a mistake to place too much reliance on indexes of aggregate production, which of their nature soften the pattern of economic fluctuations. The periodic depressions that were afflicting the economies of all advanced countries were real enough to their victims. The near regularity of these contractions gave them an appearance of inevitability, and economists bent their efforts to understanding them and predicting their timing. Much of the difficulty was linked to the increasing commercial and financial interdependence of nations. The world volume of trade had grown spectacularly (three times between 1850 and 1880, three times again from 1880 to 1913), in spite of the late-19th-century tendency to higher tariff protection. Trade entailed credit, so that every major commercial centre held large amounts of commercial paper, secured in principle by real commodity transactions but often resting on personal arrangements (accommodation paper) and mutual confidence. With the development of the limited-liability corporation and large-scale trading in company shares and debentures, a new field of speculation and credit opened up, the fortunes of which were closely linked with those of the commercial sector. The potentialities for enrichment were much enhanced, but so were the dangers of overoptimism and the penalties of miscalculation. Typically, the upswing of the cycle was marked by a rapid rise in the volume of trade and investment, with a concomitant increase in commodity prices (including the rate of interest) and a growing reliance on credit. When demand, however, was no longer able or willing to meet these prices, orders would fall off; stocks would accumulate; bills of exchange and other instruments of credit would not be paid; and a downward spiral of diminishing confidence would produce a general contraction. This process could not be limited to one area or country; all its trading partners would necessarily suffer in turn.

The international monetary system

The communicability of these fluctuations was enhanced by the workings of the international monetary system, which rested on the gold standard and generally free convertibility. The primary concern of the central banks was to protect specie and bullion reserves, thereby ensuring continued convertibility. In time of contraction, however, the calls for credit drew heavily on these reserves, for, when confidence failed, only hard metal would serve as payment. To counter this drain, the central banks tightened credit and the money supply when they were most needed and loosened them when the emergency had passed. By raising interest rates in time of crisis, moreover, each money market strained the resources of the others by siphoning off funds. To be sure, if matters became desperate, one central bank would rescue another by direct loan. More often, however, it was *sauve qui peut*, with each country defending its reserves at the expense of its own citizens.

The international gold standard depended, in the last analysis, on the maintenance of a stable balance of payments among the economic powers. To some extent, convertibility conduced to that result by imposing restraints on all: any nation that bought more than it could afford would see the rates of exchange turn against it, lose reserves, and have to stop overspending. Yet, the tacit acceptance of this discipline would never have survived any real imbalance of international exchange. Even in the 19th century, those countries, such as Austria and Russia (or even Britain during the Napoleonic Wars and the United States during and after the Civil War) that could not back their paper currency with sufficient metal simply eschewed convertibility; and, with World War I, the gold standard became a matter of history and, for some, of ideological adoration. There were several attempts to restore it subsequently, but none succeeded, for the simple reason that no sovereign nation was prepared to accept the self-privation that it implied.

A second Industrial Revolution. Insofar as the Industrial Revolution spread during these decades to newly industrializing countries, its outlines were substantially the

same as they had been earlier; that is, the first advances typically took place in food processing and textiles (cotton especially), with further extension into metallurgy, machine building, and chemicals. Within each of these branches, the techniques and equipment introduced were largely the classic elements: spinning machines and power looms, coke blast furnaces, and rolling mills. But the equipment itself had grown bigger, faster, more accurate and dependable, while the array of techniques had vastly proliferated, so that newly industrializing nations adopted, alongside these classic items, such newer ones as the steam hammer (an invention of the 1840s), the sewing machine, the milling cutter (for metal as well as wood), and the new technology of steelmaking. As a result of these changes, the threshold scale of production had grown considerably, and each piece of equipment cost substantially more. To be sure, it was sometimes possible, and sensible, to start new enterprises with the second-hand machines and equipment of more advanced economies. More often, however, one took what machine manufacturers and industrial engineers were prepared to deliver, and few of them were ready to build for their customers outmoded plants and equipment. In short, though the lineaments of these later industrial revolutions were the same as earlier, the content was quite different.

In the meantime, the technology of the advanced nations continued to evolve in the same direction, although the gains achieved in the older branches were shrinking relative to previous advances. The spindles could turn a little faster; the furnaces could be built higher; but the increment entailed a smaller proportional increase in profit than the radical changes of an earlier era. The place of the old staple industries was shrinking, and the rate of growth of a given economy depended on its ability to develop newer lines of production.

These constituted, toward the end of the 19th century, an innovational cluster important enough to warrant perhaps the designation of a second Industrial Revolution. As before, the critical advances occurred in the fundamental sphere of energy conversion and distribution: on the one hand, the production and use of electricity as a source of motive power and, on the other, the invention of the internal-combustion engine as a prime mover. The first had revolutionary consequences both for industrial technique and domestic life. Not only was electrical transmission of energy more efficient and safer than conventional direct drive by shafts and belts but it also could be effected over long distances, so that the production of energy could be centralized and power distributed to small units (households and shops, as well as mills and forges) as needed. Electrical current, moreover, was far more versatile than steam power, for it could be used not only to produce motion but also light, sound, and heat and made possible a revolution in communication (telegraph, telephone, radio, television, radar, telex, etc.), as well as some notable advances in chemical manufacture. By the end of the century, almost two-thirds of U.S. factories were using electrical power equipment to some degree; by 1914, three-quarters. Whereas steam power was necessarily concentrated in large production units, hence, in a few power-intensive industries, electric motors were useful and profitable in a much wider range of activities; thus, in 1889, the five leading power-using branches (lumber, food, iron and steel, textiles, paper) held 73.5 percent of total primary-power capacity in manufacturing; by 1929 this share had fallen to 58 percent. In Europe, Germany took the lead both in production and use of electrical power and the production of electrical equipment. German industry was marked by thorough exploitation of by-product energy produced by heat-generating industries, such as metallurgy, large distribution nets with high-use factors, and early standardization of current characteristics.

Electrical power

The second major innovation in the generation of power was the internal-combustion motor, so called because the driving force is obtained by controlled explosions within the motor itself. The first such device was Étienne Lenoir's gas engine of 1859, but the critical conceptual advance came with Alphonse Beau de Rochas' four-stroke cycle, which found application in Nikolaus August Otto's "silent" gas engine of 1876. This was a tremendous success from the outset, for it was far more efficient than the steam engine when working intermittently or at less than full load, conditions typical of small industry. Its principal shortcoming was that it was tied to its source of fuel, whereas the promise of the internal-combustion motor lay in its compactness and potential mobility. These possibilities were realized with the adoption of oil fuel. The result was a revolution in transportation (marine engines, the airplane, and, above all, the automobile), which liberated land transport from rail routes and the individual traveller from dependence on common carriers. The impact of the automobile was only beginning to be felt in the period before World War I; but the introduction of the cheap passenger car (Henry Ford's Model T, from 1908) was a promise (or warning) of things to come. The automobile quickly became the most important single product of the manufacturing industry, generating investment and growth by its demand for raw materials and highways (including bridges and tunnels) to match the burgeoning volume of traffic. Output of petroleum products, rubber, cement, asphalt, steel, glass, and even textiles (for upholstery) waxed (and sometimes waned) in sympathy with the demand for and use of automotive vehicles of all sorts, to say nothing of a large service sector (insurance, law and litigation, medical repair, auto repair, garages, and parking lots) devoted partially or exclusively to meet the needs and correcting the misdeeds of auto users. At the same time, the process of manufacturing automobiles called for innovations in materials and techniques important in themselves and susceptible of application in other industries: line assembly, new alloys, special-purpose tools, advances in grinding and stamping, and standardization of parts. By 1913 there were 1,258,000 motor vehicles registered in the United States, 208,000 in Great Britain, 125,000 in France, 93,000 in Germany—a total of 1,700,000. Twenty-five years later, on the eve of World War II, the total for those four countries had reached almost 36,000,000, a gain of over 21 times (13 percent per year).

Internal-combustion engine

As important as the automobile industry came to be, two other new branches of manufacture can be compared to it in range of influence and potential for growth. One was organic chemicals and the manufacture of synthetics. Again, the pre-World War I years gave just a hint of what was to come: celluloid (John Wesley Hyatt, 1868); artificial silk or rayon (Hilaire, comte de Chardonnet, 1889–90); bakelite, the first synthetic resin (Leo Hendrik Baekeland, 1909); cellophane (J.E. Brandenberger, 1912). These artificial fibres and plastics were the first fruits of what would be a new kind of industry, based on increasingly heavy investment in laboratory research and turning out new products to order, primarily for consumer use.

The second was the equally versatile electronics industry, in its infancy in the early years of this century, when its only application was still wireless communication. The key innovation here was the tube or valve, so called because it channelled the flow of electric current in one direction and made possible the construction of complex circuits. Synthetics and electronics did not fully develop until the interwar years and the post-World War II period and might better be seen as part of a new major cluster of innovations (including atomic energy and computers) constituting a third Industrial Revolution.

Along with these and other innovations in technique and product went new patterns of industrial and business organization. The old trend to bigness continued as better but costlier production equipment raised the price of entry and offered new economies of scale, while improvements in communication enhanced the possibilities of centralized management. In the United States, capitalization per manufacturing establishment increased from $4,334 in 1849 to $10,992 in 1879 to $153,278 in 1919. Economies of scale—in marketing and administration as well as production—were only part of the story, however. Equally important were market considerations, in particular, the desire to stabilize profits and minimize competition, either by absorbing or eliminating competitors

The trend toward bigness

or by organizing cartels, trusts, and similar combinations in restraint of trade. In Europe, the merger and cartel movements were especially strong in Germany, first, because the close ties of German manufacturing with investment banks allowed the latter to play the broker's role in combinations (which usually generated substantial fees and promotional profits); second, because German managers were usually bureaucrats who identified with their functional role and took well to group discipline; third, because the German government was, on balance, favourable to developments that it viewed as promoting stability and control. By 1905 there were some 366 industrial cartels operating in Germany; some controlled prices, others output, many both. So powerful was the trend to bigness and combination and so influential the role of the so-called great banks (themselves the result of numerous mergers) that contemporary observers thought they saw the appearance of a new stage in economic development, the era of finance capitalism.

Finance capitalism

The term was somewhat premature. Although similar trends were to be found in the other major European industrial economies, they were substantially weaker. In both Britain and France, for example, the stress on family ownership and identification posed a barrier to both financial intervention in industrial management and to disciplined participation in cartel agreements. In Britain, when mergers did come, they were often a response to contraction rather than an act of expansion: independent firms reluctantly combined their resources to meet falling prices and profits and then spent years fighting over the appropriate allocation of authority and returns.

In the United States, giantism was to a certain extent the expression of a national style, an aspiration to bigness on the part of industrial tycoons, such as Andrew Carnegie in steel, the Guggenheims in mining, John D. Rockefeller in oil, the brothers James and Benjamin Duke in tobacco, Philip Danforth Armour and Gustavus Swift in meatpacking, Edward Henry Harriman in railroads. Financial interests also played their role: Kuhn, Loeb & Co. in railroads and, most prominently, J.P. Morgan and Company, the bank that founded the General Electric Company, the International Harvester Company, and the United States Steel Corporation—to cite its three largest creations. The usual technique in these promotions was to "water" the stock—that is, to issue shares for an amount far beyond the real assets and cost of the component companies. The difference, which is said to have been as high as $62,500,000 for U.S. Steel, went into the promoters' pockets. Not that the stock was not worth its watered value: if the amalgamation was well planned and yielded sufficient control of the market and (more rarely) savings in costs, the new combine might well make enough money to justify the price of its shares; in most instances, thus, it was the consumer who paid for the "water." By the Sherman Anti-Trust Act of 1890, combinations in restraint of trade were prohibited. Nevertheless, it took decades of litigation to dissolve the largest trusts, and, even then, the giant corporations that remained found ways to cooperate informally.

Industrialization on the eve of World War I. The position of the different nations in this process of social as well as economic transformation may be measured by the position of industry relative to agriculture and the ratio of urban to rural population. In regard to the first, one can discriminate among four categories.

First, there were countries in which the industrial labour force was already greater than that in agriculture: Great Britain, which had reached this position by 1820, is the best example, but Belgium had pushed its industrialization almost as far. The United States and Germany, in spite of substantial advantages in agriculture, had long since entered this group. France had not quite made it and would not do so until after World War II. Each of these countries had increased product per head by three to five times over the course of the preceding century (see Table 3).

Second, there were countries whose agricultural population was still about twice as great as its industrial: Sweden, Italy, and Austria all fell into this category. Each nation had developed by this time a large, modern industrial sector, and Sweden and the Czech lands, particularly, had enterprises that were technologically at a level with the best European practice.

Table 3: Growth of National Product, Selected Countries and Periods (Constant Prices)

	rate of growth per decade (percent)		coefficient of multiplication in a century	
	total product	product per head	total product	product per head
England and Wales				
1700 to 1780	5.3	2.0	1.7	1.2
1780 to 1881	28.2	13.4	12.0	3.5
France				
1831–40 to 1861–70	26.3	21.6	10.3	7.1
1861–70 to 1891–1900	15.7	13.5	4.3	3.6
Germany				
1850–59 to 1910–13	29.2	16.3	12.9	4.5
United States				
1839–43 to 1910–14	48.2	16.8	58.9	4.9

Source: S. Kuznets, *Modern Economic Growth* (1966); *Economic Growth of Nations* (1971).

Third, there were countries that had entered on the Industrial Revolution but were still primarily pre-industrial. Russia was a good case in point, with a few completely modern plants, especially in the young iron-and-steel industry of the Donets Basin. Hungary also had, in the Ganz electrical works of Budapest, one of the more progressive enterprises in Europe. But these were predominantly peasant societies; skilled industrial labour was scarce; and industry was still dependent on strong tariff protection and artificial support.

Fourth, there were countries almost exclusively dependent on agriculture: such Balkan nations as Greece and Bulgaria, the colonial lands of Asia and Africa, and the more backward countries of Latin America—in short, what would later be called the Third World. With rare exceptions (*e.g.*, India's cotton manufacture), industrial production consisted almost entirely of domestic handicrafts and artisanal work.

Concomitant with the shift of resources to industry went a movement of population to urban centres. The development of a predominantly urban society entailed substantial costs to physical and psychic health, while the convenience of concentration was offset by the impediments of high density of living and movement. Even so, urbanization contributed to the increase in productivity and wealth by intensifying contacts and information flows between innovators in knowledge and its applications and by nourishing and propagating new attitudes toward family size and population control.

This last was of the greatest consequence, because it reinforced what was already a wide gap between the industrial nations and the rest of the world. By the beginning of the 20th century, the very size of the gap between industrial and pre-industrial nations constituted a barrier to development by the latter, while the differential accessibility of the new technology made their task harder by bestowing on them those benefits that translated into lower mortality rates rather than higher output. Just when the industrial nations were completing their "demographic transition" from high death rate cum high birth rate to low death rate cum low birth rate and winding up their population explosion, the undeveloped parts of the world were entering on theirs.

In sum, the Industrial Revolution, though consisting in a cluster of changes in technique in a small part of the productive system, ramified throughout that system and, transforming it, transformed the quality of life and the values of society; starting in one country, it proliferated throughout the world—a far more ecumenical movement than even the most successful religious faiths. It did so unevenly, for better or worse. Those countries that have become rich in the process (though they have pockets of poverty) are now having second thoughts; technology is much decried as an enemy of man. The poorer nations cannot afford the luxury of renunciation; they have noth-

ing to renounce. They are all, or almost all, striving to effect their own industrial revolutions. (D.S.La.)

III. Economic developments since World War I

THE AFTERMATH OF WORLD WAR I

The outbreak of war in 1914 ended a distinct era in economic history and ushered in a period of instability, uncertainty, and unexampled change that could hardly have been foreseen before the conflict began. Trends that might have been gradual were telescoped into a few years. A further impetus was given to the concentration of production in large-scale units and to the decline of competition. State intervention in the economy, prominent for the first time under modern conditions during the war, was to assume a more definite and permanent character. Technological change became more firmly linked to the giant advances in the physical sciences and permanently molded life in the advanced regions of the world.

To write the economic history of this great transformation poses many problems, for, despite statistical and literary evidence on a scale unprecedented for any previous period, it is no easier for historians to agree, the very nearness of the phenomena being observed provoking controversy rather than consensus. It is, moreover, inadequate to consider only the advanced countries of western Europe and North America when observing the development of the world economy. New light may be cast on the economic past of the now advanced countries by considering the history of the developing countries.

Only the most technologically and industrially advanced nations could wage wars of the type of 1914–18 and 1939–45. Germany's ability to hold out so long was a tribute to its industrial strength as much as to its military competence in both wars. The decisive role of the United States as an arsenal and its rapid industrial growth in both wars is again highly significant. France held out in World War I thanks to its allies' economic support; its collapse in World War II was economic as well as military. With or without the atomic bombs, Japan would have been defeated because of its overstretched resources.

In this period, economic history cannot be separated from military history, and into both the major wars there entered deep economic rivalries. War also provided an overwhelming stimulus to maximize production and make full use of scientific knowledge. Mechanization, rationalization, and standardization were imposed by the conditions of war production, and the interruption of normal trade channels intensified the search for substitute materials and the development of synthetics, from which are descended some of the fast growing industries of the period. The "new industries" early in the century expanded rapidly. In the United States as well as Europe new plants had to be erected both in existing industrial areas and in areas little touched by industrialization.

The conditions of war production

The increase of government control. Once they realized that modern economies could engage in long wars, the belligerents had to intervene rigorously in the economic arena. Market forces were superseded or supplemented by state direction and coordination. The overriding aim became the equipping of the armed forces, and, since their needs rapidly became voracious and unlimited, output at any price became the watchword and contracts were placed on a cost-plus basis to encourage maximum production. Trade unions and employers' associations were drawn in, and controls had to be extended to the distribution of necessities to the home front. The overall problem was one of acute shortage and the substitution of administered priorities for free market forces.

This greatly affected politicians and administrators accustomed and disposed to an automatically operating market economy. Not even in Germany had a comprehensive apparatus of controls been envisaged at mobilization. It was assumed that, when the fighting ended, a speedy return to the market would be possible. From 1918 the state relaxed its hold on the economy—prices surged forward in a burst of inflation, and a short postwar boom began, to end in mid-1920.

Postwar reconstruction. Destruction resulting from the war on land was chiefly confined to the main battle zones in Belgium and France. The victorious Allies expected that Germany would pay for this destruction, but this turned out to be a vain hope: reparations caused bitter controversy in the coming years and animosity between the victor powers. What was, however, of greater significance was the ability of the advanced industrial countries rapidly to reconvert their industries and rebuild devastated areas. French reconstruction endowed the northern and eastern *départements* with a new infrastructure and re-equipped and modernized agriculture and industry within a few years. Likewise, German industry recovered rapidly.

The diminution of European power. It was the world economy that experienced the most profound changes. The most outstanding was the financial and industrial upsurge of the United States and the decline of Europe, transformed from a creditor to a debtor. The world monetary and trading system centred on Europe had been seriously dislocated, and frontier changes had increased the tariff barriers dividing the Continent; and there was political uncertainty in central and eastern Europe. The Bolshevik Revolution had taken Russia out of world capitalism and threatened to spread westward in the postwar atmosphere of discontent and near chaos. Finally, industrialization had gone ahead in other parts of the world, closing old markets and rearing up new competitors. Most ominous of all was the emergence of Japan as a vigorous competitor in world markets.

Of major and immediate significance in the postwar decade was the need to restore the old pattern of trade and payments. This was taken to mean, first, the restoration of the gold standard upon which prewar currency stability and convertibility had been based. But war finance had everywhere caused inflation in the belligerent countries, and there could be no return to gold without currency stability. Second, Europe was quite unable to resume its prewar position as an international creditor. U.S. capital played a vital part in European reconstruction and subscribed heavily to the international loans that enabled Germany to make scaled down reparations payments. Stabilization and expansion in the 1920s was thus underpinned by an outflow of funds from the United States to Europe. This proved to be a most unstable arrangement. Funds returned to take advantage of the capital gains realizable on Wall Street and when the crash came in 1929 one of the main props of the central Europe banking structure was removed, and the stage was set for the financial collapse of 1931.

The role of U.S. capital

Inflation. War finance had been inherently inflationary, and, once controls were removed, prices began to soar. Moreover, social breakdown in eastern Europe and Russia had resulted in money printing and inflation. The occupation of the Ruhr by French troops in 1923 in an effort to enforce reparations payments precipitated a hyperinflation in Germany that broke all records. In the final stages Germany resorted to the printing press on a massive scale, and the rise in prices was such that workers collected their wages in suitcases and spent them at once before the paper marks became worthless. In November 1923 a determined and successful effort was made to stabilize the currency, and Germany entered a new period. Industry had meanwhile wiped off its debts and made considerable windfall profits. Thanks to this and to the inflow of American funds, industry was modernized and expanded. Inflation had disastrous effects for those who held their savings in state bonds or fixed interest securities or who lived on fixed incomes. At the same time, speculators and profiteers did well, and some made fortunes overnight. The "new rich" of the inflationary years prospered; their social juxtaposition with impoverished middle class people helps to explain the political upheavals that followed. The outstanding feature of the inflation was its arbitrary effect: undermining the old virtues of thrift and savings, which had been capitalism's most valuable moral supports.

Countermeasures. As governments sought to control inflation they prepared to return their currencies to a gold base. This was done with a greater measure of international consultation than hitherto. Britain returned to gold

—at the prewar parity—in 1925 in an effort to "look the dollar in the face" and permit the City of London to resume its prewar financial pre-eminence. It required a conscious effort to deflate prices and incomes, and the effort ran into stiff working class resistance. Its effect on exports was highly adverse, and it was an important factor behind the General Strike of 1926. The French, with the franc only four-fifths of its prewar value, were unable to return to gold at the old parity. When the franc was stabilized in 1926 and the move was legalized in 1928, to avoid unemployment it was slightly undervalued, which favoured exports and the tourist industry.

In the 1920s something like currency stability and convertibility were restored on the basis of a modified gold standard. Reconstruction was completed in war-damaged areas; industries were restored and modernized; old trade links were re-established and new ones opened. Industrial expansion was resumed. American capital flowed into Europe, and by 1930 a number of countries were enjoying boom conditions. But there were some shadows. The British economy was suffering from the loss of export markets. The traditional industries were ailing, and the overvalued pound had harmed the export trade. Germany, like Britain, had a seemingly intractable unemployment problem with almost 2,000,000 persons jobless even before the slump.

The changing world picture. The primary producing countries had boomed during and immediately after the war and had increased their capacity. The recession that began in 1921 after the postwar restocking boom was a sign that things had changed. Most of the producing countries were in a colonial or dependent relationship to the advanced countries, and in a number of them a nationalist movement had appeared on the scene. Where some industrial growth had taken place, as in India, it called for protection of the home market and government encouragement of industrialization. The seeds of change were being sown. Some economists have linked Europe's interwar difficulties with the disappointing performance of primary producers. Had the European countries adapted to changes in technology and world demand, they would have increased their demand for primary products and enabled their producers to buy more manufactured goods. As it was, world trade was the worst sufferer from the war and the biggest casualty in the Depression of the 1930s. To make matters worse, the industrial countries protected their own producers, and, when the slump came, there was a general retreat to the home market with higher tariff barriers, quotas, and administrative devices to keep imports down. Primary producers responded by encouraging their own industries as well as keeping up prices by restriction schemes.

Nationalist movements

Outside Europe, Japan, the only Asian country before 1914 to have begun to lay the foundations for an advanced industrial economy, experienced rapid economic growth in the 1920s. Japanese businessmen seized the opportunity afforded them by the cutting off of European exports to Asian markets, and, despite the recession of 1921–22 and the Tokyo earthquake of 1923, the industrial sector expanded rapidly. The economy was dominated by a few giant, monopolistic firms peculiar to Japan, but much industry continued to be organized in the old forms in workshops or on the domestic system. Wages were low and labour abundant. Creating technologically advanced industries necessary to carry on the war with China was typical of the 1930s in Japan. The economy depended intimately, however, upon foreign trade. In the 1920s it had benefitted from the demand for silk, and Japanese industry was later able to undercut its competitors in textiles and cheap consumer goods.

Technological and industrial progress. The introduction of industries based on new technologies was to have profound effects on the relative economic development of the industrial countries and particular regions in each country. During the 1920s the United States took the lead in the new fields. It was favoured with vast natural resources; a large, varied, skilled, and educated labour force; a large and expanding internal market; and a high average income. Such conditions encouraged production of standardized articles on a large scale and continuous improvement in technology to keep down labour costs and to take advantage of expanding markets. The accumulation of capital and its concentration into giant business units had proceeded rapidly from the latter part of the 19th century and took a further giant leap as a result of the war. U.S. productivity was the highest in the world and that country's system of manufacturing became a model. U.S. industry pressed into new fields opened up by technological advance, and a whole range of machine-made products passed into general consumption. This transformation was seen in the widespread use of the automobile, the telephone, the consumption of electricity, and the kitchen equipped with electrical appliances. The new consumer durables, symbols of modern living, were most fully developed and most widely used in the United States. Correspondingly, innovations of no less importance appeared in complementary fields: merchandising, advertising, packaging, retail selling, and consumer credit. The United States mirrored the prosperous future of the other advanced countries, although the full flood of Americanization was delayed until after World War II.

The U.S. in the 1920s

Compared with urban U.S., even the most highly developed areas of Europe seemed archaic. Their incomplete industrialization had left in existence parts of the traditional culture that moderated the operation of market forces. In Europe, a successful businessman might retire early or endeavour to pass his children into the ranks of the still highly regarded leisure class. Most European countries except Britain still had a large peasant sector, with which was associated many activities still in an artisan stage, and a cluster of petty capitalist activities in the small towns that dotted the rural landscape. The structure of agriculture, the distribution of income, the whole character of the market did not yet favour the mass production typical of the 20th century. The main emphasis of European industry in the 1920s was on textiles, iron and steel, coal mining, heavy engineering, and shipbuilding. Economic historians have pointed to the slow shift of resources from the old declining sectors to industries based on the new techniques of the kind that were growing in the United States as being a fundamental weakness of Europe's interwar economy. This drag on Europe's growth also limited the demand for raw materials and foodstuffs from the less developed countries.

The slow shift of resources

Market forces seemed powerless to effect the kind of transformation needed, and state intervention often merely confirmed acquired positions. In any case, the conditions favouring the growth of the new industries in the United States could not be produced in other environments. A further shift of resources out of agriculture and toward urban living was necessary. Only a general rise in incomes could open a mass market for such new products as automobiles and domestic appliances. New systems of distribution and consumer finance had to be adopted, and capital had to be attracted by the prospect of profits. These changes came about slowly as output recovered and began to expand with increased productivity; but the obstacles were not simply within the range of economic policy: they were more broadly sociological and cultural. Changes in social structures and life-styles, accelerated by World War II and helped by time passing, were vital and perhaps decisive.

The beginnings of social welfare. The economic experience of the war years did not shake the general assumption that, in peacetime conditions, state intervention was unnecessary and harmful. But it had been shown that in certain circumstances business might benefit from the activities of the state. Only in the United States was belief in the virtues of rugged individualism still unchallenged, and even there it was more myth than reality. Until the Depression of the 1930s, however, there was no desire, even in Europe, to see the state permanently endowed with more extensive economic powers. It was not the success of state intervention in the war but the breakdown of market forces in the 1930s that opened the way for regulated capitalism.

The extent of state intervention varied greatly; some countries had operated the railway system for many

years, and there were state monopolies for fiscal purposes. But the main economic intervention continued to be the tariff. Even Britain had imposed protective tariffs on luxury items during World War I and in the 1920s began "safeguarding" industries from unfair foreign competition. At the same time, there was a steady underlying trend toward greater state involvement in the economy, influenced by the growing Socialist and trade union movements. The way was being prepared for the wider intervention by the state characteristic of the Depression decade.

The use of tariffs

State provision for social welfare had a much wider measure of acceptance in the 1920s and steadily became more systematic. It was understood that the sanctity of private property and the pursuit of profit might be threatened if society did not make adequate provision for its casualties. By the 1920s most states had social security or insurance systems of various degrees of comprehensiveness. Increasingly, such welfare provision became a matter of legislation laying down minimum standards rather than the old haphazard mixture of charity, poor relief, and mutual insurance. Wartime experience, in which all members of a society were taken to be in solidarity against a common enemy and thus entitled to its support, exercised an important influence on the development of social security systems.

The state's tendency to assume social responsibilities was persistent and continuous in the period after World War I, and, with the establishment of the International Labour Office in Geneva, a beginning was made in establishing uniform standards between countries. Partly under labour union pressure, industrial regulation became more comprehensive and better enforced. Public and local authorities were also assuming greater responsibility in transport and the provision of water, gas, and electricity; in education; and in housing, made necessary by the chronic and widespread postwar shortage. Investment in human capital, if not yet a common phrase, was being taken more seriously than ever before; but, like all similar trends, it was uneven; practice often lagged behind legislation, especially in the poorer countries.

Paradoxically, the richest country of all lagged far behind. In the United States, where the new technologies were most profoundly altering social life, where business structures and patterns of consumption earliest assumed the characteristics of modern capitalism, social security was neglected until the Depression. Apart from some accepted regulatory functions, neither the federal nor the state governments did much in the social and economic arenas. The whole American tradition—and the majority of public opinion, including organized labour—upheld the principles of laissez-faire liberalism. The most advanced capitalist country lacked a system of social security that every advanced industrial country in Europe took for granted, and it was also without an independent political party of the working class.

During the 1920s, as recovery and inflation took place, production and productivity again began to grow in the advanced countries, and incomes rose. Meanwhile, the new technologies began to influence social life more fully. Road transport made its effects felt through the bus and the truck; rural areas were brought into closer contact with urban influences; and personal and labour mobility increased. Marketing and distribution changed as farm produce could be collected more regularly and moved more quickly to urban markets. Electric power brought new flexibility to industry and opened an industrial vocation to hitherto predominantly agrarian areas. The cinema and the radio transformed entertainment patterns as well as the supply of information or propaganda. Professional sports became a growing industry. Service trades of all kinds grew most rapidly in the advanced countries with high incomes.

The new technologies

This expansion of the "tertiary sector" was one of the most notable trends of this period. Consumers, with increased incomes, spent more on services. Former household activities became specialized marketed services; retail outlets multiplied; old-style services expanded, and new ones appeared. As the activities of governments, national and local, became more complex, so their employees increased; and even business employed a growing bureaucracy of its own. In other words, the white collar element in the labour market grew and was symbolic of changes in life-style in societies with a relatively high and growing per capita income. Expectations grew, too, even faster than possibilities for their realization. Workers became better organized and more demanding, installment selling enabled people to spend ahead of their incomes, and the Depression caught all these developments in full flood.

THE GREAT DEPRESSION

The stabilization of capitalism after World War I was short-lived, for the world economy plunged into the worst depression in its history. Although heralded by the Wall Street crash in the autumn of 1929, it was some years before it became apparent that this was no ordinary cyclical downturn. It afflicted every capitalist country, and its depth and length still challenge those seeking its explanation. Even on the outbreak of World War II, despite the recovery from the trough of 1932–34, the Depression had not been conquered, and capitalism had not found new long-term forces of expansion sufficient to overcome the trends toward depression and stagnation. So far from new government policies for full employment and economic growth being effective, it seemed rather that purely national remedies, the resort to beggar-my-neighbour policies and, more especially, the drive to rearm—which arose out of the political upheavals of the Depression—prepared the way for the new war.

What does seem to be accepted is that, although depression was centred in the United States, it was the special and intimate relationships established between the U.S. and other capitalist economies after World War I that accounted for its rapid diffusion and harsh impact. It is equally apparent that the stabilization that had taken place in the other countries and the boom conditions of many, in the late 1920s, concealed serious weaknesses. The British economy's failure to resume capital exports in the old way, the decline of its basic industries, the consequences of the overvaluation of sterling—all weakened the world economy and thrust greater responsibilities upon the United States. The investment boom in European countries following reconstruction could only continue so long as markets could be found; a cyclical downturn was inevitable sooner or later. The problems of the primary producers (now weak sellers as a result of overproduction) reduced the market for manufactured goods and slowed their growth rate.

In the return to gold, some currencies had been overvalued, others undervalued, and the Depression came before adjustments could be made. Moreover, international financial relations were no longer on prewar lines. War debts and reparations weakened capitalist countries, leaving the United States as the major creditor. The 1920s' boom in Europe had been financed by an outflow of U.S. funds, much of the investment being on a short-term basis and already being withdrawn in the last phase of Wall Street speculation before the crash. Any investigation of the causes of the Depression must consider the changed role of the United States as now the main support of prosperity in the capitalist world. Once the U.S. economy had slumped, there was no prospect of maintaining prosperity elsewhere.

The effects of Wall Street speculation

The collapse of the world market. The reversal of the dollar outflow exposed the deficiencies in the economies of the other capitalist countries. What was striking, above all, was the disintegration of the world market as each country sought its own salvation. The contraction of world trade was the most general expression of the crisis, hitting hardest the primary producers and those industries depending upon exports. Thus, outside the United States, the Depression was worst in Germany and Britain. The sharpest contraction was in the export trades and the capital goods industries, as well as in shipping and shipbuilding. Activities turned mainly toward the domestic market suffered less, and so "sheltered" and "unsheltered" industries were distinguished. The former, least affected by de-

pression, were also the easiest to help. The recovery of the export and capital goods industries was much more difficult to engineer, and in many cases permanent contraction was the remedy used.

For most countries, the more prosperous conditions had been before 1929 the sharper was the decline in the following years, but a large agrarian sector—as in France—could provide some shelter from the blizzard. Germany was the worst hit European country, and there, as in the United States, the Depression was a social and political calamity. In Britain, however, the abandonment of gold and the devaluation of sterling; the adoption of protection that stimulated home industry; a delayed development of the technologically more advanced industries; a building boom assisted by cheap money; and the ability to take advantage of the low prices of imported foodstuffs and raw materials while using accumulated reserves to avoid a balance of payments problem—all enabled it to weather the Depression better than some other advanced industrial countries.

The United States was the home of the Depression, and it persisted there longest. The prolongation of the Depression is linked with its causes. Although the stock market crash had acted as a detonator, there were powerful forces in the U.S. economy that made the crisis more than a cyclical decline. Moreover, the collapse of the boom raised doubts about the long-term prospects, hitherto viewed optimistically. The Depression was thus traumatic because it contradicted deeply held beliefs in unlimited expansion and also brought poverty, hardship, and hopelessness to millions in the world's richest country.

That the United States was particularly slump prone was not a popular view in the 1920s, when the country seemed on a course of boundless expansion. Indeed, optimistic forecasts were made up to and even after the crash of 1929. U.S. economist John Kenneth Galbraith, in *The Great Crash, 1929* (1955), ascribed the Depression diffidently to five main factors: (1) unequal income distribution, which limited the consumer goods market and made investment and luxury spending erratic (bad news from Wall Street sent both down); (2) fraudulent business, which abounded in "a kind of flood tide of corporate larceny"; (3) the banking structure, which contained too many independent units, causing bankers to yield to overoptimism or to act foolishly—and once one bank failed the whole system collapsed; (4) the new creditor position of the United States, which, besides resulting in unwise foreign lending, was coupled with high tariffs so that foreign countries could not service their debts by increasing exports to the United States, resulting in default and the inability of foreigners to continue to pay for U.S. exports; and (5) the poor state of economic intelligence.

Five causative factors of the Depression

Government response. The historian must give due weight to the social consequences of economic fluctuations and the way in which governments, by their response to the situations thereby created, may complicate matters further. It was generally assumed by economists, at least well into the 1930s, that government intervention should be minimal, the assumption being that government intervention could not correct, and might aggravate, a market disequilibrium. Indeed, some felt the slump was partly a consequence of government interference. Even when governments did perforce begin to intervene, the orthodox majority continued to be suspicious of the trend or even to blame it for the economy's failure to recover more quickly and completely. Since the 1930s economic opinion has moved to the view that governments should have intervened with appropriate weapons to remedy what was seen as the basic cause of depression—a deficiency in demand.

What is clear is that the Depression had such severe social consequences that even reluctant governments could not stand aside. The price of nonintervention was social revolution, and intervention was, or could be, a weapon of counterrevolution, as under the Nazi regime in Germany. Intervention, therefore, took place pragmatically, and its forms were at first "orthodox" in that they consisted of old, well-tried methods. First, the tendency toward economic nationalism was reinforced as governments imposed tariffs or raised existing ones and added quotas and other devices to protect home producers. As the Depression continued, protection tended to be used increasingly aggressively to bargain with other countries and penetrate foreign markets. The disintegration of multilateralism, which was often replaced by bilateral trading agreements, accompanied the decline of world trade. The immediate effect of all these measures was to reduce the volume of international trade and create conditions of trade war.

Parallel to this went the breakdown of the gold standard. In the autumn of 1931 the overvalued pound sterling was driven off gold, starting a succession of devaluations as countries learned that an undervalued currency gave a temporary advantage in the struggle for markets. With their objective base for parity destroyed, currencies became subject to government action, and some form of exchange control had to be established if balance of payments difficulties were not to be exaggerated by speculative currency movements. Countries like France, which delayed establishing exchange control, did so to their cost. Nazi Germany, on the other hand, by means of rigid controls, could operate parallel exchange rates to its advantage and at the expense of weaker trading partners.

The conditions of recovery. The Great Depression of the 1930s shook capitalism to its foundations and was the major formative influence on political and social as well as economic life. Already undermined by World War I and its aftermath, many of the old certainties derived from 19th-century liberalism disappeared. Despite millions of unemployed throughout the decade—in many of the most advanced areas of the world—and the triumph of Fascism in Germany and the subsequent drive to war, most years of the 1930s were, however, years of recovery, though recovery from a very low level. Indeed, the slump itself had prepared the conditions for recovery through the normal cyclical process: after a few years, in which investment falls to a low level or even becomes negative; when stocks are used up; when equipment grows obsolete and has to be replaced; when prices reach bottom, the possibilities for profitable investment begin to improve. Unsound firms have disappeared, the financial excesses of the boom have been liquidated, a shift of resources from the unprofitable sectors has been affected, old plant has to be renewed, and confidence revives. Thus, the classic conditions for recovery began to operate independently of government action.

Recovery came slowly and in many countries was far from complete by World War II. Until there was a revival of world trade, the primary producers could not raise their incomes. On the other hand, in the advanced countries, even those as export dependent as Britain and Germany, it was possible to find a substitute for exports. In Germany this was chiefly war production. In Britain, although rearmament in the late 1930s helped recovery, it had earlier depended upon an expansion of the home market, the shift to new industries, and the continuous rise in service activities, even in the "depressed areas."

As a major importer of foodstuffs and raw materials, Britain was able to take advantage of the deterioration in the terms of trade of primary products. This enabled consumers to spend more of their income on nonfood items and expanded the consumer goods and services market, creating new opportunities for profitable investment and employment. It assisted manufacturers directly and consumers indirectly through the relative cheapening of imported raw materials. Given that this took place with a resumption of technological advance and innovation, Britain could enjoy a limited recovery. Almost all persons could thus improve their real income, at the expense, in part, of primary producers abroad. Where those primary producers were to a large extent at home, as in France, these conditions for recovery did not exist.

All governments realized that something had to be done to relieve what amounted to a national catastrophe. The Soviet Union's example had begun to turn thoughts toward planning, even of capitalism, as a possibility. In the early years of the Depression the instinctive reaction of

The orthodoxy of political reaction

politicians was still strongly orthodox. In Britain, proposals for positive steps to overcome unemployment through public works, were rejected by a Labour government. The watchword virtually everywhere was to keep the budget balanced and avoid interference with free market forces. Policy was thus generally deflationary: government expenditures were vigorously reduced, and the more consistently such remedies were applied, as in France, the more they aggravated the disease. In some cases, however, governments did inadvertently run budget deficits through emergency action, and where they did so they began, possibly unwittingly, to assist recovery. The real innovation came when what was done in some cases by accident was pursued systematically.

The German remedy. Two experiments are worth noting as of particular importance. Both had the character of models—to be emulated or to be avoided—and both have preoccupied economists and historians since, though a definitive account of neither has yet been written. One is the National Socialist policy adopted in Germany on Adolf Hitler's assumption of power in 1933, which prepared the most powerful industrial nation in Europe for a new war. The other is the New Deal, which Franklin D. Roosevelt proposed to the American people after his inauguration at almost the same time. These two experiments can also be considered as alternative ways of saving capitalism. Both shed light on the nature and efficacy of remedies for unemployment and depression that, during and after World War II, were to acquire immense prestige both among professional economists and the general public.

In Germany, the recovery and expansion of the 1920s were abruptly ended by the Depression, when American investment was cut off and short-term funds withdrawn, and when the collapse of world trade drastically reduced exports. Between 1929 and 1932 national income fell by 40 percent and unemployment soared to more than 6,000,000. All of the weaknesses of the German economy were exposed, particularly its unhealthy dependence on foreign loans and credits. As German governments followed each other in rapid succession, grappling halfheartedly with insurmountable problems, the National Socialist movement gained increasing support from the middle class and the peasantry. Blaming Germany's ills on the Versailles settlement and international Jewish finance and promising to save property owners from Bolshevism, Hitler's party came to power in January 1933 and implemented a new and energetic policy to save German capitalism. It is difficult to separate Nazi economic policy from its political and international policies, but there is little doubt that it was supported by the leaders of German big business and banking. It destroyed the free trade unions and political parties and abandoned the "socialist" elements in the National Socialist program.

In some respects Nazism continued old policies. The mark's exchange value was upheld, and much trade was conducted through bilateral bargaining. The state supported the depressed sections of the economy in consultation with business leaders. To relieve unemployment the Nazis initiated a policy of public works, including the famous *Autobahnen* ("expressways"). But such measures were but the prelude to much bigger rearmament expenditures that Hitler launched in defiance of the Treaty of Versailles. With state control of wages and, later, prices, unemployment was absorbed and idle resources used without significant inflation. The opening up of a new market for heavy industry in armaments production restored profitability and confidence in German capitalism but at the price of increasing militarization of the economy and to the accompaniment of an aggressive external policy that made war inevitable.

Foreign trade and exchange dealings controls were tightened to enable vital raw materials to be obtained on favourable terms. Exports had to be increased to pay for these imports. At the same time, militarism led to the development of new industries producing armaments irrespective of costs. Domestic consumption was held back by pegging wages at about their 1933 level. Prices rose rather more, despite controls.

It was no secret that German recovery was based on a war economy. All participants in World War II experienced a similar process: but the German experience was novel and, by some, held up as an example. The Nazis established a full employment economy but one geared to war and with war as its inevitable outcome. Nor did the Nazi leaders try to disguise their intentions, which they rationalized as a need to acquire for Germany more *Lebensraum* ("living space"). The Nazis did not establish a new economic system; they simply imposed greater state control on an economy that still rested upon private ownership and private profit. Industrially, however, it was a highly concentrated economy, and it was big business that benefitted most directly from Hitler's policies.

Hitler's quest for *Lebensraum*

The New Deal policy. The urgent situation created by the Depression in the United States and the blow inflicted on cherished ideas about rugged individualism imposed striking departures in that country's government policy. To say that they were imposed is to emphasize the fact that they did not arise from the application of an economic theory but were in the tradition of American pragmatism. Not until near the end of the 1930s did the doctrine of the economist J.M. Keynes begin to influence policy, and the measures that were taken until then fell considerably short of the Keynesian prescription.

When the Depression struck, a Republican administration under Pres. Herbert Hoover held office and met the catastrophe with puny relief measures to help farmers and the unemployed. Tariffs were raised in 1930, and, when these measures failed, Hoover set up the Reconstruction Finance Corporation to make loans to banks, financial institutions, and other enterprises in trouble. In the first year the economy continued to decline, and, when Franklin D. Roosevelt took office in 1933, average income was about half what it had been before 1929; 90,000 businesses had failed; banking was on the verge of collapse; and 15,000,000 people were unemployed. American capitalism seemed headed for collapse if not revolution.

Roosevelt's economic ideas, such as they were, were purely orthodox. During his presidential campaign he had repeatedly attacked Hoover for permitting a budget deficit. Thus, the electorate did not vote for the New Deal, which is best seen as the sum of pragmatic measures the new administration enacted to grapple with the emergency. It was intended to sustain capitalism while recovery took place through the operation of market forces. In part it consisted of a huge relief program financed by public funds and organized by a host of ad hoc bodies. It did not establish a planned economy but aimed to restore the profitability of business so that idle resources and manpower could be used. But so minimal had been the role of the state heretofore that a series of reforms had to be enacted to enable the government to carry out its task. Once Roosevelt declared that he was offering a New Deal, the program was pushed through with dramatic swiftness. U.S. public opinion was prepared to try anything that offered better prices, profits instead of losses, and more jobs. Each section of the economy was dealt with by a special agency. The gold standard was abandoned and the dollar devalued. The Agricultural Adjustment Act helped farmers through price supports, subsidies, and the restriction of production. The National Industrial Recovery Act embraced a series of measures based on a code of "fair competition" for each industry. Again, the intention was to raise prices by reducing competition. At the same time, wages were also to be raised.

Budget deficits

The New Deal included large public works outlays that brought the federal budget into deficit, but these deficits were too modest to account for the recovery of the economy that slowly began. Although Roosevelt set out to save U.S. capitalism—and succeeded—his policies were strongly opposed by many businessmen and by conservative opinion, possibly because he was supported by organized labour and most wage earners and appeared more radical than he was. In return for this support the administration made concessions to trade unions and authorized a modest social security program. It seems

that once the worst of the Depression had passed, business confidence may have been weakened by Roosevelt's policies, or by the menace that many businessmen believed these policies held over them.

In any case, U.S. recovery received a severe setback in 1937, when government expenditure declined. The lesson was learned, and by 1938 the doctrines of Keynes had made headway, both among economists and in the administration. Pump priming and countercyclical spending were making their way as respectable fiscal practices, but unemployment remained high, and tendencies toward stagnation persisted until World War II began to have its effects.

The New Deal enabled the dire emergency that confronted Roosevelt on his inauguration to be overcome without disaster and without making inroads into private property rights or political freedoms. It did not fully succeed in bringing about recovery. In fact, it took the colossal expenditures of World War II finally to overcome the Great Depression. In this respect the New Deal can be contrasted with Nazi policy, which found the complete answer to the Depression: arms expenditure limited only by physical resources. This was the answer, too, that American capitalism found after 1940.

THE INDUSTRIALIZATION OF THE SOVIET UNION

The Bolshevik Revolution, which in 1917 took Russia out of the capitalist world, and Russia's transformation into a leading industrial power have had enormous significance in the 20th century. In the early years the very existence of the new regime weakened capitalism; and later the viability of the alternative system was demonstrated in the carrying out, through nationalized property relations, of planned large-scale industrialization. Whether or not Premier Joseph Stalin's tyrannical methods were necessary to the process, the results were impressive, indicating to less developed countries that there was an alternative to capitalist industrialization. Elsewhere, too, planning was turned from vague concept to practical reality, acquiring exceptional prestige even in capitalist countries.

The early days: War Communism and the NEP. When the Bolsheviks took power they could hardly have confronted a worse economic situation. When the revolutionary wave in Europe subsided, they were left with a situation unforeseen in Marxist thinking: a workers' party held power in a predominantly backward peasant country ringed by hostile capitalist states. The country was torn and divided by civil war; industry and the transport system were virtually in ruins; famine and disease stalked the land.

When World War I broke out, Russia had a small but advanced industrial sector, and capitalism had made deep inroads into agriculture. In the industrial centres were large concentrations of workers on which the Bolsheviks drew for support. The countryside was experiencing severe social strain as the old village society was undermined by the spread of capitalist relations. The Stolypin reforms, begun after the Revolution of 1905, hastened this process by enabling peasants to withdraw from the *mir* (village commune) and favouring the better-off peasants, or kulaks. The considerable advances, before 1914, in both industrialization and in modernizing agriculture made the tsarist autocracy and its officialdom increasingly anachronistic. The old order was drifting toward a crisis that the war brought to a head once the strain of battle losses began to tell.

The Stolypin reforms

It was in this situation that the February Revolution of 1917 took place, leading to a Provisional Government dominated by the middle class. As the economic situation deteriorated, the masses came into increasing conflict with the Provisional Government, enabling Lenin and the Bolsheviks to lead a successful insurrection and to establish "the dictatorship of the proletariat" in November. Workers and soldiers under Bolshevik direction had been the spearhead of the revolution in the towns, but in the countryside the peasantry had acted independently to seize land owned by the nobility. While endorsing this peasant revolution, the new regime also proclaimed the nationalization of the land. For the moment, control of agriculture remained in the hands of the peasants, and the feeding of the towns then, and for something like a decade, depended upon extracting a food surplus from a mass of peasant holdings still largely cultivated in the old way.

The economic steps taken by the Bolsheviks in their first year included nationalization of all industries employing more than five people, of banks and foreign trade, and of the land. A large part of the country, however, was outside the new Soviet government control, and civil war continued until 1920. There was, in fact, virtual economic breakdown, with many factories closed or producing only a trickle. Priority was given to the front for supplies available in what was substantially a siege economy named "War Communism." The government appropriated what revenue it required either by requisition or by printing money. The black market engulfed supplies, stocks were hoarded, and galloping inflation began.

War Communism was born of an emergency situation in which the economy had collapsed and all efforts were concentrated on survival. Market relations were sharply curtailed, and wages and other payments were made in kind. Grain surpluses were requisitioned, if necessary by force. As production fell and goods disappeared from the shops, workers who had not been drafted into the army or employed in the new administration drifted back to the villages. The hardships of the period were only made bearable by promises of a better life in the future. Discontent, already growing, now became more open. The mutiny at the Kronstadt naval base in February 1921 was the final danger signal. War Communism was abandoned and replaced by the New Economic Policy (NEP).

The NEP was generally admitted as a retreat. Its basis was that it permitted private enterprise in internal trade and to some extent in industry. It also encouraged the peasants to produce a surplus for sale and offered opportunities for their enrichment. Nepmen (private businessmen) and kulaks accumulated capital. The debate in the Communist Party was about how far this should be permitted and, more fundamentally, how fast Soviet industrialization should proceed.

Central planning. It was axiomatic that the U.S.S.R. could only overcome its historical backwardness through industrialization and that this should take place under the control of a central planning agency and on the foundation of a nationalized industry. The NEP stood by this principle. In fact, the State Planning Committee (Gosplan) was set up in February 1921 to "work out a single general state economic plan and methods and means of implementing it." What was at issue was the rate of industrialization, the tempo of growth, which could not in practice be separated from the path that was to be chosen.

The establishment of Gosplan

During the "industrialization debate" of the 1920s, all the great questions about economic growth that interested Western economists a quarter of a century later were thoroughly discussed. After Lenin's death in 1924, Joseph Stalin strengthened his hold on party and state apparatus and formulated the theory of "Socialism in one country"—concentration on building up the Soviet economy at the expense of the international goals of Communism. He was supported by Nikolay Ivanovich Bukharin and others on the right wing of the party, which assumed that NEP would last for a considerable time. Industrialization, therefore, could take place only by slow stages in which the peasantry was encouraged to market its surplus and nationalized industry produced the consumer goods it required in exchange. These perspectives were opposed by the left opposition, whose leader was Leon Trotsky. "Socialism in one country" was rejected as a revision of Marxism because it placed internal interests before those of world revolution. The policy of concessions to the kulaks and the adoption of a slow industrialization tempo were equally condemned, and a program for more rapid industrialization and planning and the encouragement of voluntary collectivization was substituted. Finally, the left opposition was defeated, and Stalin emerged as the master of the party apparatus. Free debate was stifled,

and the dead hand of orthodoxy gripped Soviet political economy.

Meanwhile, the New Economic Policy enabled recovery to proceed fastest of all in agriculture but also in industry. Once the prewar levels had been attained (from about 1925), decisions had to be made on the rate of growth. This depended upon how much grain surplus the peasants were prepared to part with. And that in turn depended upon the ability of industry to supply industrial goods. The main emphasis would have to be upon consumer goods production, and capital formation would be a dependent variable and not the main determinant. If, however, economic policy continued along these lines, industrialization would be slow, and the U.S.S.R. would be vulnerable to capitalist pressure or even attack. Meanwhile, the growing prosperity of the kulaks and the Nepmen would undermine the nationalized property relations upon which the state was founded.

Stalin's change of course

These dangers prompted Stalin to make a brusque alteration of course, taking over much of the policy of the defeated left opposition and applying it with a brutality they had never advocated. From 1928, therefore, the Soviet Union suffered a new period of strain and sacrifice in which, under conditions approaching civil war, agriculture was collectivized and the first stages of planned industrialization were forced through. From this cauldron the U.S.S.R. emerged to face the German attack of 1941 with a much more powerful industry than before. The basis had been laid for it to become the second largest industrial power, but the social and human price had been tremendous.

The drafting of the First Five-Year Plan was begun by Gosplan in 1927; the starting date was from October 1 of the following year, but the campaign did not get under way until 1929. At the end of that year, the plan targets were revised upward at about the same time that Stalin ordered forced collectivization.

A whirlwind descended on the Soviet countryside. The peasants were forced into new collective farms set up without adequate technical means or experienced leadership, and kulaks were forced off the land and deported in large numbers to Siberia. Their resistance was shown by their mass slaughter of their animals, which set back the production of meat and milk products for years. Soviet industry was unable to supply enough equipment, fertilizers, and insecticides to raise the technical level of the new collective farms above that of the former peasant holdings, so agricultural production declined sharply, and bitter resentment long remained among the peasantry.

In industry there was certainly more enthusiasm for the change; but the emphasis of the first and subsequent plans was upon endowing the Soviet Union with heavy industry, not with increasing consumer goods production. This made high growth rates possible, since a large part of production went into new investment. Sacrifices were imposed not only on the peasantry but also upon the working class, including the huge new labour forces drained from the countryside into overcrowded industrial towns. Propaganda explained that these sacrifices would provide a better life in the future.

Social upheaval and oppression. At the price of a gigantic social upheaval the economy was transformed and entered upon a long period of expansion. The First Five-Year Plan, only unevenly fulfilled, was followed by a second in 1933 and a third still incomplete when war came. These plans maintained the priority of heavy industry, with a shift toward defense. While the capitalist world was plunged into depression, the Soviet Union grew rapidly. The growth record was impressive. Although per capita consumption may have declined and agriculture stagnated, industry—and especially heavy industry—showed a formidable advance. The development of social and community services supplemented real wages. Large numbers of people acquired new skills, as well as literacy, while a new managerial and technical elite was trained.

The Second and Third Five-Year Plans

Industrialization was achieved without foreign capital. The expected revolution in Europe had not taken place, and the isolation of the U.S.S.R. was intensified with the rise of Nazi Germany and the growth of armaments in the capitalist countries in the late 1930s. It remained economically inferior to the advanced capitalist countries despite the five-year plans.

Soviet life in the 1930s was lived under the shadow of political repression and terror unjustified by the pressures of industrialization. The arbitrary arrests, purges, trials, prison camps, and mass executions were necessary instruments for Stalin in maintaining his personal ascendancy over the new ruling stratum and the masses. But Stalinism was a burden on the economy; and it is no accident that the height of the purges is paralleled by a dip in the indexes of industrial production.

Purges and executions

Industrialization in the Soviet Union was consciously planned with the whole economy centrally controlled. Material balances, not prices, determined the allocation of resources, and the plan targets were fixed in quantitative terms. Anomalies and abuses, sometimes gigantic, did not prevent the main objectives of the plans from being realized. But the type of planning used—virtually, production at any cost—showed its weaknesses as the output pattern became more sophisticated. Its defects were most apparent in the sectors closest to the consumer, in light industry and retail distribution. After the ravages of World War II had been repaired, the planning methods inherited from the Stalin era came under more critical scrutiny, and modifications and reforms were introduced.

Many criticisms can be made of Soviet planning. The priority given to heavy industry and armaments meant that the Soviet consumer had to suffer. Labour discipline was severe. The high output encouraged in the 1930s by the Stakhanovite Movement made for inequality of earnings. Complaints of wasteful use of productive capacity and materials and of mismanagement and corruption have frequently been made by the leaders themselves. Especially during the purges, initiative and responsibility were impaired by the fear of punishment. Bureaucratic heavyhandedness, unevenness of production and work loads made necessary by sudden spurts to achieve plan targets, production of goods not really required, and difficulties in coordinating the requirements of different enterprises—these are among the criticisms made of Soviet planning.

The later stages of Stalinism. By the time Germany attacked the U.S.S.R. in June 1941 the five-year plans had given the Soviet Union a powerful industrial base. Whatever the revulsion against Stalin's crimes and the excesses and mistakes of the plans, it had been shown that an alternative to a privately-owned economy was viable. But the Soviet Union was still considerably behind the advanced capitalist countries in per capita output and labour productivity, although the gap had been narrowed. Nevertheless, without the accelerated industrialization that the plans made possible the defeat of the U.S.S.R. in 1941 or 1942 would have been certain.

At the end of the war the Soviet Union faced a tremendous task of reconstruction that again postponed the improvement in consumer goods output; but the war had spurred industrial and technological advance, and the planning apparatus was well able to meet the postwar challenge. Within a few years, however, resources were once again being diverted by the Cold War. Stalin's death in 1953 initiated a period of economic reforms. With consumer needs still generally receiving a low priority, the Soviet Union showed its strength in other directions. The sustained and rapid growth of the 1950s, the development of nuclear weapons, and finally the launching of a space satellite caused concern and even alarm in Western countries: the Soviet Union did seem to be catching up to the West. The Soviet economy displayed a number of weaknesses, some inherited from the Stalin era, others arising from its continued separation from the world market and from the difficulties of providing a larger quantity and variety of consumer goods.

Reforms after Stalin's death

The postwar years saw the establishment of nationalized and planned economies in eastern European countries. East Germany and Czechoslovakia apart, these countries had been predominantly agrarian, though, except for Albania, they had some industrial towns or areas. Yugo-

slavia broke with the Soviet Union in 1949; Albania adopted a pro-Chinese policy after the Sino-Soviet schism of the 1960s; and Romania has tended to pursue its own course. Each of the east European countries adopted an economic policy closely following the Soviet example, even when inappropriate for its particular circumstances. The Stalinist model of "Socialism in one country" was applied in literal detail in what appeared to be an attempt to establish replicas of the Soviet Union in the "Peoples' Democracies."

The strains and tensions in Soviet society after Stalin's death were even more pronounced in a number of the east European countries. When the East European Council for Mutual Economic Assistance (CMEA) was established in 1949, much was heard about the need to develop "an international Socialist division of labour." In fact, the economic development of Czechoslovakia, Poland, Hungary, Romania, and Bulgaria had been adversely affected not only by a fundamental change in their international economic relations—away from Germany and western Europe and toward the Soviet Union—but also by their comparative isolation from each other. Even in the 1960s, the degree of coordination and integration provided for by CMEA left each country very much a separate economic entity. While achieving rapid rates of growth and overcoming their former underdevelopment to a considerable extent, development of these countries was uneven.

The economic history of the Soviet Union and the east European countries must be viewed against events in the capitalist world. The Cold War reduced the volume of exchanges with the capitalist countries and thus the possibility of remedying their deficiencies through trade. The United States, in particular, imposed an embargo on strategic exports to the area. Although the output of consumer goods did increase, partly under the threat of such popular uprisings as occurred in Berlin in 1953, in Hungary and Poland in 1956, in Czechoslovakia in 1968, and in Poland again in 1970, the boom in the capitalist world had increased the supply of consumer goods available to the masses of the population in western Europe still faster. In shop-window competition the east European countries continued to be outclassed. Automobiles and many consumer durables were only for a privileged few. Necessities might be in adequate supply, cultural goods—such as books and records—cheap, social services adequate, and accommodation cheap if not abundant, but the range of goods the Western consumer could spend his money on was still not available.

WORLD WAR II AND AFTER IN THE INDUSTRIALIZED NATIONS

The improvised character of the economic policies of the belligerents in World War I was almost entirely absent in World War II. War experience had been digested, and further lessons had been learned from the Great Depression. Perhaps most important of all, World War II had been anticipated with some deliberation by the main participants. In the end, it was the sheer weight of resources that the Allied powers were able to bring to bear that stopped the German military machine and demolished the Japanese "New Order" in Asia.

The impact of economic mobilization

In the capitalist states, economic mobilization imposed unprecedented centralized control and planning directed to maximize output for the war effort while distributing sufficient civilian supplies to maintain efficiency and morale. Production was regulated by a priorities system, labour was directed, and scarce commodities rationed. Prices and wages were controlled and foreign trade curtailed to strategic needs. The market mechanism was largely superseded, but private property rights were preserved, although they might be interfered with on a compensatory basis when the war effort demanded it. Although some war factories were state owned, military contracts were generally fulfilled by private industry on a profit-making basis. The war economy thus required considerable state control, but it remained, in Germany as in Britain and the United States, a capitalist economy.

The desire to knock out the enemy's economic nerve centres and to demoralize, or even destroy, the workers upon which the armed forces depended account for much of the destructiveness of World War II. Modern technology provided supremely powerful weapons that, even before the atomic bomb, could reduce whole cities to rubble in a few hours. The wide-ranging nature of the fighting on the Western Front after the Allied landings in June 1944, as well as in the East, together with the effects of bombing, extended the area of destruction far more than previous wars as well as making it more complete. At the end of the war, therefore, the European countries and the Soviet Union faced an unprecedented mass of destruction. To this were added the losses from the sea war and the dislocation and disruption of prewar trade patterns.

The recovery. The recuperative powers of advanced industrial societies, however, are equally remarkable. The effect of bombing on machines and industrial equipment was less severe than on dwellings and other buildings. Given the human skills available and an abundant supply of people with the incentive to reshape shattered lives and earn enough to buy necessities, it was possible to restore industry and reconstruct the social infrastructure comparatively quickly. From the ruins emerged, in fact, a larger and more advanced industry than in peacetime. It was even claimed that the more devastated regions, such as western Germany, had the advantage over the less battered industries in Britain, which brought much obsolescent equipment and plant intact into the postwar years. Such an explanation of British industrial problems, which were to limit its growth in the period after 1945, was, however, too facile.

The European belligerents endured very severe upheavals and hardships in the later stages of the war and the first years of peace. This was a period of general penury, of rationing and the black market, during which, in some places, money virtually lost all value, and people returned to a barter economy. Only United States loans and aid prevented economic and social breakdown. Meanwhile, large transfers of population took place, particularly from eastern Germany and Czechoslovakia to the West, and large numbers of displaced persons had to be resettled. The many millions of rootless and penniless people provided, however, a mobile and adaptable labour force. German recovery, particularly, benefitted, but in the first years of postwar reconstruction Europe as a whole was helped by labour from this and other sources. Growth was not held back by labour scarcity, and the bargaining power of workers was held in check until full employment levels were reached.

At the end of the war a tremendous disparity existed between the United States and the dislocated economy of impoverished Europe and Asia. Only the United States could supply the foodstuffs, raw materials, semifinished products, machinery, and machine tools that were in scarce supply. The capacity of the belligerent countries to earn foreign exchange, particularly dollars, had been greatly impaired; and they needed loans or aid if reconstruction was to be carried through without balance of payments difficulties. U.S. lend-lease, under which supplies had been made available during the war, ended shortly after Japan's capitulation. The subsequent U.S. loans to France and Britain proved inadequate. Finally, spurred by the growing tension of the Cold War, the European Recovery Program, or Marshall Plan, was launched in 1947 to continue reconstruction and rescue capitalism in western Europe.

The Marshall Plan

The outstanding fact was that the reconstruction boom was succeeded, not by a crisis of overproduction and a return to high unemployment but by expansion that continued through the 1950s and 1960s. The explanation of this latest phase of capitalist development is not made easier by the fact that the period of the Great Depression has not yet been adequately explained. It is possible only to trace the main lines of development and indicate some of the major problems.

Postwar monetary systems. The governments of the Allied countries in the postwar world saw that the currency chaos and beggar-my-neighbour policies of the 1930s must be avoided. First, the reconstruction of the world monetary system was needed, and the United Na-

tions Monetary and Financial Conference held at Bretton Woods in July 1944 sought to do this through the establishment of the International Monetary Fund and the International Bank for Reconstruction and Development (World Bank). The former set out to establish monetary stability and to remove exchange restrictions that hampered world trade. It established new rules for adjusting exchange rates without a return to the old gold standard. Although each national currency was to be expressed in terms of gold, it was the dollar price of gold and the convertibility of the dollar that together provided the essence of the postwar monetary system.

Bretton Woods set the postwar international monetary scene, but there were also important new factors. The recovery of the world market was much more rapid than after World War I. In an atmosphere of expansion international trade was liberalized through the General Agreement on Tariffs and Trade (GATT), and other agencies on an international or regional basis made possible a greater degree of international economic cooperation. Some, such as the Organization for European Economic Cooperation (OEEC), set up by the countries benefitting under the Marshall Plan, were mainly consultative bodies. More important was the movement toward economic integration in Europe that led to the Treaty of Rome in 1957 and the setting up of the European Economic Community (EEC, or the Common Market) in 1957.

Whatever its inherent problems (which emerged in the 1960s), the new monetary system did make possible the recovery and world expansion of the quarter century after World War II. It permitted an increase in the monetary supply on a scale impossible under the pre-1914 gold standard or the interwar gold bullion standard. The new gold exchange system was seriously criticized, particularly by France under Charles de Gaulle, when, in the 1960s, successive international currency crises and the weakening of the dollar were calling it into question. There was, however, no agreement on a substitute, apart from tinkering or modifications to suit the changing conditions created by the growth of western Europe and Japan's economic power. In particular, for the U.S. government, the gold parity of the dollar, established in 1934, appeared to be sacrosanct. It was this continued overvaluation of the dollar upon which U.S. capitalism's world position seemed to depend during the 1960s.

The Bretton Woods agreement confirmed the pre-eminence of the dollar, backed by the huge gold holdings in Ft. Knox, which lasted until the end of the 1950s. Sterling, despite Britain's diminished role, remained a key currency, which, like the dollar, could be used by other countries as part of their reserves for international payments. During the postwar period of dollar scarcity the dollar was as good as gold, in demand all over the world; but the new arrangements allowed a worldwide inflationary trend assisted by the outflow of dollars from the United States. This outflow, beginning as reconstruction loans and then Marshall Aid, was continued in other ways: as military expenditure, by military and economic aid to countries threatened by "Communist aggression," and by direct and portfolio investment. Much of this huge American foreign spending and capital outflow assisted the rebuilding of European capitalism and then its powerful resurgence from the mid-1950s as well as the remarkable upsurge of Japan. In the long run, as the U.S. balance of payments became increasingly unfavourable, the dollar scarcity turned into a dollar surplus; and the mark, the yen, and other currencies became strong currencies. At that point, the Bretton Woods system was doomed.

The first years after World War II were years of reconstruction, and progress on the whole was rapid. Industrial capacity had been built up during the war, and it was often possible to restore production quickly. Governments injected new demand into the economy, and this, together with the pent-up purchasing power of consumers and the reserves accumulated by business, fuelled the postwar boom. The war had shown that government spending could have positive economic results.

Expansion and full employment. From the early 1950s a long-term expansion began in the industrialized countries. Particularly rapid growth rates were achieved by advanced countries that were incompletely industrialized. Western Germany made a spectacular recovery and resumed its place as the hub of industrial northwest Europe. Japan exploded into an industrial giant, showing an astonishing versatility in exploiting the most advanced techniques and production methods. The Scandinavian countries continued with their solid prosperity. The fringe countries of southern Europe, still relatively backward, began to speed into the 20th century. Such older industrial areas as Britain and Belgium experienced growth, too, but at a rate that compared unfavourably with that of the rest of western Europe. The United States also grew less rapidly, thus causing some concern, especially in the years after the launching of the Soviet space satellite, but the huge size of the economy meant that continuing growth produced a steady increase in a per capita income well above that of western Europe and Japan.

Background of postwar expansion

The roots of the sustained growth of the advanced industrial countries in the 1950s and 1960s seem to have been laid in the war and even as far back as the Depression. A long period in which investment was at a low level and, in some fields, actually negative had cleared the way. In some countries the war had added to industrial capacity; machines and machine tools were available to produce new consumer goods. Japanese firms, for example, turned from producing range finders and bombsights to cameras and photographic equipment. Despite the destruction in Germany, there was more plant and equipment available for production than in the 1930s. The United States had considerably increased its industrial capacity. British industry, which had not entirely thrown off the Depression in 1939, entered the postwar world in a sellers' market, and unemployment was minimal until the late 1960s.

Interrelated developments in varied fields tended to make industry much more science based than ever before. This enabled the logic in the industrial system to be worked out more fully, culminating in the fully automated computer-controlled process. But highly sophisticated machines had only a partial application and represented an enormous outlay of capital. Perhaps more important in raising total output was the steady increase of productivity levels over almost the whole of industry. Into this there entered not only straight technology but also changes in business organization, administration, and management, and a better educated and more adaptable labour force. The workers were geared by the incentive of the increased range of goods available to the needs of the industrial machine that produced them. New methods of wage payment were devised to obtain increases in productivity.

Postwar capitalism was characterized by almost uninterrupted output and productivity growth and high levels of investment. As the expansion continued its momentum, workers had to be mobilized from further afield. Germany and France each attracted some 3,000,000 immigrants from southern Europe, Turkey, Algeria, and even sub-Saharan Africa to meet a chronic labour shortage. In particular, steady and prolonged employment became as much a part of the expected order of things as insecurity and unemployment had been in the 1930s.

Full employment greatly enhanced trade union power as well as making employers compete for labour. Some economists claimed that "wage push" was a main factor in the inflationary pressures of the 1960s; but the whole situation was inherently inflationary, and governments that attempted to deal with it merely checked expansion and created a recession. It was generally accepted, therefore, that some inflation was unavoidable if growth and full employment were to be maintained. Government expenditures, especially on armaments, when financed by credit expansion, fed the inflationary trend and were probably its basic source.

The return of inflation

Inflation was undoubtedly linked to the increase in the money supply that government policies required. The ability to increase the money supply was made possible by the currency arrangements established by Bretton Woods and particularly by the outflow of dollars from

the United States, which continued during the 1960s. By the end of that decade not only was the postwar monetary system facing new difficulties but inflation also began to grow bigger and less controllable. Internationally, the renewed strength of the European and Japanese economies, coupled with U.S. balance of payments deficits, had made the parities between the main currencies increasingly unrealistic. The dollar was palpably overvalued in terms of gold and other currencies, and the mark and the yen in particular were undervalued; but any revaluation threatened the high exports upon which West German and Japanese economic growth depended.

The main characteristic of economic development in the advanced industrial countries was thus the apparent overcoming of the trends toward stagnation of the 1930s. War and war preparations, impressive technological changes, the increased role of government policy to maintain growth, and full employment offered continuous opportunities for new profitable investment. Best equipped to take advantage of these opportunities were the big corporations, and so, especially in the most modern industries, the field tended to be dominated by a few giant firms. Mergers and take-overs became familiar occurrences, and American corporations moved into the European market.

Internationalism. *The Common Market.* The European Economic Community, or Common Market, was in the first place a customs union whose members reduced (and were finally to abolish) duties on their own trade while a common external tariff was to be maintained against imports. The original signatories of the Treaty of Rome—France, West Germany, Italy, Belgium, The Netherlands, and Luxembourg—also intended to develop policies to achieve full economic integration. It was hoped that the establishment of a unified market area would yield benefits comparable to those gained by the United States. The history of the Common Market is inseparable from the political trends of the period. While Gen. Charles de Gaulle was in power in France (1958–69), he opposed the development of supranational policies and opposed two attempts at entry into the Common Market by Britain. In 1959 Britain and six other European countries formed the European Free Trade Association (EFTA), which reduced tariffs on imports received from member countries until they were finally eliminated in 1970.

The Treaty of Rome

International corporations. The giant international firm was another focal point of public interest. Although the big British, French, German, and other firms were also operating internationally and the Japanese were extending their international connections, the majority of the "multinationals" had their headquarters in the United States. These firms were particularly strong in the industries that were in the vanguard of technical progress, such as computers, motor vehicles, and electronics.

Luxury goods and the tertiary sector. As incomes in Europe and Japan rose, so more was spent on luxury goods, and the extent to which a country was equipped with different types of such goods became a good index of its per capita income. Even advanced European countries were poorly equipped with washing machines, refrigerators, vacuum cleaners, and the like after World War II, and thus a big field was open for new investment and large-scale production, with U.S. capital often taking a leading part. As a result, European consumption moved closer to U.S. patterns. Europeans also spent more on services of various kinds, so that there was a continuous growth of the "tertiary" sector in Europe as well as the United States, and industrialization itself resulted in a fall in the proportion of the occupied population working in manufacturing industry.

The uncertain future. The 1960s—a period clouded by the war in Vietnam and continued high military expenditure—saw a long period of uninterrupted expansion in the United States that ended in a recession in 1969–70. The Japanese economy boomed through most of the decade, as did those of the Common Market countries. Britain faced a series of balance of payments crises and had to devalue the pound sterling in 1967. Moreover, in Britain, in Italy, and then more widely in Europe, as well as in the United States, unemployment began to rise, and the rate of economic growth began to fall. The combination of these phenomena with continued, and even accelerated, inflation, as well as the increasing strains on the world monetary system, seemed to indicate that the economy of the industrialized countries was changing course. In December 1971, the U.S. devalued the dollar by 8.57 percent, but its balance of payments deficit continued, and a further devaluation was necessary in 1973.

THE DEVELOPING NATIONS

Until recently, economic history has been written mainly from the standpoint of those countries in western Europe and North America that, in the 19th and 20th centuries, achieved relatively high per capita incomes as a result of industrialization. The rest of the world was seen as a source of raw materials and as a market for manufactured goods. Since World War II a great change has come about. The disparity between the rich and the poor countries has become a matter of international concern, and the question of economic growth and development has become a major branch of economics.

The developing countries entered the world market as colonies or economic dependencies of the leading capitalist countries. Despite wide differences, certain fundamental characteristics can be distinguished that make them "developing." Of primary importance is the fact that they entered the modern period without the possibility of autonomous economic growth because of their colonial or semicolonial status. Their economic history, therefore, was not a delayed repetition of that of the developed countries but had distinctive features.

The economic and technical superiority of the European capitalist countries enabled them to dominate most of the globe. The development of other areas of the world was determined by the needs of the advanced capitalist countries. Their role in the world market was to supply minerals and plant and animal products to the industrializing countries. Outside capital opened ports for tradè, built transport links into the hinterland, made investments in mining and plantations, and created a labour force of native or immigrant labourers. A modern sector—essentially a geographical extension of the European metropolis or of American capitalism—was inserted into the existing economy; the result was a dualistic economy in which this modern sector existed alongside, or within, a precapitalist peasant economy.

The position of the modern sector

The phase of economic atrophy. European administrators, concession seekers, traders, and investors were not interested in promoting balanced growth or development, and the social side effects were only of concern if they threatened public order and the safety of investments. Although the developing countries differed greatly, they were all dependent on the world market. They remained primary producers, with their economies hinged to the world demand for one or two minerals or export crops. The scope for industrialization was limited, because as the home market grew, the demand for manufactured goods was met largely by imports. When these were cut off during World War I, an impetus was given to manufacturing industry in such countries as India and Brazil. The catastrophic fall in exports during the Depression of the 1930s, followed by another world war, gave further encouragement to local production, but industrialization remained slow and uneven, despite state encouragement during and after World War II.

At the outbreak of World War I, European colonialism was at its height; virtually the whole globe had been divided by the great capitalist powers. After the defeat of Germany its somewhat meagre share of the colonial spoils was disposed of to the other powers as "mandates" of the League of Nations. Britain and France extended their interests in the Middle East, of growing importance as a source of petroleum.

Wartime conditions had generally encouraged faster industrial development in the dependent countries, but the main stimulus was felt by primary producers. The wartime increase in productive capacity proved greatly in ex-

cess of the needs of the market for most of the interwar period. Consequently, what had been the main if not the only engine for growth—rising exports—faltered and stopped. All primary producing countries faced problems of excess capacity, except those with oil.

The emergence of the native bourgeoisie. The most important social change in the developing countries in the interwar period was the emergence of a native bourgeoisie that became increasingly self-conscious and active in seeking home rule or independence. It was strengthened by the increasing opportunities for local investment in trade and industry and sought to extend them by removing the restrictions and restraints arising from colonial or semicolonial status. After World War I the Western-educated intelligentsia became increasingly articulate in formulating political demands and in expressing the aspiration for material improvement in such countries as Indonesia, India, and Indochina. At the same time, the rise of the modern sector, the accelerated disintegration of the traditional society, and the population explosion produced movements of mass discontent and the threat of revolution.

When the national bourgeoisie joined forces with the masses, the national movement in the colonies became a force to be reckoned with, as in India in the 1920s. During and after World War II, nationalist movements took root in many colonial countries, accusing colonialism of deliberately retarding economic development and claiming independence as a precondition for social advance and higher production and material standards. Industrialization became the battle cry of awakened nationalism in the colonial and semicolonial countries, and it was their main policy goal after independence. The equation of colonialism with poverty and of independence with improved living standards was a powerful factor in the success of the nationalist movements. At the same time, the weakened condition of the European colonial powers after World War II made it impossible to hold down all the colonies by force.

India and Burma became independent in 1947 and 1948, the former being at once partitioned into the Indian Union and the new state of Pakistan. In the following years most of the former colonies acquired their independence, in most cases by a peaceful transition. Colonialism of the old type thus steadily disappeared; the winds of change blew over Africa; new states appeared in Southeast Asia. The Latin American states, although developing, had been politically independent since the early 19th century, and this suggests that political sovereignty could not of itself remove economic dependence. What was sometimes called the Third World—to distinguish it on the one hand from the advanced capitalist countries and on the other from the noncapitalist states of eastern Europe, the Soviet Union, and China—was still greatly involved with the capitalist world market.

The disappearance of colonialism

Links with the capitalist world. This involvement took a number of forms. First, there was the stake of foreign capital in mines, oil wells, plantations, and even manufacturing industry. Short of measures of expropriation, these forms of economic colonialism remained. Thus, the newly independent states not only continued to export mainly primary products but also control of their production and sale, and therefore the profits, remained with the extra-territorial companies based in the United States or in the old colonial metropoles. Second, new industries and economic growth usually required foreign participation, if not control. Practically every newly independent state sought foreign private capital and thus had to offer guarantees against expropriation.

Moreover, the political aspects of the question were complicated by the victory of the Chinese Communists in 1949 and by the wars in Malaysia, Korea, and Indochina. Military and economic aid, especially by the United States, became part of the policy of containment intended to keep the former colonies and semicolonies within the orbit of world capitalism and to strengthen resistance to "Communist aggression." This often meant supporting regimes that resisted the kind of changes involved in development policy.

While direct government aid had an overtly political motive, private capital flowed more or less according to economic opportunity. It was attracted predominantly by the existence of raw materials, especially nonferrous metals and petroleum. Developing areas lacking such resources found it difficult to attract private capital; and, in any case, countries undertaking development required imports of machinery and materials and, in some cases, foodstuffs, involving chronic balance of payments deficits that could only be covered by outside aid and investment. Their economies were thus crucially underpinned by capitalist governments and corporations whose motives have primarily been to ensure supplies of raw materials on favourable terms, to sell commodities in a privileged market, and to open up profitable fields for investment. Industrial growth and industrialization in the developing countries have thus depended primarily upon the initiative of their own governments.

The example of India. Although there is no typical developing country, India epitomizes in a dramatic form many of the problems of development and economic growth. After 1951 India launched a series of five-year plans intended to modernize agriculture and establish an industrial base with a large measure of state participation ostensibly designed to establish a "Socialist pattern of society." In the early 1970s India contained about one-sixth of the human race, its population was growing at about 20 percent each decade, and its per capita income was one of the lowest in the world. Something like 70 percent of the population was occupied in agriculture or living off the land. Despite the existence of modern industrial plants, only about 2 percent of the population was employed in industry. While the masses remained illiterate, there was a highly educated elite and an old established industrial and commercial bourgeoisie. Agriculture was still mainly carried on by primitive methods, while the village, the caste, and the traditional culture provided the framework for the lives of the rural masses.

India's five-year plans

Indian planning has differed fundamentally from the Soviet in that, although it has set targets, it has been able to influence directly only that part of industry under state control. The major part of the economy remains privately owned and unplanned in the sense that the resources it will absorb, the foreign exchange it will require, and its contribution to total income cannot be foreseen. Industrialization depends upon capital accumulation, and in India it requires the import of capital equipment, especially in the early stages, and tends to increase dependence upon the world market both for materials and for markets for manufactured goods. Since it has not been possible to produce a sufficient surplus for investment in India or to direct the surplus produced into the requisite channels, development has continued to depend upon a flow of foreign capital and aid. Foreign capital, therefore, remains of great importance, some of the most modern sections of industry being foreign owned or joint enterprises with foreign firms.

Under the five-year plans there has been a considerable industrial expansion, and optimists feel that a basis has been laid for future industrialization. But, the rapid rate of population increase has meant that per capita income may have fallen during the 1960s. In some areas there has been a significant improvement in cultivation methods, favouring the larger and richer farmer; but the fact that there has not been much change in the agrarian system remains a severe handicap. The spheres of state and private capital have been determined pragmatically. Although some say that Indian planning has subserved the profit-making interests of the private sector, others believe it has inhibited private enterprise. In any case, there is little doubt that private capital has found rewarding outlets in trade and finance as well as in industry.

The Indian experience remains the most impressive attempt to industrialize a developing, former colonial country, without the expropriation of capitalists and landlords, while conserving the forms of parliamentary government. Its success or failure will be of importance not only to Indians but also to the increasingly awakened peoples of the Third World everywhere.

The experience of Latin America. Although each Latin American country has a unique history, the area as a whole provides an interesting comparison with India. Becoming independent in the early 19th century, Latin American states developed economically as suppliers of export staples to the developing world market dominated first by Britain and later by the United States. Where the production of the exports gave rise to urbanization and the growth of a labour force and a local bourgeoisie, a domestic market for manufactured goods came into existence. A basis therefore existed for factory industry, notably in textile production. At the same time, the richer segment of the population was able to afford more sophisticated manufactured imports.

The effects of the Great Depression on Latin America

Latin America was strongly affected by the Great Depression, which reduced world demand for primary products. Cutting off earning power stimulated a number of countries to industrial activity that was essentially producing substitutes for imports without necessarily reducing the previous forces favouring industrialization. The situation during World War II tended in the same direction, and a general expansion of the home market was made possible by the renewed demand for primary exports. The more advanced countries, such as Brazil and Argentina, displayed such trends most clearly.

Since World War II most Latin American countries have had serious economic difficulties, the extent and nature of which has varied according to the country's primary exports. In the 1950s, for instance, stagnation in the agrarian–pastoral economy of Argentina and Uruguay led to balance of payments difficulties and a low growth rate in the manufacturing sector. But other countries, such as Brazil, also had acute instability and inflation problems. The shift toward industry has been too slow to increase per capita income in view of the rapid rate of population growth. Thus, despite a considerable absolute increase in manufacturing output, the region as a whole suffers from under-development. In particular, while there is a reasonably prosperous urban middle class to provide a market for manufactured consumer goods, the great mass of the peasantry and urban wage earners exist in poverty. This duality of the market limits the economies of large-scale production and causes sharp social tensions.

Problems of dependence on primary products

The Latin American economies continue to depend on the export of one or two primary products and are variously dominated by U.S.-based foreign corporations that account for a good deal of manufacturing production serving the higher income groups in the home market. Profits are either used to expand local plants or repatriated to the United States. Industrialization tends, therefore, to be extremely lopsided, despite some state-encouraged investment in capital goods industries. The great mass of the population has benefitted little if at all from industrialization. Their poverty limits the possibility for home market growth and helps to explain the Latin American countries' disappointing economic performance in the 1960s. The general picture is one of great social contrasts in which modern factories and urban areas coexist with primitive agriculture and mass poverty aggravated by a high rate of population increase. Powerful vested interests both in the traditional ruling class and the newly emergent elite are opposed to structural change. The U.S. corporations and the U.S. government are also opposed to political changes necessary to break the social deadlock.

Africa. By 1914 Africa had been divided up between the European powers and brought into the world market through the development of rail and water transport and the establishment of port and harbour facilities. There was great unevenness of economic development reflecting the type of primary product being exploited and the extent of white settlement. In some colonies production of export staples was in the hands of peasant farmers; in others, of plantations established with European capital. Where mineral resources were developed the pattern was to recruit black labour, usually by near coercive methods. Economic control of the African colonies was firmly in the hands of extra-territorial companies, and as trade grew it was often conducted mainly by immigrants, such as Indians in East Africa or Syrians on the west coast. The scope for the emergence of a native bourgeoisie in black Africa was therefore limited.

Primary production in Africa followed the price fluctuations in the world market: stimulated by World War I, production declined during the Depression but expanded during World War II and the postwar boom. Periods of boom encouraged further investment, most of which concentrated on minerals production, largely in the countries of white settlement—the Rhodesias and South Africa. The rise of largely European urban communities created a home market that encouraged industrialization. At the same time, black labour was cheap and abundant, and, under the pressure of a developing money economy, indigenous tribal society was undermined, and a large, uprooted urban proletariat came into being.

Economic development over Africa as a whole was extremely uneven down to World War II, many areas remaining backward and primitive. With the outbreak of the war a new era in African history began, still marked by great unevenness but bringing it rapidly into a world economy with an unlimited and apparently permanent demand for African exports. Inevitably, however, the existing political and social structure had much to do with the way in which this new prosperity was distributed and used. In the countries of white supremacy, existing patterns were confirmed and maintained by policies of discrimination against Africans. In the black African colonies, on the other hand, nationalist movements grew in importance, led by the educated elite that colonial governments had reared and drawing support from the developing bourgeoisie and the peasantry. During the 1950s and 1960s the British, French, and Belgian colonies gained independence, while Portugal continued to maintain its colonial hold despite nationalist movements.

The growth of nationalist movements in Africa

The frontiers of the new black African states corresponded with those of the former colonies. That meant fragmentation, and tribal rivalries within the new states, such as Nigeria and the Congos, could break out into civil war. The states were weak politically and economically and remained dependent upon markets and capital in the advanced capitalist countries. Despite postwar changes and economic expansion, they remained primary producers and, despite a boom in exports, still lacked the basis for rounded economic development. Meanwhile, they experienced all the problems of development, which were aggravated whenever world demand for their staple exports slackened.

The outlook for the future. The less developed countries will undoubtedly become of growing significance in the closing decades of the 20th century. Political independence has not enabled them to industrialize rapidly nor to adapt inherited structures to the modernization of their societies. It has failed to break their economic dependence on a world market dominated by the advanced capitalist countries in which they must sell a small number of specialized primary products. For more rapid growth a number of conditions have to be satisfied. Industrialization—the means to rising incomes—requires capital, either domestically generated or obtained from the more advanced countries. Agricultural output has to be increased to feed a growing population, and masses of the rural unemployed and underemployed must be provided with more productive work. Population increase must be controlled. Education has to be improved and technical skills disseminated on a wide scale. Markets for the produce of these countries have to expand, and a place has to be found on the world market for growing exports of manufactured goods if their industrialization is to succeed.

As the developing countries have found out, such objectives are not easy to achieve. They require coherent programs and plans; they are opposed not only by the old property-owning classes and traditional vested interests but also by those of the bourgeoisie that are satisfied with the existing relationship with foreign capital. Thus, more rapid change may require a social revolution, but the fear of social revolution frightens off foreign capital or may invite retaliatory measures by the extra-territorial corporations or by foreign governments. The problems of

development, therefore, are basically political problems, the key to which is not to be found exclusively in the developing countries themselves.

The ability of the new nations to overcome hunger, disease, and illiteracy, let alone to provide for the mass of the people a level of material existence considered tolerable in the advanced countries has not yet been proven. The economic history of the rest of the 20th century undoubtedly will be concerned very much with their struggle to do so. In particular, different methods of development are under test, as are the possibilities of overcoming the problems of development without social revolution in the developing countries and a fundamental change of approach in the advanced countries.

(T.Ke.)

BIBLIOGRAPHY

From c. 1500 to c. 1750: The most satisfactory survey at a readable level is HERBERT HEATON, *The Economic History of Europe*, rev. ed. (1948). *The Cambridge Economic History of Europe*, vol. 4, *The Economy of Expanding Europe in the Sixteenth and Seventeenth Centuries*, ed. by E.E. RICH and C.H. WILSON (1967), treats the context of expansion in greater depth (population, discovery, prices, science, government policy, economic ideas, etc.). CARLO M. CIPOLLA (ed.), *The Economic Decline of Empires* (1970), usefully brings together a collection of essays on the phenomenon of national decline, including that of Italy, Spain, and The Netherlands in this period. On European agriculture, B.H. SLICHER VAN BATH, *De agrarische Geschiedenis van West-Europa, 500–1850* (1960; Eng. trans., *The Agrarian History of Western Europe, A.D. 500–1850*, 1963), is both comprehensive and original, linking agrarian development closely to price and demographic movements. Mercantilist ideas and policies in Europe are the subject of ELI F. HECKSCHER, *Der Merkantilismus*, 2 vol. (1932; Eng. trans., *Mercantilism*, 2nd ed., 2 vol., 1955); though dinted by criticism, this remains a classic, especially on French and German history. HENRI SEE, *Les Origines du capitalisme moderne* (1926), is a valuable and pithy French contribution to the evolution of commercial, industrial, and finance capitalism. Though nearly half a century old, GEORGE L. BEER, *The Old Colonial System, 1660–1754*, 2 vol. (1912), remains an excellent reference book on early European colonial rivalries. VIOLET BARBOUR, *Capitalism in Amsterdam in the Seventeenth Century* (1950), gives a valuable domestic view of Dutch growth that can be complemented by the vivid study of their overseas empire in CHARLES R. BOXER, *The Dutch Seaborne Empire, 1600–1800* (1965). For the Low Countries as a whole, see J.A. VAN HOUTTE, *Economische en Sociale Geschiedenis van de Lage Landen* (1964). For the most recent and highly detailed study of Antwerp, see HERMAN VAN DER WEE, *The Growth of the Antwerp Market and the European Economy*, 3 vol. (1963). FRIEDRICH LUTGE, *Deutsche Sozial- und Wirtschaftsgeschichte* (1960), is useful for all aspects of German economic history. JAIME VICENS VIVES, *Apuntes del curso de historia económica de España* (1955; Eng. trans., *An Economic History of Spain*, 1969), is the best general survey of Spanish economic problems. For the now classic view of the impact of American silver, see E.J. HAMILTON, *American Treasure and the Price Revolution in Spain, 1501–1650* (1934). A good general survey of France during the *ancien régime* is provided by the *Histoire économique et sociale de la France*, vol. 2, written by a distinguished team of French historians and edited by FERNAND BRAUDEL and ERNEST LABROUSSE (1970). Professor Braudel also wrote the remarkable study of southern Europe in the 16th and 17th centuries, *La Méditerranée et le monde méditerranée à l'époque de Philippe II* (1949). The relationship of French economic development and government policy is the subject of several studies by the American scholar CHARLES W. COLE. The two most important are *Colbert and a Century of French Mercantilism*, 2 vol. (1939, reprinted 1964), and *French Mercantilism, 1683–1700* (1943, reprinted 1965). CHARLES WILSON, *England's Apprenticeship, 1603–1763* (1965), deals with all aspects of England's development leading to the Industrial Revolution. The influence of commercial policy is the special concern of LAWRENCE A. HARPER, *The English Navigation Laws* (1939, reprinted 1964); while P.G.M. DICKSON, *The Financial Revolution in England, . . . 1688–1756* (1967), is the most recent study of an often neglected but vital subject—public finance.

From c. 1750 to 1914: The literature on industrial change and economic development is enormous, and any selective bibliography must of necessity slight important aspects of the subject. A good overview of the process, with heavy quantitative emphasis, is provided by SIMON S. KUZNETS in *Economic Growth and Structure: Selected Essays* (1965) and *Modern Economic Growth: Rate, Structure and Spread* (1966). For the history of technology proper, see MELVIN KRANZBERG and CARROLL W. PURSELL, JR. (eds.), *Technology in Western Civilization*, 2 vol. (1967), offering an uneven but invaluable array of specialized essays. Also useful, though more impressionistic, is SIGFRIED GIEDION, *Mechanization Takes Command* (1948, reprinted 1969). There are a number of works that treat of development on a comparative basis: thus DAVID S. LANDES, *The Unbound Prometheus: Technological Change and Industrial Development in Western Europe from 1750 to the Present* (1969); and WILLIAM ASHWORTH, *A Short History of the International Economy, 1850–1950*, 2nd ed. (1962). W.W. ROSTOW (ed.), *The Economics of Take-off into Sustained Growth* (1963); and M.M. POSTAN and H.J. HABAKKUK (eds.), *The Cambridge Economic History of Europe*, vol. 6, *The Industrial Revolutions and After*, pt. 1–2 (1965), also provide material on the experience of different countries, set side by side, however. On the relations between population and economic change, the best international coverage is provided by DAVID V. GLASS and D.E.C. EVERSLEY (eds.), *Population in History* (1965). There are also important studies of different national experiences: On Britain, PETER MATHIAS, *The First Industrial Nation: An Economic History of Britain, 1700–1914* (1969); PHYLLIS DEANE, *The First Industrial Revolution* (1965); ERIC J. HOBSBAWM, *Industry and Empire* (1968). PHYLLIS DEANE and W.A. COLE, *British Economic Growth, 1688–1959*, 2nd ed. (1967), strongly emphasizes the analysis of statistical time series. DONALD N. MCCLOSKEY (ed.), *Essays on a Mature Economy: Britain After 1840* (1971), is a valuable collection of essays in the so-called new economic history and stresses the analysis of Britain's alleged economic decline in the last part of the 19th century. On the European economy, the old classic of JOHN H. CLAPHAM, *Economic Development of France and Germany, 1815–1914*, 4th ed. (1936), is still useful. It should be supplemented by the articles from European writers that have been translated into English and edited by FRANCOIS CROUZET, W.H. CHALONER, and W.M. STERN, *Essays in European Economic History, 1789–1914* (1969). Some of the essays in ALEXANDER GERSCHENKRON, *Economic Backwardness in Historical Perspective* (1962), are helpful in organizing the data on the European experience. Also interpretative is TOM KEMP, *Economic Forces in French History: An Essay on the Development of the French Economy, 1760–1914* (1971). RONDO CAMERON, *France and the Economic Development of Europe, 1800–1914* (1961), is a study in the migration of capital and entrepreneurship. On the economic history of central Europe, the best sources are not in English: PIERRE BENAERTS, *Les Origines de la grande industrie allemande* (1933); RUDOLF BRAUN, *Industrialisierung und Volksleben*, 2 vol. (1960); WOLFRAM FISCHER (ed.), *Wirtschafts- und Sozialgeschichtliche Probleme der frühen Industrialisierung* (1968). A good general introduction to agricultural economics and change is to be found in the old but still useful JOHN A. VENN, *The Foundations of Agricultural Economics*, 2nd ed. (1933). The best overall history is B.H. SLICHER VAN BATH (*op. cit.*). On the agricultural changes of the 18th and 19th centuries, English historians have made the greatest contribution. See, for example, J.D. CHAMBERS and G.E. MINGAY, *The Agricultural Revolution: 1750–1880* (1966); also E.L. JONES (ed.), *Agriculture and Economic Growth in England, 1650–1815* (1967). MICHEL MORINEAU, *Les Faux-semblants d'un démarrage économique, agriculture et démographie en France au XVIII^e siècle* (1971), is a skeptical view of alleged agrarian advances in France in the 18th century and a useful comparison of the French with the British experience. Overall histories of transportation tend to be popularizations. The classic study of WILLIAM T. JACKMAN, *The Development of Transportation in Modern England*, 2 vol. (1916; 2nd ed. 1 vol., 1965), is still useful. Perhaps the best study of this kind for the United States, though dealing with a limited aspect of the subject, is LOUIS C. HUNTER, *Steamboats on the Western Rivers* (1949). The most exciting works in this field in recent years have been the monographs of the new economic historians, which have tended to minimize the contribution of the railroad to American economic development: ROBERT W. FOGEL, *Railroads and American Economic Growth* (1964); ALBERT FISHLOW, *American Railroads and the Transformation of the Ante-Bellum Economy* (1965). These studies have given rise to controversy and further research; the final results are not yet in. On banking, both RONDO CAMERON *et al.*, *Banking in the Early Stages of Industrialization* (1967); and RAYMOND W. GOLDSMITH, *Financial Structure and Development* (1969), offer comparative perspectives. LELAND JENKS, *The Migration of British Capital to 1875* (1927), remains a classic, for both content and style. The most substantial study of European banking and economic development is by MAURICE LEVY-LEBOYER *Les Banques européennes et l'industrialisation internationale dans la première moitié du XIX^e siècle* (1964). There is no overall treatment of the social consequences of industrialization. A few

works, however, may be cited for either their originality or quality of presentation. The student of labour history will still find much of value in SIDNEY and BEATRICE WEBB, *Industrial Democracy* (1897, reprinted with slight changes 1920). Dealing with an earlier period and stimulating and provocative for its moral indignation is EDWARD P. THOMPSON, *The Making of the English Working Class* (1963). On the French side, GEORGES DUVEAU, *La Vie ouvrière en France, sous le Second Empire* (1946), gives a comparably personal evocation of working class life. JOHN R. KELLETT, *The Impact of Railways on Victorian Cities* (1969), deals with a previously neglected aspect of the relationship between economic change and the condition of life.

The diffusion of the Industrial Revolution from its original centre in western Europe to the rest of the world is a subject in itself. For some of the countries treated in this article, the following works are useful: On the United States, H.J. HABAKKUK, *American and British Technology in the Nineteenth Century: The Search for Labour-Saving Inventions* (1962); and NATHAN ROSENBERG (ed.), *The American System of Manufactures* (1969). For Japan, WILLIAM W. LOCKWOOD (ed.), *The State and Economic Enterprise in Japan* (1965), is the best overall treatment; more quantitative and less interpretative in emphasis is HENRY ROSOVSKY, *Capital Formation in Japan, 1868–1940* (1961). For Russia, two recent studies place special emphasis on the diffusion of technology: WILLIAM L. BLACKWELL, *The Beginnings of Russian Industrialization, 1800–1860* (1968); and JOHN P. MCKAY, *Foreign Entrepreneurs and Russian Industrialization, 1885–1913* (1968). There is a growing literature on the economics of technical change. See, for example, the book of that title edited by NATHEN ROSENBERG (1971); also EDWIN MANSFIELD, *The Economics of Technological Change* (1968); W.E.G. SALTER, *Productivity and Technical Change* (1960); and JACOB SCHMOOKLER, *Invention and Economic Growth* (1966). Finally, the reader should be directed to two works that it is easy to omit because they are so obvious: ADAM SMITH'S *Wealth of Nations*, and vol. 1 of KARL MARX'S *Capital*.

1914–70: SHEPARD B. CLOUGH and THOMAS and CAROL MOODIE (eds.), *Economic History of Europe: Twentieth Century* (1968), a useful collection of documents with explanatory text for the period 1914–46; WILLIAM A. LEWIS, *Economic Survey, 1919–1939* (1949, reprinted 1969), still the best brief account of this period for the student; HEINZ W. ARNDT, *Economic Lessons of the Nineteen-Thirties*, 2nd ed. (1963), a pioneer study of the causes and consequences of the world economic depression; INGVAR SVENNILSON, *Growth and Stagnation in the European Economy* (1954), an influential study made under the auspices of the United Nations by a Swedish economist—a diagnosis of Europe's economic ills between the wars; PAUL ALPERT, *Twentieth Century Economic History of Europe* (1951), one of the first textbooks covering the first half of the century; JOHN MAYNARD KEYNES, *The Economic Consequences of the Peace* (1919, reprinted 1971), a prophetic attack on the Versailles settlement by an economist who was to become one of the most important intellectual forces of the period; A.G. KENWOOD and A.L. LOUGHEED, *The Growth of the International Economy, 1820–1960* (1971), basic facts and explanation of the world trading and monetary system for the student; CARLO M. CIPOLLA, *The Economic History of World Population* (1962), an interesting study of the population problem by a well-known economic historian for the nonspecialist; SIDNEY POLLARD, *The Development of the British Economy, 1914–1967*, 2nd ed. (1969), the fullest account of the problems of the world leader of pre-1914 years under pressure and in decline; LOUIS M. HACKER, *The Course of American Economic Growth and Development* (1970), the last four chapters provide a useful account of the period; SUMNER H. SLICHTER, *Economic Growth in the United States* (1963), a critique of Keynes and the stagnationist thesis; JOHN KENNETH GALBRAITH, *The Great Crash, 1929* (1955), a brilliantly written description of the gyrations of the stock market in 1929 and the sequel; MAURICE H. DOBB, *Soviet Economic Development Since 1917*, 6th ed. (1966), a very full study of Soviet industrialization, with apologetic overtones; ALEC NOVE, *An Economic History of the U.S.S.R.* (1969), a more critical view with emphasis on recent developments; WILLIAM W. LOCKWOOD (ed.), *The State and Economic Enterprise in Japan* (1965), a collection of essays that provides an introduction to the reasons for Japan's spectacular emergence as an industrial giant; ANGUS MADDISON, *Economic Growth in Japan and the U.S.S.R.* (1969), a summary picture and comparison with statistical backing—makes a good starting point for the student; CELSO FURTADO, *The Economic Development of Latin America* (1970), the problems of underdevelopment in historical perspective; GUNNAR MYRDAL, *Asian Drama*, 3 vol. (1968), a massive study of the poverty of nations by a renowned scholar—poses the problems in a generally pessimistic framework; MICHAEL M. POSTAN, *An Economic History of Western Europe, 1945–1964* (1967), the postwar recovery and economic boom as seen by the doyen of British economic historians (a pioneer work); ROBERT LEKACHMAN, *The Age of Keynes* (1966), the life and times of the man who, more than any other, set his mark on the economic policies of the capitalist countries in the period since the depression of the 1930s.

(C.H.Wi./D.S.La./T.Ke.)

Economic Planning

This article deals with central economic planning by governments, and not with the economic plans of individuals and business firms. It is divided into the following sections:

I. The nature of economic planning
II. Economic planning in Communist countries
  Origins of planning in the U.S.S.R.
  Structure of the Soviet planning system
  Planning in other Communist countries
  Assessment of Soviet-type planning
III. Economic planning in developed non-Communist countries
  Origins of planning in the West
  Objectives of planning
  Stages of planning
  Assessment of planning in developed countries
IV. Economic planning in developing non-Communist countries
  Approaches to development planning
  Difficulties in development planning

I. The nature of economic planning

The various forms of planning

By the late 1960s the majority of the world's countries conducted their economic affairs in the framework of a national economic plan. This was a far-reaching change from a generation earlier, when national economic planning existed only in the Soviet Union and, to some extent, Nazi Germany. It should be emphasized that planning takes various forms. In the Soviet Union it is based on state ownership of the means of production, while in western Europe it is thought to be consistent with a large degree of private ownership. The essential idea behind economic planning is that certain key economic decisions should be made, or at least influenced, by some central authority, rather than left to the free play of market forces. In theory it would be perfectly possible to have a high degree of central economic planning in an economy in which everything was privately owned. It would also be possible in theory to have an economy in which, though the means of production were owned by the state, there would be practically no central economic planning in the usual sense of the word. The basic point—that state ownership is not the same as economic planning, and that private ownership is not the same as laissez-faire—is an important one for understanding the history of the idea of economic planning and also for understanding the attitudes toward planning in the modern world.

For the purposes of this article, distinction is made between economic planning in Communist countries, in developed non-Communist countries, and in developing non-Communist countries. Historically, the idea of economic planning was associated with the criticism of capitalism as a system of anarchy and greed. The critics did not give much thought to how the economy would be run after capitalism had been abolished; most of them professed to see no difficulty in organizing the society that would follow. When in 1917 the new Soviet government found itself the owner of all the means of production, it had no blueprint as to what to do next. The evolution of central economic planning in the Soviet Union was largely a pragmatic affair; methods were tried and discarded, and new ones were introduced. The decision in 1927 to undertake rapid and large-scale industrialization required the centralizing of control, since only the government could undertake the task of marshalling the productive resources of the country to achieve its ambitious aims. In western Europe economic planning is adapted to, and can benefit from, the existence of a diversified economic structure, a dynamic class of business managers, and a long tradition of political and economic

liberty. Consequently, although planning implies an extension of the economic responsibilities and activities of the state, the mainspring of economic growth remains the private sector. Only rarely does the state intervene directly in the affairs of individual firms. Economic planning remains indirect and takes the form of collaboration between the public and the private sectors. Producers and consumers are free to adapt their activities to changes in market conditions and relative prices.

Communist methods of planning since the mid-1950s have been in a state of flux, and the highly centralized administrative type of planning inherited after the war from the U.S.S.R. by all the newly established Communist states has undergone considerable modifications. In Yugoslavia planners have followed policies very different from those of the Soviet model, and differences also exist in the practice of other eastern European countries. In the Soviet Union itself, a debate concerning the most appropriate means for implementing plans has been under way for some years, but by the early 1970s, despite numerous efforts on the part of the government to reorganize the machinery of planning, fundamental obstacles had still to be overcome.

The knowledge of these difficulties has given rise in the developing non-Communist countries to a repugnance to Soviet planning methods, while the methods used in the developed non-Communist countries are felt to be not directly applicable, either. There is no settled planning doctrine in these countries, and the approach of governments has remained empirical. In practice, this has meant that the state played a major role in setting up new industries and in modernizing agriculture, particularly in countries of recent independence. The state budget was a major source of savings, supplemented frequently by the local currency counterpart of foreign aid. But the absence of a highly qualified civil service placed limits upon the extent and efficacy of state action. Thus, in urban areas, privately owned businesses continued to supply most local consumer goods. In agriculture, peasant proprietorship or large private estates—particularly for export products—remained the general rule. (Jo.Hac.)

II. Economic planning in Communist countries

ORIGINS OF PLANNING IN THE U.S.S.R.

Soviet improvisations

The kind of economic planning that is practiced in the Soviet Union and in most other Communist countries developed during the 1920s and 1930s in the struggle to industrialize the U.S.S.R. The Bolsheviks had seized power in 1917 without any clear notion as to how an economy should be run. No guidance was to be found in the writings of Karl Marx other than the assertion that a socialist society would operate the economy for the common good, which suggested that it would create organs of economic administration to replace the market system of capitalism. In the future Communist society there would be no money, no profit motive. No wages would be necessary to stimulate effort. It would be "from each according to his ability, to each according to his needs." Economics, a science of exchange relationships or value, would wither away or be replaced by a kind of higher management. "People will be able to manage things very simply, without the intervention of the famous 'value'," wrote Marx's collaborator Friedrich Engels in *Anti-Dühring* (1877–78). In the same tradition, the Bolshevik leader N.I. Bukharin wrote in 1920: "As soon as we deal with an organized national economy, all basic 'problems' of political economy, such as price, value, profit, etc., simply disappear . . ., for here the economy is regulated not by the blind forces of the market and competition but by the consciously implemented plan."

The leader of the Bolsheviks, Lenin, shared these somewhat utopian expectations. It is not clear from his pre-1917 writings just what kind of economic organization he had in mind for Russia should he achieve power. As it turned out, the revolution of October 1917 was accompanied by economic breakdown, a refusal of cooperation from officials and managers, civil war, and uncontrollable inflation. Partly under the stress of these circumstances, partly from ideology, the Bolsheviks moved to establish thoroughgoing state control over industry and trade, nationalized all economic property including land, declared all private enterprise illegal, and demanded that the peasants deliver all farm surpluses to state procurement organs. Money lost all value.

On paper, this period of "war communism" was one of centralized planning. All economic units, except the peasant producers, were subjected to orders from the governments's Supreme Council of National Economy (V.S.N.Kh.). But this initial essay in planning was a failure—except insofar as it facilitated the concentration of the few available resources for the civil war fronts. Rationing functioned spasmodically, there was famine, and output fell drastically.

From Lenin to Stalin

The controversies of the 1920s. In 1921 Lenin introduced the New Economic Policy (NEP). Small-scale private manufacturing, private trade, and free sale of peasant surpluses became legal once again, while the state retained the "commanding heights" (large-scale industry, foreign trade, banking, transport, etc.). The state sector continued to be operated under the aegis of V.S.N.Kh. by trusts and enterprises with state-appointed managers. In 1921 the State Planning Committee (Gosplan) came into existence to advise the government and its economic alter ego, the Council of Labour and Defense; but planning was still a shadowy process. Trusts and enterprises had considerable autonomy and were free to make agreements and grant credits to one another. The planners made forecasts, and government policy decisions influenced the level and direction of state investments; but there was no integrated system of production and allocation planning, even in the state sector, while the private sector was not directly planned at all. In 1924 only 35 percent of the national income, 1.5 percent of agricultural production, less than half of all retail trade, and three-quarters of industrial output were "socialized"; the rest was private.

In 1926–28 a vigorous discussion raged about the future basis of planning. Two schools of thought developed, one advocating "genetic" and the other "teleological" planning. The former, composed of the more moderate and cautious planners, believed that plans should be based on existing trends in the economy and reasonable projections thereof. The latter considered that drastic measures were necessary to speed up the industrialization of the country, and this "teleological" school produced extremely ambitious drafts of the First Five-Year Plan. The radicals conceived the plan as taking precedence over all previous economic decisions so as to enable a sharp break with the past. It was their view that won.

Priority for heavy industry

First Five-Year Plan. The First Five-Year Plan (1928–1932), which was later said to have been carried out in four years, called for immense investments in heavy industry; for example, steel output was to be more than doubled by 1932. Amid great confusion, the planning mechanism was overhauled, and gradually a "command economy" was established. In this system, subordinate units of the economy (*e.g.*, industrial enterprises) operated in accordance with administrative instructions, and they did not decide their inputs or outputs by negotiation with other enterprises, these being determined by the planners. The sole effective criterion of management decision became conformity to plan; *i.e.*, to the instructions issued by the central administrative planning organs (see below). In this way, the political authorities achieved a high degree of control over resource allocation and were able to enforce their priorities. Consequently, when the First Five-Year Plan ran into trouble, the government was able to insist on the fulfillment of most of the plans for key sectors of heavy industry, at the cost of a drastic fall in availability of consumers' goods.

STRUCTURE OF THE SOVIET PLANNING SYSTEM

Control from the top

A rearrangement of the planning system was the necessary consequence of the new tasks it was called upon to perform. In 1932 three People's Commissariats (for heavy, light, and timber industries) replaced the V.S.N.Kh; these were further split, and by 1939 the industries of the U.S.S.R. were run by 21 People's Com-

missariats (the numbers varied in subsequent years). Each commissariat (renamed ministry in 1946) controlled a branch of industry, either directly or through a ministry in one of the union republics. (There were also purely local, small-scale industrial establishments, controlled by republican or local soviets.) The ministries issued instructions to "their" enterprises, organized the supply of materials and components, and also disposed of the output. This organization according to sectors facilitated the handling of problems common to each separate industry, but necessarily put a heavy burden on the coordinating organ, Gosplan.

At the apex of the system stood, and still stand, the leaders of the Communist Party, who decide the policy objectives in economic as in other matters and who make choices as to the means of achieving those objectives. It was the party leaders who decided, in 1957–58, to make drastic changes in fuel policy, to launch the virgin-lands campaign in Siberia (1954), and to greatly expand the chemical industry (1963). The government's role is to carry out the decisions of the party. All key appointments in the economic hierarchy are made or confirmed by appropriate party committees.

The work of Gosplan. It is Gosplan's task to "translate" the politically determined objectives into a consistent set of plan targets. There must be coherence between production and supply at all times, as well as between investment plans and the current production of capital goods. Foreign trade must also be taken into account, as a drain on available resources (exports) and as a source of needed goods (imports). The planners proceed by drawing up a series of material balances, which express anticipated supply of, and demand for, all key commodities. The successive versions of the plan are revised until a general balance is attained, since it is no use planning an increase in production of any item if the necessary additional machinery, raw material, and fuel cannot be made available. The task is of especial complexity in the short term (*i.e.*, within a period of a year), since the plan must take the form of millions of consistent instructions to thousands of enterprises to produce, deliver, transport, and process millions of commodities of a great many shapes, sizes, and types.

Balancing supply and demand

Needless to say, all these decisions must be made somewhere in all economic systems. The Soviet type of "command economy," however, developed under Stalin, provides no criterion for decentralized decision making such as is provided, however imperfectly, by the market mechanism in Western capitalist countries. Consequently, the coordination of all these decisions must be consciously achieved by the planners. The enterprises and other local organs can put forward their own plans and suggestions, as well as applications for material inputs, and this certainly influences what the planners do and know; but in the last analysis, only the centre can be aware of the needs of the economy as a whole. This became apparent after the reforms of May 1957 scrapped the industrial ministries and sought to base planning on territorial economic councils (*sovnarkhozy*) in the republics and regions, with Gosplan as the coordinating body. The object, insofar as it was economic, was to strengthen inter-industrial coordination, regional, republican, and national, at the expense of ministerial industrial "empires." It was soon found necessary to tighten central control, and the resultant proliferation of coordinating agencies and state committees greatly complicated an already highly complex planning process. This led in 1965 to the abolition of the territorial councils and the restoration of the industrial-sector ministries. In the 1960s there were vigorous and far-reaching discussions of reforms in the system, designed to overcome some of its known weaknesses, which spring essentially from the sheer volume of decision making. The latter had led to a considerable de facto dependence on low-level decisions and also on information that it often suited the enterprise management to withhold: For example, since bonuses and promotions depend on plan fulfillment, it "paid" management to argue for a modest plan, to conceal potential, and to hoard materials and equipment (in case supplies did not match production plans, which, owing to the fragmentation of the planning mechanism, they quite often did not).

Proposed reforms. The growth of the economy, itself evidence of the success of the system as an engine of growth, has led to greatly increased complexity. The fact that various components of the plan, originating in different offices, often fail to match one another gives subordinate organs a choice as to which orders to obey.

Reformers in the 1960s were concerned, first of all, with finding ways of adapting the newly developed electronic computer and programming techniques to the tasks of the planners, both in the process of allocation and in making rational choices between alternatives; they were also concerned with seeking criteria for decentralization. They hoped to achieve the latter by utilizing the profit motive and by making prices reflect in some measure supply and demand conditions and opportunity cost, instead of being based on average costs of some previous year. Discussions and cautious experiments were continuing, though no fundamental changes had been attempted by 1971. Detailed production and supply planning was still to a great extent centralized, and prices still failed to reflect supply and demand conditions. The idea of a socialist market or competitive price system remained unacceptable. Speeches at the 24th Party Congress (March–April 1971) reaffirmed the refusal of the leadership to make the basic changes in the centralized "directive planning" system, while seeking to make it more efficient and technologically advanced. The amalgamation of enterprises into large industrial associations, or trusts, was expected to improve matters. The reformers were not silenced, however; in particular, the mathematical economists continued to urge a combination of computerization and the market system.

Agricultural planning

Agricultural planning has had a peculiarly difficult history. With priority given to industrialization, agriculture during the regime of Joseph Stalin was essentially treated as a source of cheap food and materials for the cities. The peasants were, in fact, expropriated by force in the period 1930–35, and the bulk of them were compelled to join collective farms (*kolkhozy*), which were nominally a form of voluntary cooperation. While in Soviet ideology, state farms, operated like factories with wage labour, were preferred to collective farms, they remained of relatively minor importance until after 1954. Mechanization was for many years confined to a very few crops and especially to grain growing. The entire system was primarily designed to ensure deliveries of produce at low prices, and the planners and administrators concentrated on procurements, while production plans were seldom, if ever, fulfilled. Under Nikita Khrushchev in the late 1950s and early 1960s there was a substantial change of policy, with greatly improved prices and a major investment program designed to restore agriculture to health. After initial successes, especially in the period 1955–58, reality again fell far short of planned growth. Endeavours to stimulate output on collective and state farms often took the form of centrally organized campaigns, which led to the imposition by party officials of practices and crop patterns that were unsound or harmful in particular areas. After the fall of Khrushchev, farm prices and incomes were increased and arbitrary intervention was diminished, with resultant gains in output. Peasant incomes greatly improved, but at the same time costs of production rose. Since, for political reasons, retail prices were not increased, a very large subsidy had to be paid out of the state budget (over $13,000,000,000, at the official rate of exchange, in 1971).

PLANNING IN OTHER COMMUNIST COUNTRIES

In other European Communist-ruled countries the Soviet system was extensively copied, even in minor details, until 1956. After that date it is not possible to generalize. In Poland there was talk of drastic reforms in the direction of a looser, market type of socialism, but shortages and low productivity checked the would-be reformers. Alone among this group of countries, the Poles did not collectivize the peasants, and so Polish agriculture was dominated by the private sector. In Czechoslovakia, a de-

centralizing reform, launched in 1958, was followed by some disorganization, caused partly by difficulties in foreign trade, on which that country is very dependent; and in the period 1960–64 the economy underwent a severe crisis, with growth virtually at a standstill. At the end of 1964 the Czechoslovaks adopted a sweeping reform, which went further toward decentralization. Attempts to work out a form of market socialism similar to that in Yugoslavia were halted by the Soviet intervention of 1968.

Efforts to decentralize

Yugoslavia. The Yugoslav Communists developed their own conception of socialist planning after their break with Moscow in 1948. The collectivization of agriculture was abandoned in the early 1950s. The control of the state-owned enterprises was given to workers' councils that would decide their own production programs according to profitability, with prices subject to negotiation. Investments were partly controlled by the enterprises themselves out of profits or by the central planners, partly financed from bank credits. An economic crisis in 1966–67 led to still greater measures of "liberalization," despite some dangerous growth in unemployment. The central authorities still retain a measure of overall economic control through their power to regulate the supply of money and bank credit, and through taxes. The authorities also have some control over price levels, including the power to fix a few key prices. The fact remains that Yugoslavia has for long had substantially more of a market system, and much less of administrative regulation, than other Communist countries.

Hungary. In 1968 Hungary adopted a system of market socialism that, while not going as far as Yugoslavia's, left each enterprise management very largely free to determine its own production program. The central planners were no longer to set obligatory production targets. While some prices remained controlled, others were set free. Enterprises were also given some freedom to buy and sell abroad, and efforts were made to link Hungarian prices with those on world markets. Profits became the principal measure of managerial success, and bonuses based on profits had an appreciable effect on managers' and workers' incomes. Large-scale investments were still controlled by the central planners, but the enterprises were required to finance roughly 40 percent of all investments out of their own resources.

The Soviet type of planning ran into serious difficulties in foreign trade, not least in trade arrangements between Communist-ruled states. Since each country planned centrally, and since (apart from Yugoslavia) none had, by the mid-1960s, a price system that reflected scarcities or needs, trade planning encountered a double obstacle—on the one hand, comparative cost calculations were all but impossible, and on the other hand, the necessary coordination required a supranational authority that could enforce a "socialist international division of labour," and no such authority existed. Efforts to create an effective international planning body encountered opposition from the less developed countries, notably Romania, which refused to accept any limitation on its own industrialization plans. A further difficulty arose when some countries adopted market-oriented reforms and others did not. A Council for Mutual Economic Assistance (Comecon or CMEA) had existed since 1949, supposedly coordinating the planning and trade of the European Communist states, but their trade remained predominantly bilateral rather than multilateral. A resolution adopted by Comecon in August 1971 aimed once again at integrating the plans, prices, and economies of the member countries.

China. Chinese Communist planning at first also followed the Soviet pattern. In 1958, however, came the "great leap forward," an effort to speed up progress by shifting rural manpower into manufacturing. This failed, and later it was admitted that production statistics in industry and especially agriculture had been greatly exaggerated. The quarrel with the Soviet Union contributed to industrial difficulties (the U.S.S.R. had been by far the largest supplier of capital equipment to China) and led the Chinese Communist leadership to devise its own planning methods, adapted to a vast country with poor communications and a low stage of economic development. In the mid-1960s the government stopped publishing regular economic statistics, and knowledge of Chinese economic performance was still scanty in the early 1970s. Production was thought to have been adversely affected by the internal political struggles of the late 1960s known as the Cultural Revolution.

ASSESSMENT OF SOVIET-TYPE PLANNING

The Soviet type of planning grew up under the special conditions prevailing in the U.S.S.R. and was adapted to the task of speedy industrialization of a poor country, with strong emphasis on heavy industry, explicable partly by the logic of industrialization (steel and machinery are more conducive to industrial growth than textiles and jam), partly by concern for military potential. The system made it easy for the authorities to attain a high rate of forced saving and investment and a rapid buildup of basic industries, though at the cost of neglecting for many years the elementary needs of the citizens. The plans were backed up politically by a one-party state and Stalin's system of terror.

Advantages and disadvantages of Soviet planning

Planning of the Soviet type has certain advantages compared to other systems of economic organization. The growth of the Soviet economy in the early 1930s, for example, was in striking contrast to the situation of the "capitalist" world at that time. Through their control over the level of economic activity, financial institutions, and personal incomes, the planners should be able to avoid inflation, the bugbear of economic development in many other countries. In their investment planning, they should be able to avoid the type of wasteful duplication that sometimes occurs under private enterprise through lack of knowledge of other people's investment intentions. The system should also make it possible to reconcile the conflict between short-term sectional profitability and long-term advantage. Insofar as the investment plans of most basic industries depend in the last resort on a quantitative estimate of future demand, the Soviet system is better adapted to making such estimates, since so much of the additional demand is a consequence of the planners' own decisions.

In practice, of course, Soviet-type planning is not always able to realize these potential advantages. There have been repeated instances of overinvestment, followed by the abandonment or freezing of partly finished projects. Inflation, with large, unplanned increases in prices and incomes, affected the U.S.S.R. in the 1930s, though price stability was reasonably well maintained during 1950–67. Experience has also shown that the separate administrative units into which a nationalized economy must be divided can take as narrow and short-term a view as any capitalist entrepreneur. Thus, the most easily accessible forests have been cut, the richest sources of iron ore exhausted, and fallow land put under the plow in order to fulfill current plans, with little consideration of the consequences. A centralized system of material balances is not insurance against erroneous forecasting; the demand for coal was greatly overestimated at the beginning of the 1950s in eastern and western Europe alike. The material-balances approach exercises a conservative influence, perhaps because it is simplest to plan on the assumption that the various technical coefficients will remain constant. The economy grows, but preserves its existing pattern, until the need for change has been overwhelmingly demonstrated. Thus the Soviet planners were very slow to turn from solid fuels to oil and gas and to shift from natural materials to synthetics and plastics. When change does come, it often has to be ordered by the top party leadership, and a political campaign is launched to catch up on arrears caused by past neglect. One must bear in mind, of course, that Western economies also have many imperfections. A theoretical model of centralized planning works as smoothly and as efficiently as a theoretical model of a perfectly competitive market, but neither exists in the real world.

Following the death of Stalin in 1953, the Soviet economic system was presented with new problems. It

showed itself unable to cope effectively with the finer adjustments required in a sophisticated industrial economy. It dealt very clumsily with the satisfaction of consumer demand, though this became more important in the changed political conditions, not only in the Soviet Union but in most of its European allies. The system was in urgent need of reform to save it from being overwhelmed by the complexities inherent in an attempt to base decision making on plans; *i.e.*, on command. Yet Soviet experience may have a marked appeal to those in developing countries (see below) who seek to achieve rapid and drastic changes. Insofar as rapid industrialization involves social and political coercion, the fact of such coercion in the history of Communist countries may be a source of attraction rather than revulsion. (A.No.)

III. Economic planning in developed non-Communist countries

Planning in a capitalist framework

Non-Communist countries that practice some explicit form of economic planning include Belgium, Finland, France, Germany, Ireland, Italy, The Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Great Britain, and Japan. Planning in these countries differs from the Soviet type of planning in that the economies are predominantly run by private enterprise, and the planning authorities seek to guide and influence private decision making rather than to make the decisions themselves.

ORIGINS OF PLANNING IN THE WEST

Until World War II there was no serious attempt at economic planning outside the Soviet Union. During the depression of the 1930s, many governments were forced to intervene vigorously in economic affairs, but in a manner that amounted to economic warfare; this intervention took the form of giving increased protection to domestic producers against competition from abroad; of acquiescing in the formation of cartels and other arrangements among producers to raise prices and reduce competition; and of higher levels of government spending, some of it for relief and some of it for armaments.

At the end of the war there was a shift to the left in the politics of some of the countries, and with it a turn to more positive forms of government intervention. In Great Britain the Labour Party secured a large majority in Parliament in 1945, and with it a mandate for policies aiming at more social equality. In Scandinavia, particularly in Sweden, moderate left-wing traditions in government made a transition to planning politically acceptable. In France, left-wing groups, including the Communist Party, emerged as the dominant political force after 1945 with programs of far-reaching social reform. More important, a group of eminent public servants, engineers, and business leaders—continuing a tradition of French 19th-century capitalism known as Saint-Simonianism—were in favour of the state taking a leading role in economic affairs.

While the initial impulse to planning came from the political left, actual decisions by governments to plan were based on practical considerations rather than on political doctrine. The decision to plan most often followed a crisis in a country's economic affairs, as was the case in France after World War II, when there was an urgent need to reconstruct and modernize the economy. In the United Kingdom, the setting up of a medium-term plan accompanied the emergency measures taken to deal with a balance of payments crisis in July 1961; and the Labour government's National Plan of September 1965 was formulated in similar circumstances. In Belgium and Ireland dissatisfaction with the past performance of the economy was a major reason for planning. Belgium had not shared in the European prosperity of the 1950s, and accordingly, in 1959, the government adopted a plan aimed at an increase of 4 percent a year in the gross national product, practically double the rate achieved from 1955 to 1960. Its planning methods were modelled on those of France (see below).

The French example also influenced planning in other European countries. In Great Britain, a Conservative government undertook, during a balance of payments crisis in July 1961, to set up a National Economic Development Council to draft a five-year economic plan that would emphasize much more rapid economic growth. The Netherlands, which had been very successful since the war in achieving balanced economic growth, initiated five-year plans in 1963 through the medium of the Central Planning Bureau, which had for some years been advising on national budgetary policies. Italy had first turned to planning in the 1950s, when a plan for the development of southern Italy was launched; later, attempts were made to extend this example of regional economic planning into a plan for the national economy. Even in West Germany, where the Christian Democratic governments had emphasized a policy of strengthening the free market, a need for some central management of the economy was increasingly recognized.

It may be said in general that planning in the developed countries does not represent an attempt to apply preconceived doctrines. Although left-wing parties are more ready to propose increased intervention by the state in economic affairs than are the more conservative ones, the attitudes and responses of governments in the 1950s and 1960s were mainly determined by practical considerations. The key factors were dissatisfaction with the performance of the national economy as compared with the economies of other countries, and the emergence of serious obstacles to a continuation of the upsurge that had begun in the late 1940s.

OBJECTIVES OF PLANNING

Growth, balance, and consistency

The plans drafted in the developed countries all postulate target rates of growth that are high by historical standards (usually 4 or 5 percent annually). They require that a choice be made in allocating rates of increase to the various kinds of expenditures on available goods and services (private and public consumption, social investment, directly productive investment, stocks, and exports). Further, as the plan has to be balanced, the total of these components of overall demand must fall within the probable total available supply of goods and services, after allowing for the desired level of the current balance of payments surplus. The rates of increase of some types of demand—directly productive investment and stocks in particular—are fixed within fairly narrow limits by technological considerations once the overall rate of growth of output has been chosen. But an important area of choice concerns the respective rates of growth of private consumption by individuals and what has come to be called collective consumption; *e.g.*, education, health, urban facilities, provision for culture.

Another planning objective that has become very prominent (and that figures in the plans of Great Britain, The Netherlands, France, and Italy, to mention only a few) is the correction of imbalances in regional development.

One of the most important functions of economic planning is to achieve consistency among different economic objectives. Some desirable goals are likely to conflict with others. While it may be possible, for example, to stimulate the economy so as to obtain sharply higher levels of output and employment, the measures required to do this may also produce rapidly rising prices, which in turn will lead to rising imports and falling exports; the result may be a balance of payments crisis.

STAGES OF PLANNING

Public discussions of objectives

Choosing the targets. One of the chief merits of national planning is that it publicizes the choices before a country and encourages discussion of them. In France, the Parliament is consulted on the broad outline of the five-year plan, which is presented in terms of a number of alternative balanced sets of objectives. More government spending on social services, for example, can be shown as implying a slower rise in individual incomes after taxes, or—a rather more difficult choice to make explicit—a higher growth rate can be shown as requiring greater willingness on the part of producers and consumers to adapt themselves to changes in markets and technology. Since 1964 the French government has pre-

sented the Parliament with an outline of the main options it favours for economic policy during the coming period. The document is appraised by an advisory body drawn from business, agriculture, the trade unions, and the professions and then submitted to a vote in Parliament. A year later, the complete plan, drafted according to the guidelines approved by the Parliament, comes back again for a final vote.

In other countries, the choice of objectives may be left to the government, which thus makes the plan a part of the program upon which it will stand or fall at the next election. Or the government may prefer to detach itself almost entirely from the plan. In West Germany, where planning is less explicit, projections of economic trends are set out in a technical document that does not have the formal status of a government draft and circulated to the governments of the *Länder* (states) and to employers and trade unions. Such projections are thought to have some influence on public thinking and on the expenditure plans of business and government. In Great Britain, since the Conservatives returned to power in 1970, there has been less enthusiasm for public discussion of planning objectives than there was under the preceding Labour government. The Dutch procedure is to leave the main responsibility for drafting projections of economic trends to the Central Planning Bureau. In that country, however, business and labour have generally been ready to take account of such projections when drawing up their own plans. In Belgium, after the period of strained relations between the main language groups during the 1960s, regional considerations have been very much to the fore in all planning discussions. In 1970 the institutional setting of the new plan for the years 1971–75 included three regional economic councils whose views were influential in the establishment of the national targets.

An important issue for the European countries in the 1970s has been the impact on their economies of the European Economic Community (Common Market). The British government's reluctance to commit itself to economic targets for several years ahead arose from uncertainty as to whether Britain would join the EEC. The French planners gave this same uncertainty as one of their reasons for keeping the targets in the plan for 1971–75 few in number.

Bringing the public into the planning process

Detailed drafting. Since targets in this type of planning do not constitute orders to producers and consumers to do or not to do particular things, the plan has to take account of what private firms say they intend to do, or could do, during the period of the plan. Drafting a plan, therefore, requires arrangements for bringing the representatives of the private sector—both employers and workers—into the planning process. This is usually done by setting up a tripartite body (the High Planning Council in France, the Council for Economic Planning in Sweden, for example) where representatives of the private sector are brought together with representatives of the government. These bodies have much greater responsibility for economic policy making than the older consultative committees that exist in all developed countries. In exchange, the representatives of the private sector are expected to commit their organizations to helping carry out the decisions made.

Detailed discussions on planned targets and ways of achieving them are frequently entrusted to tripartite committees, one for each major sector of activity. The French have 21 such committees, called modernization commissions, some of which—the horizontal commissions—deal with specific topics such as investment and its finance, regional policy, or manpower questions, the others being specialized by sectors of industry. In the United Kingdom, similar committees for major branches of industry were set up after 1963. It is important, although difficult, to ensure that the targets of the branches as regards investment, employment, and output do not conflict with each other and are consistent with the targets for the economy as a whole. The French try to avoid such conflict by giving each planning commission an outline of the overall plan as a starting point for its own projections. This procedure is said to facilitate coherence, since the individual commissions all work toward the same targets.

Government departments in most countries draft programs several years in advance for expenditure on such public projects as education, road building, urban improvement, and hospital construction. But it often happens that these programs are not mutually consistent. One of the advantages of an overall plan is that the confrontation and coordination between the various programs has to be done in an explicit manner. The task can be difficult; watertight administrative compartments are not always easy to break down.

Efforts to incorporate into the plan all that can be known about the intentions of the private and public sectors need to be supplemented by more general economic analysis. As has been stressed already, the plan should be, above all, coherent; and the best way of ensuring this is to build up the planned targets within the framework of the national income accounts. Since World War II, such accounts have been drawn up on an annual basis in all the developed countries. From this it is only a step to projecting the national accounts for several years ahead.

Recent innovations in planning

Earlier types of economic planning leaned heavily upon the method of economic balances. This consists of setting out the quantities of economic resources that will be available during the plan period and comparing them with the quantities demanded by the plan. Four balances are of key importance: the demand for and supply of goods and services; of savings; of manpower; and of foreign exchange. The notion of balance is a valuable one in planning, since no plan can be successful if it outruns the available resources. The method has its difficulties, however, because of the numerous interactions between different sectors of the economy, with the consequence that an adjustment in one set of balances requires adjustments in the others—a complex and time-consuming process. The method of balances is also unable to throw light upon a more fundamental aspect of economic decision making, the need to choose among alternative courses of action on their own merits.

Other methods of planning that have in varying degree replaced the method of balances include mathematical model making and cost–benefit analysis. A mathematical model consists of a series of equations that describes the structure and working of the national economy, enabling various sets of targets to be "tried out" by feeding their values into a computer. It is too early, however, to claim for economic models any clear superiority over the more pragmatic method of economic balances. The most systematic use of models has been in The Netherlands and France. In the Scandinavian countries, a strong mathematical tradition among economists has favoured their adoption.

Cost–benefit analysis, sometimes known as the planning–programming–budgetary system (PPBS), represents an effort to improve the planning of government expenditures. Starting from the fact that public expenditures are not sensitive to the economic considerations of price and profitability but that they nevertheless use up scarce resources that have economic value, PPBS attempts to provide rules of calculation when deciding upon the allocation of these resources. A first step is to break down public expenditure into its main functions and to divide each of the latter into programs that can be identified with government policy objectives. Then an effort is made to evaluate the effectiveness of each program in achieving its declared objectives, together with a consideration of alternative ways of achieving the same objectives and the costs of those alternatives. The United States government pioneered in the application of PPBS to government activities in the 1960s. Great Britain introduced it in the Ministry of Defence in the late 1960s and then began to extend it to other departments, particularly in education and science. In France, the government decided to apply the system in 1968, first in the Ministry of Defense and then in relation to energy, town planning, and such departments as posts and telegraph.

By the early 1970s PPBS had become an integral tool of national economic planning.

Carrying out of plans. In their planning, all the non-Communist countries leave a large margin of initiative to individual producers and consumers; *i.e.*, they rely upon market mechanisms rather than upon direct controls. There is no contradiction between this state of affairs and the existence of a plan. First, as already noted, the fact of associating the private sector with the drafting of the plan encourages its representatives to be dynamic in their behaviour, a prerequisite for any successful growth policy. Second, when a plan is drafted in this way, all sectors are encouraged in making their own plans to adopt the same general assumptions about the growth of the economy. In France, this aspect of planning is called "generalized market research." Third, participation in drafting the plan can help to make the representatives of the employers and workers more conscious of existing obstacles to growth and encourage them to find constructive solutions for overcoming them.

Problems of flexibility

But the government is not freed from participating actively in carrying out the plan. The government's budget and its monetary policy must be used in such a way that overall demand rises as steadily as possible along the desired growth path. The problem of avoiding checks to growth caused by balance of payments difficulties is likely to call for more direct means of influencing the private sector of the economy. That is why governments in developed countries are increasingly seeking to work out "incomes policies" to prevent undue increases in costs and prices. In some countries, planning and incomes policy tend to be linked.

If certain objectives in the plan are not likely to be achieved, it may be necessary to revise some of the targets. In Great Britain the overall growth target in the 1965–70 plan had to be abandoned as early as 1966 because the estimates of productivity growth had been set too high. The Fifth French Plan (1966–70) had to be revised following the social upheavals in May 1968. Ireland and Italy have also had to revise plans because of unanticipated crises.

Direct government intervention

There is no easy solution to the problem of how to reconcile five- or six-year plans with the need to adjust economic policy to the phases of the business cycle. In Japan, a country that has experienced several sharp though brief recessions since the end of the postwar reconstruction period, governments are willing to cut back their public expenditure programs and to take restrictive fiscal and monetary measures because they believe that the economy will soon rebound once the restrictive measures have done their work. Japan's remarkable economic growth has borne out this confidence. In the less flexible economies of western Europe, governments are more circumscribed in their action. France, Germany, and The Netherlands have instituted systematic reviews of their plan targets every one or two years and are quite prepared to revise the targets if necessary.

In addition to these general measures for implementing plans, some governments attempt to influence more directly the behaviour of private firms by granting depreciation allowances for investments and tax exemptions for some activities, such as expenditure on research. In France, which has gone further than most countries in using such measures to implement planned targets, large requests for medium-term credit from the banks are granted after consultation with the planning office, and private firms have to receive the permission of the Ministry of Finance before issuing debentures on the capital market.

But it is in the field of regional development policy that governments intervene most actively to influence the decisions of the private sector of the economy. There is widespread use in western Europe of tax incentives and grants from public funds to encourage private firms to expand in less developed regions. Frequently, permission to set up a new factory, or to expand an existing one, in already congested areas where labour shortages are acute is refused. In addition, governments endeavour to stimulate growth in the less developed areas by improving transport facilities, housing, and urban infrastructure generally. These efforts have not shown any significant success in raising incomes or reducing unemployment in the less developed regions. This does not mean that the regional policies have been without effect, however, for without them the regional disparities might have worsened. Present-day plans for regional development attempt to concentrate on a few selected "growth points" where it is thought that favourable geographical characteristics or the presence of raw materials or other advantages offer the most hope of success.

ASSESSMENT OF PLANNING IN DEVELOPED COUNTRIES

Economic growth as an object of policy

The acceptance of economic planning in a large number of countries and by a variety of political parties and governments indicates that it is now a permanent feature of national policy in many of the developed industrial countries. By the 1970s planning had become more flexible and selective than in earlier years. The general consensus was that the government should seek to create conditions in the private sector of the economy that would encourage growth; this would include measures to establish and maintain competition. The corollary was that governments should try to avoid applying detailed controls over the private sector in peacetime, since these lead to reduced efficiency.

Some critics of planning have charged that the planners put too much emphasis on measures to accelerate economic growth, overlooking the social costs involved. A difficulty with simple growth targets is that they do not measure the increase in side effects such as pollution, noise, and the destruction of nature; on the contrary, they show the expenditures on combatting these effects as part of the growth itself. (For example, expenditures on conservation or smog abatement are included in the statistics of national income and gross national product.) Similar contradictions are found in the easy equation of economic growth with the general welfare: it is possible for income per head of the population to rise while the incomes of certain groups fall; quite frequently, some groups, such as the aged, handicapped, unemployed, and certain ethnic groups, do not share in the increasing prosperity of their country. The general welfare obviously includes elements such as health, housing, education, and economic opportunity as well as economic growth. This concern with the qualitative aspects of economic growth has left its mark upon the objectives written into the economic plans, which increasingly spell out general social aims.

Governments have also adopted a more flexible approach to the setting of targets. During the 1960s, mainly under the influence of French practice, targets for the private sector were often spelled out in detail. But experience showed that elaborate targets were rarely achieved, although they were likely to be considered by public opinion as representing firm commitments by the government. Since then, care has been taken to distinguish between firm targets and estimates. The firm targets are set only for areas over which governments have a considerable degree of control. While the government may have some influence on output in manufacturing industry, this depends more directly upon such things as the state of business conditions abroad, the purchasing habits of consumers, trends in prices and incomes, and so on. It is noteworthy that Japan, which holds the record for economic growth since World War II, has never used detailed output targets in its multiyear plans. (Jo.Hac.)

IV. Economic planning in developing non-Communist countries

Since the end of World War II, it has become an accepted practice among the governments of the developing countries to publish their "development plans." These are medium-term plans, usually for a five-year period. The aim is to select a period long enough to include projects spanning a number of budget years but not so long as to delay periodic assessment of the development effort stretching over a series of plans. The development plan attempts to promote economic development in three

Why poor countries plan

main ways: (1) by increasing the overall rate of investment; (2) by carrying out special types of investment designed to break bottlenecks in production in important sectors of the economy; and (3) by trying to improve the coordination between different parts of the economy. Of these, the third is perhaps the most important and the least understood function of economic planning. The other two functions of planning cannot be efficiently carried out without effective economic coordination between the different government departments and agencies within the public sector and the private sector. In order to improve coordination it is necessary to spread reliable economic information to indicate the future course of the government's economic intentions and activities so that the people concerned, both in the public and the private sectors, may make appropriate plans of their own to bring them in line with the government's plan. In fact, this may be regarded as the main reason for publishing development plans, although this point is not always clearly appreciated by the governments that issue them.

APPROACHES TO DEVELOPMENT PLANNING

From simple to comprehensive planning

Listing of development projects. The newly independent countries, just starting to plan their economies, usually begin with a simple type of development plan. In most cases this is merely an ad hoc list of individually conceived social and economic projects that the various government departments have submitted for the plan. So long as the projects are well selected (say, to break some obvious bottlenecks in production) and are well designed in a technical sense, such a simple plan may be quite serviceable. But it tends to suffer from a number of weaknesses arising from insufficient coordination. (1) Since the projects are drawn up on a piecemeal basis in separate government departments, there is usually no systematic attempt to compare the relative costs and benefits of the schemes proposed by the different departments on a uniform basis. As a consequence, the collection of projects included in the plan may or may not represent the most productive pattern of investing the available resources of the government. (2) A lack of coordination frequently leads to wasteful duplication and a failure to take advantage of complementary relationships between individual projects. (3) A simple listing of the projects does not provide a clear-cut system of priorities in their implementation. Typically, the projects that are relatively easy to implement are pushed far ahead of others that, although requiring a longer time to prepare and implement, may contribute more directly to the expansion of national output and government revenue. This can have serious budgetary consequences when the projects that are easier to implement generally happen to be in the field of social welfare, education, and health and—although they may indirectly contribute to economic development in the longer run—entail an increasing stream of recurring government expenditure after their completion.

Integrated public investment plan. An obvious way of remedying these defects is to formulate a more systematic plan of the public investment program as an integrated whole. In order to do this, it is necessary to begin by making a careful estimate of the total amount and time pattern of the financial resources that the government expects to receive during the plan period from domestic sources and from external loans and aid. Next, it is necessary to make realistic estimates of the costs and benefits of the alternative investment projects within the public sector as a whole so as to select the most productive combination of projects, taking into account significant complementary relationships between the different projects. In selecting the best combination of projects to be included in the plan, it is necessary to pay special attention to the time pattern of costs and benefits. A poor country, with limited sources of government revenue, would have to discount future benefits heavily relative to the more immediate benefits and would have to give priority to the type of project with quicker returns in the form of expansion in output and tax yields over the type of project that may promise higher rates of return, but only in the more distant future.

The problems of carrying out an integrated public investment program serve to emphasize the crucial role of the annual budget in development planning. At the aggregate level, with a given amount of external aid, the stream of the total investable funds available to the government during the plan period depends on its ability to raise revenue (and borrow from domestic sources) and, equally important, to control its nondevelopment, or "consumption," expenditure year by year during the plan period. At the individual project level, the fact that a project requires a number of budget years to complete does not dispense with the need for annual budgetary controls to ensure that it is being implemented in stages, according to the timetable as originally planned. Indeed, it is only through the discipline of annual budgetary controls that a medium-term development plan is likely to be kept nearer the course as originally planned.

Planning the whole economy

Comprehensive planning. Few developing countries have submitted themselves to the budgetary discipline necessary for implementing an integrated public investment program. This has not deterred them, however, from jumping from a simple type of development plan to "comprehensive" economic planning, embracing both the public and the private sectors and regulating both the aggregate level of economic activity and its detailed composition. The drive toward comprehensive planning arises from various causes: from a distrust of the automatic working of the market mechanism and its ability to promote economic development; from a desire to assert national economic independence by government control of foreign trade and investment; and from the theories of economic development, fashionable during the 1950s, that emphasize the need for a "big push" to overcome technical indivisibilities and the need for a simultaneous setting up of a number of mutually supporting projects to enjoy the benefits of technical complementaries. The economic development plans published by the developing countries in the 1960s were fairly elaborate. The trend to "quantitative" planning encouraged the use of elaborate statistical estimates and projections even when the primary statistical sources on which these computations were based were often unreliable or conjectural. Advanced mathematical techniques were also increasingly employed.

Basically, there are three parts to such a development plan: (1) the target figures for increase in per capita income and consumption to be attained at the end of the plan (with estimated figures for the intermediate years during the plan); (2) estimates of the quantities of various resources, such as capital, manpower, and foreign exchange, needed to implement the target figures (including the time profile of the rate at which these resources will be required during the plan); and (3) parallel but independent estimates and projections of the quantities and the time pattern of these resources expected to be available both to the government and to the economy as a whole during the plan period. The elaborate planning documents issued by some developing countries may be described as attempts to quantify as far as possible the information required under the three heads and to test the formal consistency of the plan. This essentially consists in asking (a) whether the total amount of available resources is sufficient to meet the total requirements of resources as set by the target figures, and (b) whether the allocation of resources planned for different sectors is consistent with the detailed target figures for the increased output of different goods and services required for consumption and investment. When the resources required by some industries are intermediate goods (the output of other industries), input–output tables are frequently used to check whether the outputs of different industries are sufficient to supply not only the target figures for final use in the form of consumption and investment but also the "indirect use" required by other industries. The more advanced planning models using programming techniques then attempt to solve the further question (c) whether the planned pattern of allocat-

ing resources is the most efficient; *i.e.*, whether it minimizes the resources needed to meet the target figures as compared with other patterns.

DIFFICULTIES IN DEVELOPMENT PLANNING

In spite of increasing professionalism in the formulation of development plans on paper, the practical performance of the developing countries in implementing development plans of any complexity has not been very encouraging. Development plans, however elaborately formulated on paper, rarely get beyond the first and most obvious practical hurdle, viz., how to equate the total amount of investable resources required to fulfill the target rates of economic development set by the plan with the total supply of investable resources. This arises from the practice of starting with some minimum "politically acceptable" target rate of economic growth while optimistically assuming that the problem of providing the necessary resources will somehow look after itself. This is the opposite of realistic planning, which should start from the supply of available resources and find out the maximum possible rate of economic growth that can be got out of these resources. Sometimes, also, the plan may be deliberately drawn up larger than can be sustained out of the available domestic resources as an elaborate window-dressing exercise to obtain a greater amount of external aid. When external aid fails to fill the planned gap in resources, the typical reaction is not to reduce the size of the plan but to take the politically less painful (and the administratively simpler) expedient of keeping to the publicly declared target rates of the plan and then trying to fill the gap in resources out of "forced saving," which it is hoped will be generated by budget deficits and inflation. Unfortunately this "forced saving" approach has not worked in most developing countries, because the public soon loses confidence in the stability of the purchasing power of money as prices tend to rise in step with increases in government expenditure. The pressure of domestic inflation increases the pressure of demand for imports, while rising domestic production costs discourage expansion of exports.

The need for realistic targets

The balance of payments problem. This disequilibrium situation may be eased by raising the rate of interest to reduce the demand for investable funds and to encourage saving, and by devaluation of the currency to discourage imports and encourage exports. But the governments of the developing countries generally prefer to maintain artificially low rates of interest (lower than the prevailing rates in the advanced countries during the 1960s) and to supply cheap loans to the public sector and to some favoured sections of the private sector engaged in modern manufacturing industry. They are also reluctant to devalue for fear of further raises in prices through speculation and the higher costs of imported goods. Thus they tend to rely heavily on detailed administrative controls and import licensing to ration the scarce supply of investable funds and foreign exchange. The attempt to control the entire economy in detail through a network of direct administrative controls inevitably results in inefficiency and delays, aggravated by the inadequacy of the administrative machinery and a shortage of competent civil servants. The closer the "integration" planned between different sectors of the economy, the greater the damage to efficient coordination, since delay in one sector causes widespread delays in others.

The overlooking of qualitative factors. The main weakness of the formal "quantitative" economic development plans is that they distract attention from various qualitative factors, such as the practical capacity of the administrative machinery to implement the plan, the degree of political stability, and the extent of public confidence in the government's willingness and ability to carry out stated aims, which are crucial for the practical success of planning. The developing countries that are not well provided with these crucial qualities may do much better with a less comprehensive type of planning, making a greater use of indirect controls through the market mechanism and concentrating attention on the breaking of bottlenecks, particularly to the expansion of production in agriculture and export industries. Some developing countries have in fact succeeded in attaining rapid rates of growth just by concentrating on these vital sectors of the economy without an elaborate paraphernalia of planning. (H.My.)

BIBLIOGRAPHY

Developed countries: The best general introduction to the development of economic planning in western Europe is M.M. POSTAN, *An Economic History of Western Europe 1945 to 1964* (1967). Of a more technical nature are: B. HICKMAN (ed.), *Quantitative Planning of Economic Policy* (1965); and JOAN MITCHELL, *Groundwork to Economic Planning* (1966). The best general treatment of planning in Japanese context is G.C. ALLEN, *Japan's Economic Expansion* (1965). An assessment of three European experiments in planning from a rather critical standpoint is GEOFFREY DENTON, MURRAY FORSYTH, and MALCOLM MACCLENNAN, *Economic Planning and Policies in Britain, France and Germany* (1968). Three aspects of planning that are very much in the foreground of present-day discussions of planning are dealt with in HAROLD A. HOVEY, *The Planning-Programming-Budgeting Approach to Government Decision-Making* (1968); E.J. MISHAN, *The Costs of Economic Growth* (1967), a critique of excessive preoccupation with growth; and E.B. SHELDON and W.E. MOORE (eds.), *Indicators of Social Change* (1968), a discussion of the techniques of measuring social aspects of economic growth.

Communist countries: General works on the Soviet-type system of planning include: A. NOVE, *The Soviet Economy*, 2nd ed. (1969); and H.J. SHERMAN, *The Soviet Economy* (1969). Very useful, though somewhat introductory in character, are M.C. KASER, *Soviet Economics* (1970); and R. CAMPBELL, *Soviet Economic Power*, 2nd ed. (1967). For the history of Soviet planning, see A. NOVE, *Economic History of the U.S.S.R.* (1969); and the highly detailed and well-documented work of E.H. CARR and R.W. DAVIES, *Foundations of the Planned Economy*, 1926–29, 2 vol. (1970). A guide to discussions of economic reform in the Soviet Union is G.R. FEIWEL, *The Soviet Quest for Economic Efficiency* (1967). Much useful material is contained in two collections of studies prepared for the Subcommittee on Foreign Economic Policy of the Joint Economic Committee of the Congress of the United States, *New Directions in the Soviet Economy* (1966), and *Recent Trends in the Soviet Economy* (1970). For the mathematical approach to central planning, see L.V. KANTOROVICH, *The Best Use of Economic Resources* (1965; orig. pub. in Russian, 1959); V.V. NOVOZHILOV, *Problems of Cost-Benefit Analysis in Optimal Planning* (1970; orig. pub. in Russian, 1967); A. ZAUBERMAN, *Aspects of Planometrics* (1967); and J.P. HARDT *et al.* (eds.), *Mathematics and Computers in Soviet Economic Planning* (1967). Works on planning in central and eastern Europe include: A. ZAUBERMAN, *Industrial Progress in Poland, Czechoslovakia and East Germany, 1937–62* (1964); OTA SIK, *Plan and Market Under Socialism* (1967), a critique of central planning by the former head of the Czechoslovak planning commission; M.C. KASER and J.G. ZIELINSKI, *Planning in Eastern Europe* (1970). For planning in Yugoslavia, see B.N. WARD, *The Socialist Economy* (1967); T. WILSON and G.R. DENTON, *Economic Reform in Yugoslavia* (1968); and J.T. CRAWFORD, "Yugoslavia's New Economic Strategy: A Progress Report," in studies published by the Joint Economic Committee, Congress of the United States, *Economic Developments in Countries of Eastern Europe* (1970). A general comparative survey is N. SPULBER, *Socialist Management and Planning* (1971).

Developing countries: The best introduction to economic planning in the developing countries is W. ARTHUR LEWIS, *Development Planning: The Essentials of Economic Policy* (1966). It explains with arithmetical illustrations the various stages of drawing up a consistent development plan, but rightly emphasizes that economic policies are more important than formal planning techniques. ALBERT WATERSTON, *Development Planning: Lessons of Experience* (1965), gives a well-documented survey of development plans in various countries and is particularly good on the administrative and fiscal problems of development planning. A useful general survey of criteria for allocating investment resources is in G.M. MEIER, *Leading Issues in Economic Development*, 2nd ed., sect. 6 (1970). The general theoretical issues of balanced growth and the critical minimum size of development programs are discussed in H. MYINT, *The Economics of the Developing Countries* (1964). W.F. STOLPER, *Planning Without Facts* (1966), combines a case study of development planning in Nigeria with a pragmatic general approach to development planning.

(A.No./Jo.Hac./H.My.)

Economics

No one has ever succeeded in neatly defining the scope of economics. Economists used to say, with Alfred Marshall, the great English economist, that economics is "a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of wellbeing"—ignoring the fact that sociologists, psychologists, and anthropologists frequently study exactly the same phenomena. Another English economist, Lionel Robbins, has more recently defined economics as "the science which studies human behaviour as a relationship between (given) ends and scarce means which have alternative uses." This definition—that economics is the science of economizing—captures one of the striking characteristics of the economist's way of thinking but leaves out the macroeconomic approach to the subject, which is concerned with the economy as a whole.

What economists do

Difficult as it may be to define economics, it is not difficult to indicate the sort of questions that economists are concerned with. Among other things, they seek to analyze the forces determining prices—not only the prices of goods and services but the prices of the resources used to produce them. This means discovering what it is that governs the way in which men, machines, and land are combined in production and that determines how buyers and sellers are brought together in a functioning market. Prices of various things must be interrelated in a certain manner; how does such a "price system" or "market mechanism" hang together, and what are the conditions necessary for its survival?

These are questions in what is called "microeconomics," the part of economics that deals with the behaviour of such individuals as consumers, business firms, traders, and farmers. The other major branch of economics is "macroeconomics," in which the focus of attention is on aggregates: the level of income in the whole economy, the volume of total employment, the flow of total investment, and so forth. Here the economist is concerned with the forces determining the income of a nation or the level of total investment; he seeks to learn why full employment is so rarely attained, even in industrialized countries, and what public policies should be followed to achieve higher employment or more stability.

But these still do not exhaust the range of problems that economists consider. There is also the important field of "development economics," which examines the attitudes and institutions supporting economic activity as well as the process of development itself. The economist is concerned with the factors responsible for self-sustaining economic growth and with the extent to which these factors can be manipulated by public policy.

Cutting across these three major divisions in economics are the specialized fields of public finance, money and banking, international trade, labour economics, agricultural economics, industrial organization, and others. Economists may be asked to assess the effects of governmental measures such as taxes, minimum-wage laws, rent controls, tariffs, changes in interest rates, changes in the government budget, and so on.

The size of the profession

In the 19th century, economics was the hobby of gentlemen of leisure and the vocation of a few academicians; economists wrote about economic policy but were rarely consulted by legislatures before decisions were made. Today, there is hardly a government, international agency, or large corporation that does not have its resident economist. According to an estimate of the National Science Foundation (U.S.), for instance, there were 11,000 economists in the United States in 1966. Clearly, much depends on how one defines the job of an economist: the list of the National Science Foundation is confined to persons whose chief competence is in any one of the recognized economic specialities. Of the 11,000 professional economists, about 4,500 were employed as teachers of economics; the rest worked in various research or advisory capacities, either for themselves, for industry, or for government. This leaves out of account many others employed in accounting, commerce, marketing, and business administration; they may think of themselves as economists, but their professional expertise falls within other fields. There are perhaps another 10,000 economists in the rest of the world—their numbers have never been counted. It would be reasonable to estimate the total number of professional economists in the world in 1970 at 20,000, a number that was apparently growing at about 5 percent per year. There were about 75 English-language journals in economics and another 25 in various foreign languages, with new ones appearing every year. This implies the publication of about 1,500 scientific papers per year, not to mention the 700 new books on economics published every year. This is indeed "the age of economists," and the demand for their services seems insatiable.

MAIN CURRENTS IN THE DEVELOPMENT OF ECONOMIC THOUGHT

The work of Adam Smith

The effective birth of economics as a separate discipline may be traced to the year 1776, when the Scottish philosopher Adam Smith published *An Inquiry into the Nature and Causes of the Wealth of Nations.* There was, of course, economics before Adam Smith: the Greeks made significant contributions and so did the medieval scholastics; from the 15th to the 18th century, an enormous pamphlet literature appeared that developed the implications of economic nationalism, a body of thought now known as "mercantilism"; for a brief period in the 18th century the French "physiocrats" developed a fairly sophisticated economic model; and several other 18th-century figures can compete with Smith for the title of "first economist." Nevertheless, Adam Smith wrote the first full-scale treatise on economics and, by his magisterial influence, founded what later generations were to call the "English School of Classical Political Economy."

Analysis of the market. The *Wealth of Nations,* as its title suggests, is essentially a book about economic development and about policies that promote or hinder development. In its practical aspects it is an attack on the protectionist doctrines of the mercantilists and a brief for free trade. But in the course of attacking "false doctrines of political economy," Adam Smith was led to analyze the workings of a free-enterprise system as a governor of human activity. In a competitive market each individual, being one among many, can exert only a negligible influence on prices; each must take prices as they come and is free only to vary the quantities bought and sold at given prices; yet the sum of all individuals' separate actions determines prices. The "invisible hand" of the market, as Adam Smith was fond of saying, assures a social result that is independent of individual intentions and thus creates the possibility of an objective science of economic behaviour. Adam Smith believed that he had found, in the competitive market, an instrument capable of converting "private vices" (like selfishness) into "public virtues" (like maximum production). But this is only true if the competitive system is embedded in an appropriate legal and institutional framework, an insight that Adam Smith developed at length but that was largely forgotten by later generations. Within this great tome on the theme of rich and poor nations was contained a simple theory of value (or prices), a crude theory of distribution, an even cruder theory of international trade, and a primitive theory of money; but with all their imperfections, these were the building blocks of classical and modern economics. The book's very fecundity gave it strength because it left so much for disciples to tidy up.

The later classical economists

Construction of a system. David Ricardo's *Principles of Political Economy and Taxation* (1817) was, in one sense, simply a critical commentary on the *Wealth of Nations;* in another sense, it gave an entirely new twist to the developing science of political economy. Ricardo invented the concept of the "economic model," a tightly knit logical apparatus consisting of a few strategic variables, an apparatus that was capable of yielding, after a bit of manipulation, results of enormous practical im-

port. At the heart of the Ricardian system is the notion that economic growth must sooner or later be arrested, owing to the rising cost of growing food on a limited land area. An essential ingredient of this argument is the Malthusian principle—enunciated in Thomas Malthus' *Essay on Population* (1798)—that population tends to increase up to the limits set by the existing supply of food, thus holding down wages. As the labour force increases, extra food to feed the extra mouths can be produced only by extending cultivation to less fertile soil or by applying capital and labour to land already under cultivation (with diminishing results because of the so-called law of diminishing returns). Although wages are held down, profits do not rise proportionately because tenant farmers outbid each other for superior land. The chief beneficiaries of economic progress, therefore, are the landowners.

Since the root of the trouble, according to Ricardo, is the declining yield of wheat per unit of land, one obvious solution is to import cheap wheat from other countries. Eager to show that Britain would benefit from specializing in manufactured goods and exporting them in return for food, Ricardo hit upon the "law of comparative costs" as proof. He assumed that, within countries, labour and capital are free to move in search of the highest returns; between countries, however, they are not. In these circumstances, Ricardo showed, the benefits of trade are determined by a comparison of costs *within* each country, rather than by a comparison of costs *between* countries. It pays a country to specialize in the production of those goods that it can produce *relatively* more efficiently and to import everything else; although India may be able to produce everything more efficiently than England, India is nevertheless well advised to concentrate its resources on textiles, in which its efficiency is relatively greater, and to import British capital goods. The beauty of the argument is that if all countries take full advantage of the territorial division of labour, total world output is certain to be larger than it will be if some or all countries try to become self-sufficient. Ricardo's law became the fountainhead of 19th-century free-trade doctrine, which would have been enough, if he had said nothing else, to give him a place in the economists' pantheon.

Ricardo's influence

The influence of Ricardo's treatise was felt almost as soon as it was published, and for over half a century the Ricardian system dominated economic thinking in Britain. In 1848 John Stuart Mill's restatement of Ricardo's thought in his *Principles of Political Economy* brought it new authority. After 1870, however, most economists turned their backs on the range of problems that had concerned Ricardo and began to re-examine the foundations of the theory of value; that is, they became interested in the theory of why goods exchange at particular prices, so that for a while they devoted almost all of their efforts to the problem of resource allocation under conditions of perfect competition.

Marxism. A few words must first be said, however, about the last of the classical economists, Karl Marx. The first volume of *Das Kapital* appeared in 1867; the second and third after his death, in 1883 and 1894. For a generation, therefore, the competitive market theorists jostled with the followers of Marx. By 1900 the intellectual battle was over, and thereafter professional economists largely lost interest in Marx. Despite the Russian Revolution, despite what amounts to official endorsement of Marxism in one-third of the world, and despite the lingering influence of Marx's ideas, Marxian economics has been moribund ever since Marx's death in 1881. If Marx may be called "the last of the classical economists," it is because to a large extent he found his economics not in the real world but in the teachings of Smith and Ricardo. They had espoused a "labour theory of value," which holds that products exchange roughly in proportion to the labour costs incurred in producing them. Marx worked out all the logical implications of this theory and added to it "the theory of surplus value," which rests on the axiom that human labour alone creates all value and hence constitutes the sole source of profits. It is an axiom in the sense that it cannot be established in terms of the theory itself: it must be imported from without. To say that an economist is a Marxian economist is in effect to say that he shares the value judgment that it is socially undesirable for some people in the community to derive their income merely from the ownership of property. Since few professional economists in the 19th century accepted this ethical postulate and most were indeed inclined to find some social justification for the existence of private property and the income derived from it, Marxian economics fell on deaf ears. The Marxian system, moreover, culminated in three great generalizations: the tendency of the rate of profit to fall, the growing impoverishment of the working class, and the increasing severity of business cycles, of which the first is the linchpin of all the others. Marx's exposition of the "law of the declining rate of profit" is invalid; with it all of Marx's other predictions fall to the ground. In addition, Marxian economics had little to say on some of the practical problems that are the bread and butter of economists in any society. This is enough to suggest why Marxian economics failed to make many converts among academic economists. Marxists will reply that the reason is simply that academic economists have always been "lackeys of the capitalist class." Perhaps so, but the fact remains that Marx has had virtually no effect on modern economic thought.

The three generalizations of Marxism

The marginalists. The marginal revolution was essentially the work of three men: Stanley Jevons, an Englishman; Anton Menger, an Austrian; and Léon Walras, a Frenchman. Their contribution was the replacement of the labour theory of value by the marginal utility theory of value; their explanation of prices began with the behaviour of consumers in choosing among increments of goods and services (see UTILITY AND VALUE, ECONOMIC). The idea of emphasizing the marginal or last unit proved in the long run to be more significant than the introduction of utility. It was the consistent application of marginalism that marks the true dividing line between classical theory and modern economics. The classical political economists saw the economic problem as that of predicting the effects of changes in the quantity of capital and labour on the rate of growth of national output. The marginal approach, however, focussed on the conditions under which these factors tend to be allocated with optimal results among competing uses—optimal in the sense of maximizing consumers' satisfactions.

Throughout the last three decades of the 19th century, the English, Austrian, and French contributors to the marginal revolution largely went their own way. The Austrian school dwelt on the importance of utility as the determinant of value and vehemently attacked the classical economists as completely outmoded. A brilliant second-generation Austrian economist, Eugen von Böhm-Bawerk, applied the new ideas to the determination of the rate of interest, putting his stamp for all time on capital theory (see CAPITAL AND INTEREST). The English school, led by Alfred Marshall, sought a reconciliation with the doctrines of the classical writers. The classical authors, Marshall argued, concentrated their efforts on the supply side in the market; marginal utility theory was concerned with the demand side, but prices are determined by both supply and demand, just as a pair of scissors cuts with both blades. Marshall, seeking to be practical, applied his "partial equilibrium analysis" to particular markets and industries.

The leading French marginalist was Léon Walras, who carried the approach furthest by describing the economic system in general mathematical terms. For each product there is a "demand function" that expresses the quantities of the product that consumers demand as depending on its price, the prices of other related goods, the consumers' incomes, and their tastes. For each product there is also a "supply function" that expresses the quantities producers will supply as depending on their costs of production, the prices of productive services, and the level of technical knowledge. In the market, for each product there is a point of "equilibrium"—analogous to the equi-

librium of forces in classical mechanics—at which a single price will satisfy both consumers and producers. It is not difficult to analyze the conditions under which equilibrium is possible for a single product. But equilibrium in one market depends on what happens in other markets (a "market" in this sense being not a place or location but a complex of transactions involving a single good), and this is true of every market. There are literally millions of markets in a modern economy, and therefore "general equilibrium" involves the simultaneous determination of partial equilibria in all markets. Walras' efforts to describe the economy in this way led the historian of economic thought Joseph Schumpeter to call his work "the Magna Carta of economics." Walrasian economics is undeniably abstract, but it provides an analytical framework for incorporating all of the elements of a complete theory of the economic system. It is not too much to say that nearly the whole of modern economics is Walrasian economics. Certainly, modern theories of money, of employment, of international trade, and of economic growth are all Walrasian general equilibrium theories in a simplified form.

The neoclassical era

The years between the publication of Marshall's *Principles of Economics* (1890) and the Great Crash in 1929 may be described as years of reconciliation, consolidation, and refinement. The three national schools gradually coalesced into a single mainstream. The theory of utility was reduced to an axiomatic system that could be applied to the analysis of consumer behaviour under various circumstances, such as a change in income or price. The concept of marginalism in consumption led eventually to the idea of marginal productivity in production, and with it came a new theory of distribution in which wages, profits, interest, and rent were all shown to depend on the "marginal value product" of a factor. Marshall's concept of "external economies and diseconomies" was developed by his leading pupil, Alfred Pigou, into a far-reaching distinction between private costs and social costs, thus laying the basis of welfare theory as a separate branch of economic inquiry. There was a gradual development of monetary theory, which explains how the level of all prices is determined as distinct from the determination of individual prices, notably by the Swedish economist Knut Wicksell. In the 1930s the growing harmony and unity of economics was rudely shattered, first by the simultaneous publication of Edward Chamberlin's *Theory of Monopolistic Competition* and Joan Robinson's *Economics of Imperfect Competition* in 1933 and then by the appearance of John Maynard Keynes's *General Theory of Employment, Interest and Money* in 1936.

The critics. Before going on, it is necessary to take note of the rise and fall of the German Historical school and the American Institutionalist school, which levelled a steady barrage of critical attacks on the orthodox mainstream. The German historical economists, who had many different views, basically rejected the idea of an abstract economics with its supposedly universal laws; they urged the necessity of studying concrete facts in national contexts. While they gave impetus to the study of economic history, they failed to persuade their colleagues that their method was invariably superior. The institutionalists are more difficult to categorize. "Institutional economics," as the term is narrowly understood, refers to a movement in American economic thought associated with such names as Thorstein Veblen, Wesley Clair Mitchell, and John R. Commons. These writers had little in common aside from their dissatisfaction with the abstract theorizing of orthodox economics, its tendency to cut itself off from the other social sciences, and its preoccupation with the automatic market mechanism. They failed to develop a theoretical apparatus that would replace or supplement the orthodox theory. This may explain why the phrase "institutional economics" has become little more than a synonym for "descriptive economics." The hope that institutional economics would furnish a new interdisciplinary social science proved stillborn. (This is perhaps not surprising, because it was by abstracting purely economic forces from the totality of social interactions that economics got so far ahead of the other social sciences in theoretical rigour.) Although there is no longer an institutionalist movement in economics, the spirit of institutionalism is alive in such works as the Harvard economist John Kenneth Galbraith's *The Affluent Society* (2nd ed., 1969) and *The New Industrial State* (1967).

Returning to the innovations of the 1930s, the theory of monopolistic or imperfect competition remains somewhat controversial to this day. The older economists had devoted all their attention to two extreme types of market structure, that of "pure monopoly," in which a single seller controlled the entire market for one product, and that of "pure competition," characterized by many sellers, highly informed buyers, and a single, standard product. The theory of monopolistic competition gave recognition to the range of market structures that lie between these extremes, including (1) markets having many sellers with "differentiated products," employing brand names, guarantees, and special packaging that cause consumers to regard the product of each seller as unique; (2) "oligopoly," markets dominated by a few large firms; and (3) "monopsony," markets with a single monopolistic buyer and many sellers (see MONOPOLY AND COMPETITION). The theory produced the powerful conclusion that competitive industries in which each seller has a partial monopoly because of product differentiation will tend to have an excessive number of firms, all charging a higher price than they would if the industry were perfectly competitive. Since product differentiation—and the associated phenomenon of advertising—seems to be characteristic of most industries in developed capitalist economies, the new theory was immediately hailed as injecting a healthy dose of realism into orthodox price theory. Unfortunately, its scope was not great enough. It failed to provide a satisfactory theory of price determination under conditions of oligopoly. In advanced economies many of the manufacturing industries are oligopolistic. The result has been to leave a somewhat undigested lump at the centre of modern price theory, a constant reminder of the fact that economists still lack an adequate explanation of the conditions under which the giant firms of rich countries conduct their affairs.

Keynesian economics. The second major breakthrough of the 1930s, the theory of income determination, was primarily the work of one man—John Maynard Keynes. Keynes asked questions that in some sense had never been asked before; he was interested in the level of national income and the volume of employment rather than in the equilibrium of the firm or the allocation of resources. It was still a problem of demand and supply, but "demand" here means the total level of effective demand in the economy, and "supply" means the nation's capacity to produce. When effective demand falls short of productive capacity, the result is unemployment and depression; when it exceeds the capacity to produce, the result is inflation. The heart of Keynesian economics consists of an analysis of the determinants of effective demand. If one ignores foreign trade, effective demand consists essentially of three spending streams: consumption expenditures, investment expenditures, and government expenditures, each of which is independently determined. Keynes attempted to show that the level of effective demand so determined may well exceed or fall short of the physical capacity to produce goods and services: that there is no automatic tendency to produce at a level that results in the full employment of all available men and machines. This fundamental implication of the theory came as something of a shock to exponents of the traditional economics who had been inclined to take refuge in the assumption that economic systems tend automatically to full employment. By keeping his attention focussed on macroeconomic aggregates, like total consumption and total investment, and by a deliberate simplification of the relations between these economic variables, Keynes achieved a powerful model that could be applied to a wide range of practical problems. His system subsequently underwent considerable refinement—some

have said that Keynes himself would hardly have recognized it—and became thoroughly assimilated into the body of received doctrine (see INCOME AND EMPLOYMENT THEORY). Still, it is not too much to say that Keynes is perhaps the only economist to have added something really new to economics since Walras and perhaps since Ricardo.

Keynesian economics as conceived by Keynes was entirely "static"; that is, it did not involve time as an important variable. But a disciple of Keynes, Roy Harrod, soon developed a simple macroeconomic model of a growing economy; in 1948 he published *Towards a Dynamic Economics,* launching an entirely new speciality, "growth theory," which absorbed the attention of an increasing number of economists.

Developments after World War II

Mathematical economics. In the 25-year period following World War II, economics was so totally transformed that those who studied it before the war might as well have lived in another world. First of all, there was an enormous increase in the use of mathematics, which came to permeate virtually every branch of economics. Previously, few economists had made use of mathematics other than differential and integral calculus. Matrix algebra became important with the advent of "input–output analysis," an empirical method of reducing the technical relations between industries to a manageable system of simultaneous equations; it was an attempt to put quantitative flesh on the bones of a general equilibrium model of the economy. A closely related phenomenon was the development of "linear programming" and "activity analysis," which opened up a whole host of industrial problems to numerical solution and introduced economists for the first time to the mathematics of "inequalities" rather than exact equations. Likewise, the emergence of "growth economics" promoted the use of difference and differential equations.

Hand in hand with the spread of mathematical economics went an increasing sophistication of empirical work under the rubric of "econometrics," comprising a combination of economic theory, mathematical model building, and statistical testing of economic predictions. The development of econometrics had an impact on economics in general, since those who formulated new theories began to cast them in terms that allowed them to be empirically tested.

The postwar years also saw a renewal of interest in the underdeveloped countries. Economists became aware that they had too long neglected "an inquiry into the causes of the wealth of nations." There was also a conviction that economic planning of one variety or another was needed to close the gap between the rich and poor countries. Out of these concerns came the field of development economics. Regional economics, urban economics, health economics, and the economics of education are other offshoots of the mainstream since 1945.

The postwar tendencies in economic thought were best exemplified, not by the emergence of new techniques or by the addition of new parts to the economics curriculum but by the disappearance of divisive "schools," by the increasingly standardized professional training of economists all over the world, and by the transformation of the science from a rarefied academic exercise to an operational discipline geared to practical advice. This transformation brought prestige to the profession but also a new responsibility: now that economics really mattered, economics had to reckon with the conflict that so often exists between analytical precision and economic relevance.

The question of relevance was at the centre of a "radical critique" of economics that developed along with the campus revolts of the late 1960s. The radical critics declared that economics had become a defense of the status quo and that its practitioners had joined the power elite. The marginal techniques of the economists, ran the argument, were profoundly conservative in their bias, and encouraged a piecemeal rather than a revolutionary approach to social problems; likewise, the tendency in theoretical work to ignore the everyday context of economic activity amounted in practice to the tacit acceptance of prevailing institutions. The critics said that economics should abandon its claim of being a value-free social science and address itself to the great questions of the day—those of civil rights, poverty, imperialism, and nuclear war—even at the cost of analytical rigour and theoretical elegance.

It is true that the study of economics encourages a belief in reform rather than revolution; economics as a science does not provide enough certitude for any thoroughgoing reconstruction of the social order. It is also true that most economists tend to be deeply suspicious of monopoly in all forms, including state monopolies, and to favour competition between independent producers as a way of diffusing economic power. Finally, most economists prefer to be silent on large questions if they have nothing to offer beyond the expression of personal preferences: most economists as economists are fundamentally concerned with the professional standards of their discipline, and this may mean in some cases frankly conceding that economics has as yet nothing very interesting to say about these questions. (It does not mean, however, that they desire to justify the status quo.) Yet the radical critique of modern economics was not to be lightly dismissed. The radical economists were posing issues that were important. At the very least it could do the economic researcher no harm to think about the social and political relevance of his work.

ECONOMICS AS A SCIENCE

The role of value judgments

Economists are sometimes confronted with the charge that their discipline is not a science. Human behaviour, it is said, cannot be analyzed with the same objectivity as the behaviour of atoms and molecules. Value judgments, philosophical preconceptions, and ideological biases must interfere with the attempt to derive conclusions that are independent of the particular economist espousing them. Moreover, there is no laboratory in which economists can test their hypotheses.

This argument raises issues for all of the social sciences. Only a very general reply can be given here. Economists are wont to distinguish between "positive economics" and "normative economics." Positive economics seeks to establish facts: Will a subsidy to butter producers lower the price of butter? Will a rise in wages in the automobile industry reduce the employment of automobile workers? Will devaluation improve the balance of payments? Does monopoly foster technical progress? Normative economics, on the other hand, is not concerned with matters of fact but with questions of policy, of "good" or "bad": Should the goal of price stability be sacrificed to that of full employment? Should income be taxed at a progressive rate? Should there be legislation in favour of competition?

Positive economics in principle involves no judgments of value; its findings may be as impersonal as those of astronomy and meteorology, two natural sciences that are also denied the advantage of conducting laboratory experiments. As the British philosopher David Hume argued 200 years ago, there is no logical way to deduce "ought" from "is" or prescriptions from descriptions; all statements of fact are ethically neutral. In that sense a value-free economics is possible (at least in principle): if economics is about the application of means to achieve given ends, there would seem to be no reason why one cannot analyze the allocation of means to achieve *any* end. This is not to deny that most of the interesting economic propositions involve the addition of definite value judgments to a body of established facts, that ideological bias creeps into the very selection of the questions that economists investigate, that what is a means from one point of view may be an end from another, nor even that much practical economic advice is loaded with concealed value judgments, the better to persuade rather than merely to advise. This is only to say that economists are human. The commitment of economists the world over to the ideal of value-free positive economics (or to the candid declaration of personal values in normative eco-

nomics) serves as a defense against the attempts of special interests to bend the science to their own purposes. The best assurance against bias on the part of any particular economist is the criticism of other economists. The best protection against special pleading in the name of science is the professional standards of scientists.

Methods of inference. But how, one may ask, are facts established in a science that cannot conduct experiments? In essence, the answer is: by means of statistical inference. Economists typically begin by describing the area under study according to what they feel to be important. Then they construct a "model" of the real world, deliberately repressing some of its features and emphasizing others; they abstract, isolate, and simplify, thus imposing order on a world that at first glance appeared disorderly. Having evolved an admittedly unrealistic representation of the real world, they then manipulate the model by a process of logical deduction, arriving eventually at some prediction or implication that is of general significance. At this point, they return to the real world to see whether or not the prediction is borne out by observed events.

But the observable events that are available to test a theory never exhaust the population of all such events: they are merely a sample of it. This raises the problem of statistical inference; namely, what can be inferred about a population from a sample of the population? The theory of statistical inference is simply an agreed-upon procedure for making such inferences, but in the nature of the case it never succeeds in removing all elements of judgment from an inference. Thus the empirical truths of economics are invariably surrounded by a band of doubt, and economists speak of them as "probable" or "likely"; they are propositions in which economists have "a certain degree of confidence" because it is unlikely that they could have come about by chance.

It follows that judgments are at the heart of both positive and normative economics. It is easy to see, however, that judgments about "degrees of confidence" and "statistical levels of significance" are of a totally different order from those that crop up in normative economics. When men say that every individual should be allowed to spend his income as he likes, that people should not be free to control material resources and to employ others, or that governments must offer relief for the victims of inexorable economic forces, they are making the kind of value judgments that laymen have in mind when they accuse economists of producing personal preferences in the guise of scientific conclusions. There is no room for such value judgments in positive economics.

Testing theories. In the past some economists tended to claim too much for their propositions. Economic models were said to be based on fundamental axioms and premises about economic behaviour that were absolutely true a priori because they were derived from an examination of one's own economic behaviour. Since the theorems of the model were deduced from these axioms by the laws of logic, the theorems also were held to be true a priori. Economic models did not need to be confronted with empirical evidence.

This extreme apriorist position may be contrasted with the ultraempiricist view, which holds that one must begin and end with observable facts; the latter approach, however, has never appealed to more than a small minority of economists. In the middle ground between these two sharply opposing views is the methodological position that has found increasing acceptance among modern economists. It argues that one must test the predictions or conclusions of a model but without worrying too much about the realism of its premises, axioms, or assumptions. Most assumptions in economic theory cannot be tested directly. For example, there is the famous assumption of price theory that businessmen strive to maximize profits. Attempts to find out whether they do, by asking them, usually fail; after all, businessmen are no more fully conscious of their own motives than other people are. A logical approach would be to observe businessmen in action. But that would require knowing what sort of action is associated with profit maximizing, which is to say that one would have drawn out all the implications of a profit maximizing model. Thus one would be testing an assumption about business behaviour by comparing the predictions of a theory of the firm with observations from the real world.

The problem of validity

This is not as easy as it sounds. Since the predictions of economics partake of the nature of probability statements, there can be no such thing as a conclusive, once-and-for-all test of an economic hypothesis. The science of statistics cannot prove any hypothesis, it can only fail to disprove it. A theory that survives a statistical test is not true as such; it is only provisionally true because it has so far resisted all attempts to falsify it. Attempts to falsify economic hypotheses never yield unambiguous results, and hence economic theories tend to survive until they are falsified repeatedly with new or better data. This is not because they are *economic* theories but because the attempt to compare predictions with outcomes in the social sciences is always limited by the rules of statistical inference.

It is not remarkable that competing theories exist to explain the same phenomena, with economists disagreeing as to which theory is to be preferred. While virtually all economists today agree that theories should be submitted to empirical testing and that the theory is to be preferred that allows predictions that conform, in a probabilistic sense, most closely to observable events, this precept can be very difficult to apply in practice. There have been periods in the history of economics when there was overwhelming agreement in the profession as to which models or theories were "true." But a period of consensus may be followed by a generation of doubt until a new departure is made that succeeds in producing a new consensus. In this, economics is not very different from physics.

Much has been written about the doubtful accuracy of economists' predictions. Of course, economists cannot foretell the future as such; only soothsayers do that. Economists can foretell the effects of specific changes in the economy, but they are better at predicting the direction than the actual magnitude of events. When economists predict that a tax cut will raise national income, one may be confident that the prediction is accurate; when they predict that it will raise national income by a certain amount in three years, however, the forecast is likely to miss the mark. The reason is that most economic models do not contain any explicit reference to the passage of time and hence have little to say about how long it takes for a certain effect to make itself felt. Short-period predictions generally fare better than long-period ones, in part because most economic models rely on propositions about the plans and intentions of economic agents, whereas the data on which the theories are tested are derived from past events. This is disappointing, but it does not mean that economics is not a science.

MICROECONOMICS

The firm and the household

Since Keynes, economic theory has been of two kinds: macroeconomics—or the study of the determinants of national income—and the traditional microeconomics. The latter approaches the economy as if it were made up only of business firms and households (ignoring governments, banks, charities, trade unions, and all other economic institutions) interacting in two kinds of markets—product markets and markets for productive services, or factor markets. Households appear as buyers in product markets and as sellers in factor markets, where they offer men, machines, and land for sale or hire. Firms appear as sellers in product markets and as buyers in factor markets. In each type of market, price is determined by the interaction of demand and supply, and the problem of microeconomic theory is to say something meaningful about the forces that make up demand and supply.

Theory of choice. At first it appears that all one can say is that everything depends on everything else. But firms and households do not behave in a vacuum. Firms face certain technical constraints in producing goods and services, and households have definite preferences for some products over others. It is possible to express the

technical constraints facing business firms by writing down a series of "production functions," one for each firm. A production function is simply a kind of equation that expresses the fact that the output of a firm depends on the quantity of inputs it employs and, in particular, that inputs can be technically combined in different proportions to produce a given level of output. A production engineer could calculate, on the basis of existing technical knowledge, the largest possible output that could be produced with every possible combination of inputs and in this way could define a boundary to the range of production possibilities open to a firm. By itself this does not tell how much the firm will produce or what mixture of products it will make or what combination of inputs it will adopt: these depend on the prices of products and the prices of inputs (or "factors of production"), which have yet to be determined. One may assume that the firm is motivated in a particular way: it wants to maximize profits, which are defined as the difference between the sales value of its output and the money outlays required to obtain its inputs. It will, therefore, select that combination of inputs that minimizes the costs of producing any given quantity of output and will select from the range of possible combinations of products that combination that maximizes its revenues. This is to say that it always tries to move along its production function, along the edge of the boundary of technical possibilities. But where it ends depends, in part, on the demand for its products. This leads to the part played by households in the system.

Households are endowed with definite "tastes" that can be expressed in a series of "utility functions," one for each household. A utility function is an equation like a production function, expressing the fact that the pleasure or satisfaction that households derive from consumption depends on the products that they purchase and on the various ways in which they combine these products in consumption to yield a given level of satisfaction. The utility function need not be specified in the same detail as a production function. One may think of it as a general description of the household's preferences between all the paired alternatives with which it will be confronted. Here, too, it is necessary to assume something about motivation to make any progress: the assumption is that households seek to maximize satisfaction, distributing their given incomes among available consumer goods in such a way as to derive the largest possible "utility" from consumption. Their incomes, however, remain to be determined.

Supply and demand curves

The purpose of production functions in economic theory is to provide an anchor in the bedrock of technology from which to derive the "supply curves" of firms in product markets and the "demand curves" of firms in factor markets. Similarly, the purpose of utility functions is to provide an anchor in subjective "tastes" from which to derive the "demand curves" of households in product markets and the "supply curves" of households in factor markets. All of these demand and supply curves express the quantities demanded and supplied as a function of prices, not because price is the only determinant of economic behaviour but because the purpose is to have a theory of price determination. Much of economic theory is devoted to showing how various production and utility functions, coupled with certain assumptions about behaviour, lead to demand and supply curves in which the quantity demanded is inversely related and the quantity supplied positively related to price. The figure depicts these relationships (curves would be just as suitable as straight lines).

Not all demand and supply curves look alike. The essential point is that most demand curves are negatively inclined, while most supply curves are positively inclined. This may seem a modest result for a great deal of effort, but the argument has powerful implications. The participants in a market will be driven automatically to the price at which the two curves intersect; this price p is called the "equilibrium" price or "market-clearing" price because it is the only price at which supply and demand are equal. If it is a market for butter, any change in the

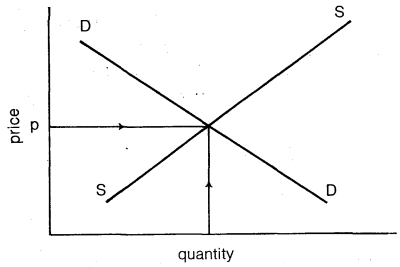


Illustration of the relationship of price to supply (S) and demand (D).

production function of dairy farmers or in the utility function of butter consumers or in the prices of cows, grassland, and milking equipment or in the incomes of butter consumers or in the prices of nondairy products that consumers buy can be shown to lead to definite changes in the equilibrium prices of butter and in the equilibrium quantity of butter produced. Better still, the effects of a government ceiling on the price of butter or of a tax on butter producers or of a price-support program for dairy farmers can be predicted with almost perfect certainty. As a rule, the prediction will refer only to the direction of change (the price will go up or down); but if the demand and supply curves of butter can be defined in quantitative terms on the basis of past data, one may be able to predict the actual magnitude of the change.

Theory of allocation. This analysis of the behaviour of firms and households is to some extent symmetrical: all economic agents are conceived of as ordering a series of attainable positions in terms of an entity they are trying to maximize. For a firm these attainable positions are essentially input combinations; for a household they are product combinations. From the maximizing point of view, some combinations are better than others; the best combination is called the "optimal" or "efficient" combination. The rule for efficient, optimum allocation may now be stated baldly: an optimum allocation is one that equalizes the returns of the marginal or last unit to be transferred between all the possible uses. In the theory of the firm, an optimum allocation of outlays among the factors of production implies that the "marginal physical product" of an additional dollar devoted to hiring the services of any one of the factors is the same for all factors; the so-called law of eventually diminishing marginal productivity, a property of a wide range of production functions, ensures that such an optimum exists. In the theory of consumer behaviour an optimum situation obtains when the consumer has distributed his given income in such a way that the "marginal utility" of each additional dollar spent on any of the products purchased is equal for all products; the "law of eventually diminishing marginal utility," a property of a wide range of utility functions, ensures that such an optimum exists. These are merely particular examples of the "equimarginal principle," which is not only at the core of the theory of the firm and the theory of consumer behaviour but also underlies the theory of money, of capital, and of international trade. In fact, the whole of microeconomics is nothing more than the spelling out of this principle in ever wider contexts.

The equimarginal principle is, of course, applicable to any decision that involves alternative courses of action. Economics furnishes a technique for thinking about decisions, whatever their character and whosoever makes them. Military planners may, for example, consider a variety of weapons in the light of a single objective, that of damaging an enemy; some of the weapons are effective against the enemy's army, some against the enemy's navy, and some against his air force; the problem is to find an optimal allocation of the defense budget, one that equalizes the marginal contribution of each type of weapon. But defense departments rarely have single objectives; along with maximizing damage to an enemy there may be another objective, such as minimizing losses

from attacks. In that case, more than the equimarginal principle is needed for a decision; it is necessary to know how the department ranks the two objectives in order of importance, since different rankings will imply different optima. But a ranking of objectives is simply a utility function or a preference function.

In other words, when an institution pursues multiple ends, decisions about how to achieve them require a weighting of the ends. Every decision involves a "production function"—a statement of what is technically feasible—and a "utility function"; the equimarginal principle is then invoked to provide an efficient, optimal strategy. This applies just as well to the running of hospitals, churches, and schools as to the conduct of a business enterprise, to the location of an international airport as well as to the design of a development plan for an underdeveloped country. This is why economists crop up in what seem to be the most unlikely places, advising on activities that are obviously not being conducted for economic reasons.

MACROECONOMICS

The different approach of macroeconomics

There is, however, an approach to economics in which the foregoing considerations do not apply. That is the field known as macroeconomics. In macroeconomics one is concerned with the aggregate outcome of individual actions. Keynes's "consumption function," for example, which relates aggregate consumption to national income, is not built up from individual consumer behaviour; it is simply an empirical generalization. The focus is on income and expenditure flows rather than on markets. Purchasing power flows through the system from business investment to consumption, but it leaks out at two places in the form of personal and business savings. Counterbalancing the savings are investment expenditures in the form of new capital goods, houses, and so forth, which constitute a source of new injections of purchasing power in every period. Since savings and investments are carried out by different people for different motives, there is no reason why "leakages" and "injections" should be equal in every period. If they are not equal, national income, the sum of all income payments to the factors of production, will rise or fall in the next period. When planned savings equal planned investment, income will be at an equilibrium level, that is, a level at which it can sustain itself; when the plans of savers do not match those of investors, the level of income will go on changing until the two do match. One can complicate this simple model by making investment a function of the interest rate; by introducing the government budget, the money market, labour markets, imports and exports, foreign investment; and so forth. But all this is far removed from the problem of resource allocation and from the maximizing behaviour of individual economic agents.

The result is a kind of intellectual schizophrenia in which the techniques of microeconomics do not carry over fully into macroeconomics and vice versa. This is widely held to be an unsatisfactory state of affairs; economists have in recent years sought to build a bridge between the individual consumer and the overall consumption function and between the individual investor and the behaviour of aggregate investment. Nevertheless, the bridge remains incomplete, and the student of economics must be prepared to work with two boxes of tools.

FIELDS OF CONTEMPORARY ECONOMICS

The following is a bird's-eye view of the main fields of contemporary economics.

Public finance. Since the time of Ricardo economists have been concerned with the incidence of taxes, that is, with determining who it is that really pays a tax. If a corporation faced with a profits tax reacts by raising its prices, it may succeed in making the consumer pay the tax; on the other hand, if sales decline as a result of the rise in price the firm may have to lay off some of its workers, and the burden of the tax will be shared by consumers, wage earners, and shareholders. This simple example shows how complex may be the actual incidence of a tax. A large part of the literature of public finance in the 19th century was devoted to such problems.

Taxes and government expenditures

Keynesian economics brought new dimensions to public finance: the older preoccupation with tax incidence gave way to the analysis of the impact of government expenditures on the level of income and employment (see FISCAL AND MONETARY POLICY). It was some time, however, before economists realized that they lacked a theory of government expenditures, that is, a set of criteria for determining what activities should be supported by governments and what should be the relative expenditure on each. One of the most exciting recent developments in the field of public finance is the attempt to devise such criteria. Decisions on public expenditures have proved to be susceptible to much of the traditional analysis of microeconomics. In the 1960s there developed a technique known as "cost–benefit analysis," which tries to appraise all of the economic costs and benefits, direct and indirect, of a particular activity so as to decide how to distribute a given public budget most effectively between different activities. This technique has been applied to everything from the construction of hydroelectric dams to the control of tuberculosis. Its exponents hope that the same type of analysis that has proved so fruitful in the past in analyzing individual choice may also succeed with problems of social choice.

Money. One of the oldest, widely accepted functions of government is control over the supply of money. The dramatic effects of changes in the quantity of money on the level of prices and the volume of economic activity were recognized and thoroughly analyzed in the 18th century, and monetary economics has ever since constituted one of the principal branches of economics. In the 19th century a complex and somewhat crudely formulated tradition grew up known as the "quantity theory of money," which held that any change in the supply of money can only be absorbed by variations in the general level of prices (the purchasing power of money). In consequence, prices will tend to change proportionately with the quantity of money in circulation. As the growth of fiat paper money gave governments increasingly effective control over the stock of circulating media, the quantity theory of money supplied an apparently simple rationale for the management of the economy: all that was needed to prevent inflation or deflation was to vary the quantity of money in circulation inversely with the level of prices.

One of the targets of Keynes's attack on traditional thinking in his *General Theory of Employment, Interest and Money* was this quantity theory of money. Keynes produced a different theory of the demand for money that implied that the impact of a change in the stock of money on the level of national income is weak and at best indirect; the effect on prices is virtually nil, he maintained, at least in economies with heavy unemployment such as prevailed in the 1930s. He put his emphasis instead on government budgetary and tax policy and direct control of investment. As a consequence, economists lost faith in monetary management and came to regard monetary policy as more or less ineffective in controlling the volume of economic activity.

In the 1960s there was a remarkable revival of the older view, at least among a small but growing school of American monetary economists. They accepted much of Keynesian economics but argued that the effects of fiscal policy are unreliable unless the quantity of money is regulated at the same time. They refurbished the quantity theory of money and tested the new version on a variety of data for different countries and for different time periods, leading to the broad conclusion that the quantity of money does matter (see MONEY).

In the early 1970s the controversy was still raging. It is notable that this debate, unlike previous debates in the history of monetary economics, was characterized by disputes over empirical findings—that is, it was focussed on the testable character of different monetary theories rather than on the manner of their formulation. Progress was made slower by the political overtones of the con-

troversy: in some countries, belief in the efficacy of monetary policy had become a kind of litmus test of political conservatism. Nevertheless, a reconciliation between Keynesians and quantity theorists needed only some agreement as to the magnitude of monetary forces and the degree of stability of the demand for money. Monetary economics seemed at last to be coming of age as an empirical discipline.

International economics. The foundations of international economics were firmly established in the 19th century. The subject has consisted ever since of two distinct but connected parts: (1) the "pure theory of international trade," which seeks to account for the gains obtained from trade and to explain how these gains are distributed among countries, and (2) the "theory of balance-of-payments adjustments," which analyzes the workings of the foreign exchange market, the effects of alterations in the exchange rate of a currency, and the relations between the balance of payments and the level of economic activity.

In modern times, the Ricardian pure theory of international trade has been reformulated by the American Paul Samuelson, improving on the earlier work of two Swedish economists, Eli Heckscher and Bertil Ohlin. The so-called Heckscher–Ohlin theory explains the pattern of international trade as determined by the relative land, labour, and capital endowments of countries: a country will tend to have a relative cost advantage in goods requiring the intensive use of the country's relatively abundant factor of production (thus land-rich Canada exports wheat) and to import goods requiring the intensive use of the country's relatively scarce factor (thus capital-poor Canada imports American automobiles). This theory absorbs Ricardo's law of comparative costs but goes beyond it in linking the pattern of trade to the economic structure of trading nations. It implies that foreign trade is a substitute for international movements of labour and capital —which raises the intriguing question of whether or not foreign trade may work to equalize the prices of all factors of production in all trading countries. Whatever the answer, the Heckscher–Ohlin theory provides a model for analyzing the effects of a change in trade on the industrial structures of economies and, in particular, on the distribution of income between factors of production. Much of the recent effort of specialists in international economics has gone toward refining the Heckscher–Ohlin model and testing it on an ever wider range of empirical evidence.

Labour. Like monetary and international economics, labour economics is an old economic speciality. It gets its raison d'être from the peculiarities of labour as a commodity. Labour itself is not bought and sold; rather, its services are hired and rented out. But since people cannot be disassociated from their services, various nonmonetary considerations play a role in the sale of labour services as contrasted with the sale of machine time or the rental of land. Yet, the bulk of the literature in labour economics was until recently concerned solely with the demand side of the labour market. Wages, the textbooks all said, were determined by the "marginal productivity of labour," that is, by the relationships of production and by consumer demand. If the supply of labour came into the picture at all, it was merely to allow for the presence of trade unions; unions could only raise wages by limiting the supply of labour.

The economics of labour and industry

After a long period of neglect, the supply side of the labour market began, in the 1960s and 1970s, to attract the attention of economists. First, attention shifted from the individual worker to the household as a supplier of labour services; the increasing tendency of married women to enter the labour force and the wide disparities and fluctuations observed in the rate that females participate in a labour force drew attention to the fact that an individual's decision to supply labour is not independent of the size, age structure, and asset holdings of the household to which he or she belongs. Second, the new concept of "human capital"—that people make capital investments in their children and in themselves by incurring the costs of education and training, the costs of searching for better job opportunities, and the costs of migration to other labour markets—has served as a unifying explanation of the diverse activities of households in labour markets. In this way, capital theory (see CAPITAL AND INTEREST) has become the dominant analytical tool of the labour economists, replacing or supplementing the traditional theory of consumer behaviour. The economics of training and education, the economics of information, the economics of migration, the economics of health, and the economics of poverty are some of the by-products of this new perspective. A field that was at one time regarded as rather cut-and-dried has taken on new vitality.

Labour economics, old or new, has always regarded the explanation of wages as its principal task, including the factors determining the general level of wages in an economy and the reasons for wage differentials between industries and occupations. Wages are influenced by trade unions; the impact of their activities is of increased importance at a time when most governments manage the economy with one eye on the unemployment statistics. The prewar fears of chronic unemployment gave way to the postwar fears of chronic inflation at or near levels of full employment. In response to this a vast literature sprang up after 1945 analyzing the inflationary pressures stemming from both the supply side and the demand side of labour markets. Whether prices were being pushed up by the labour unions ("cost push") or pulled up by excess purchasing power ("demand pull") became the issues in this 20-year long debate on inflation, a controversy that is, of course, intimately related to the quarrels in monetary economics mentioned earlier.

Industrial organization. The principal concerns of this field are the structure of markets, public policy toward monopoly, the regulation of public utilities, and, of late, the economics of technical change.

The monopoly problem, or, more precisely, the problem of the maintenance of competition, does not fit well into the received body of economic thought. Economics started out, after all, as the theory of competitive enterprise, and even today its most impressive theorems require the assumption of numerous small firms, each having a negligible influence on price. Yet the typical market structure of manufacturing today is that of oligopoly—competition among the few—and some industries are dominated by firms so large that their annual sales volume exceeds the national income of the smaller countries of western Europe. It is tempting to leap to the conclusion that oligopoly is deleterious to economic welfare, on the ground that it leads to the misallocation of resources. But some economists, notably Joseph Schumpeter, have argued that economic growth and technical progress are brought about not through free competition but through large firms and the destruction of competition. According to this view, monopoly has its origin in the need of business firms to protect themselves from the risks associated with the introduction of new products, new techniques, and new methods of marketing. The giants compete not in price but in successful innovation, and this kind of competition has proved more effective for economic progress than the price competition described in orthodox textbooks.

Although this thesis smacks of *post hoc ergo propter hoc* ("after this, therefore because of this")—giant firms have prospered in rapidly growing economies; therefore, growth is due to giant firms—it makes the merits of "trust busting" and cartel dissolution somewhat less compelling. The question is what sort of competition is socially most desirable. If each of four or five large firms in an oligopolistic industry finds it necessary to compete in terms of the quality of its products and its research or by means of better technology and superior merchandising, the performance of the industry may well be more satisfactory than if it were reorganized into a price-competitive industry. But if the four or five giants settle down to a quiet life and concentrate their rivalry on sales promotion techniques, the verdict must be less favourable. One cannot, it seems, draw facile conclusions about

the competitive results of different market structures; it is necessary to approach the monopoly problem with a generous dose of pragmatism.

The reason why there is so much uncertainty in the economic discussion of policies toward big business is the lack of a general theory of oligopoly. There are dozens of special theories applying to special cases, but there is no single, organizing framework with testable implications about the behaviour of oligopolists in general.

Agriculture. Farming has long provided economists with their favourite example of a perfectly competitive industry. But with increasing government regulation of agriculture, it also provides striking examples of the effects of price controls, income supports, output ceilings, and marketing cartels. Agricultural economics commands attention wherever governments wish to stimulate farming or to protect farmers—which is to say, everywhere.

Agricultural economists generally have been closer to their subject matter than other economists. In consequence, more is known about the technology of agriculture, the nature of farming costs, and the demand for agricultural goods than is known about any other industry. The student of economics who wants to learn how to estimate a production function or a demand curve is well advised to go to the literature on agricultural economics.

The underdeveloped countries have furnished a new laboratory for agricultural economics. Many such countries have a "subsistence agriculture," in which the farmer produces mainly for his own family consumption and brings to market only what is left over. Subsistence farmers are only tenuously linked to the money economy. They are more reluctant than commercial farmers to take the risks entailed in experimenting with new seeds, fertilizers, and farming methods: given the vagaries of the weather, they prefer to operate on the basis of the worst weather that can be expected rather than the best or even the average. While the pessimism of subsistence farmers is perfectly rational, it makes the task of predicting their response to new prices or new methods doubly difficult. The economist must also recognize that every agricultural area presents its own production, processing, and marketing problems. To suggest changes that will raise agricultural productivity in these circumstances is no easy task, and agricultural economists often find themselves warning governments not to intervene too vigorously, since much intervention in the past has been shown to be wrong or inappropriate.

The problem of economic development in Africa, Asia, and Latin America centres on the agricultural sector; one of the abiding questions is how far industrialization can proceed without there first being an agricultural revolution. For this reason, if for no other, agricultural economics has a large future.

Growth and development. The study of economic growth and development is not a single branch of economics but falls, in fact, into two quite different fields. The two fields—"growth" and "development"—employ different methods of analysis and are indeed addressed to two distinct types of inquiry.

Development economics is easy to describe. It is one of the three major subfields of economics, the other two being microeconomics and macroeconomics. Development economics resembles economic history in that it seeks to explain the changes that take place in economic systems with the passage of time (see ECONOMIC DEVELOPMENT).

The subject of economic growth is not so easy to characterize. It is the most technically demanding field in the whole of modern economics, impossible to grasp for anyone who lacks differential calculus. Its focus is the properties of equilibrium paths, rather than equilibrium states. One makes a model of the economy and puts it into motion, requiring that the time paths described by the variables be self-sustaining in the sense that they continue to be related to each other in certain characteristic ways. Then one can investigate the way economics might approach and reach these steady-state growth paths from given starting points. Beautiful and frequently surprising theorems have emerged from this experience, but as yet there are no really testable implications nor even definite insights into how economies grow.

Growth theory began with the work of Roy Harrod in England and Evsey Domar in the United States. Their joint product has been known ever since as the Harrod–Domar model. Keynes had shown that new investment has a multiplicative effect on income and that the increased income generates extra savings to match the extra investment, without which the higher income level could not be sustained. One may think of this as being repeated from period to period, remembering that investment, apart from raising income disproportionately, also generates the capacity to produce more output that cannot be sold unless there is more demand, that is, more consumption and more investment. That is all there is to the model. It contains one behavioral condition—that people tend to save a certain proportion of extra income, a tendency that can be measured. It contains one technical condition—that investment generates additional output, a fact that can be established. And it contains one equilibrium condition—that planned saving must equal planned investment in every period if the income level of the period is to be sustained. Given these three conditions, the model generates a time path of income and even indicates what will happen if income falls off the path.

More complex models have since been built, incorporating different saving ratios for different groups in the population, technical conditions for each industry, definite assumptions about the character of technical progress in the economy, monetary and financial equations, and much more (see ECONOMIC GROWTH).

Mathematical economics. Differential calculus has long been the traditional tool of mathematical economics. Many economic problems, particularly in microeconomics, take the form of maximizing some variable (such as profits) subject to a constraint (such as the production function), for which calculus supplies the simplest technique. Traditionally it was applied to problems in comparative statics. These problems include so-called endogenous variables, the values of which are determined within the model, as well as constants that originate outside the model and are called "exogenous variables" or "parameters." The object is to discover the effects of changes in one or more of the parameters upon the equilibrium situation. The latter is a situation in which all of the endogenous variables are simultaneously in a state of rest. If the value of some of the parameters is changed, the result is a new equilibrium state.

The application of mathematical methods

Much economic analysis, even when it is expressed in words, is simply the method of comparative statics. But comparative statics has its limitations: it tells the investigator *where* the system will arrive, but it does not tell him *when* it will arrive or what will happen along the way; and it cannot tell him whether, once driven out of the way, it will ever get back to its destination. In other words, comparative statics ignores the process of adjustment from the old equilibrium state to the new one, and it entirely neglects the time element in that adjustment process. The study of this process of adjustment over time is called "economic dynamics," and one may think of it as the economics of disequilibria.

Just as differential calculus is the mathematics of comparative statics, difference and differential equations are the ideal tools for handling dynamic problems. Difference equations deal with time as a discrete variable—changing only from period to period—whereas differential equations treat time as a continuous variable; the choice between them is simply one of convenience. They enable one to ask such questions as: if the system is pushed out of equilibrium, perhaps because one of the parameters of the model has changed, will economic forces drive it toward a new equilibrium position or away from one, will the time path described by the endogenous variables be steady or fluctuating, and if fluctuating, will the move-

ments be damped down or will they increase and become explosive?

Economic dynamics is one of the newer developments of mathematical economics, and often it falls short of the ambitious demands made on it. Dynamic models, for example, are typically formulated in terms of linear equations, not because the world is linear but because nonlinear equations can be very difficult to solve. Likewise, the coefficients of difference and differential equations are usually taken to be constants, again for the sake of making the mathematics of the analysis manageable. This means that if the economic environment changes as the model runs its course, its predictions will be false. An abiding danger in all mathematical economics is the tendency to adopt economic assumptions for the sake of mathematical convenience. The way to meet this danger is for economists to acquire enough mathematical sophistication so that they will not be dazzled by displays of mathematical technique.

Econometrics. Like mathematical economics, econometrics is something economists do rather than a special area of interest. Econometrics refers to the study of empirical data by statistical methods, the purpose of which is the testing of hypotheses and the estimation of relationships suggested by economic theory. Whereas mathematical economics considers the purely theoretical aspects of economic analysis, econometrics attempts to falsify theories that are expressed in explicit mathematical terms. But frequently the two go together.

The classic technique for estimating an economic relationship is that of "least squares," which is a method of fitting a trend line to a scatter of observations that minimizes the square of the deviations of the observed points from the line. To take a simple example: the Keynesian theory assumes that consumers' expenditures depend principally on income; one may interpret this to mean that consumption depends *only* on income and then test the hypothesis by trying to fit a trend line to a series of observations of income and consumption over a period of time. In so doing, one is really saying that the observations that fall to either side of the line are due either to errors in measuring the variables or to errors in specifying the relationship between consumption and income. It is essential to the method of least squares that these "errors" be randomly distributed or at any rate distributed in known ways. When this condition is violated, least squares estimates are unreliable. It is sometimes difficult to tell with economic data just how the errors are randomly distributed, and it is precisely for this reason that an econometrician is needed rather than an ordinary statistician.

A still more significant trend in recent econometrics is the tendency to move from single-equation estimates (such as the relationship between consumption and income) to systems of simultaneous equations. While consumption depends on income, income also depends on consumption; this kind of interdependence requires two equations rather than one. More generally, most economic variables are the result of demand and supply forces that simultaneously determine quantities and prices. To estimate a demand curve for butter from a single-equation regression (by relating the price of butter to the quantities of butter consumed, the incomes of consumers, and the prices of near substitutes for butter) is likely to produce a biassed answer because the price of butter is also influenced by supply conditions in the dairy industry. This creates the so-called identification problem, namely, the question of whether it is possible to identify a demand curve or a supply curve from observed price–quantity data. The use of simultaneous equation models to estimate economic relationships is by now perhaps the best way of distinguishing econometrics from economic statistics.

The foregoing discussion covers only nine major branches of economics. There are many other fields in economics, including economic history, comparative economic systems, business cycles, economic forecasting, national income accounting, managerial economics, business finance, marketing, the economics of natural resources, economic geography, consumer economics, and regional economics. The reader who wishes to investigate them will find references in RELATED ENTRIES under ECONOMICS in the *Ready Reference and Index*.

BIBLIOGRAPHY. The best introduction to methodological issues in economics is still L. ROBBINS, *The Nature and Significance of Economic Science*, 2nd ed. (1935). The modern position on empirical testing is well conveyed by M. FRIEDMAN in *Essays in Positive Economics* (1953); and by R.G. LIPSEY and P.O. STEINER, *Economics*, 2nd ed. (1969), an elementary textbook. The best way to learn about economics is to browse through an introductory textbook. The best known is P.A. SAMUELSON, *Economics: An Introductory Analysis*, 8th ed. (1970); another is A.A. ALCHIAN and W.R. ALLEN, *University Economics* (1964). There are several good histories of economic thought: H.L. HEILBRONER, *The Worldly Philosophers*, rev. ed. (1953), a pleasure to read; O.H. TAYLOR, *A History of Economic Thought* (1960), arranged chronologically; and E. WHITTAKER, *A History of Economic Ideas* (1940), arranged topically. J.A. SCHUMPETER, *History of Economic Analysis* (1954); and M. BLAUG, *Economic Theory in Retrospect*, 2nd ed. (1968), are advanced references. Excellent bibliographical articles on the great economists, as well as superb but sometimes quite difficult essays on all the leading branches of modern economics may be found in the *International Encyclopedia of the Social Sciences*, 16 vol. (1968). Further readings for advanced students are those published by the American Economic Association: *Survey of Contemporary Economics, I*, ed. by H.S. ELLIS (1948); *A Survey of Contemporary Economics, II*, ed. by B.F. HALEY (1952); and *Surveys of Economic Theory*, 3 vol. (1965).

Each branch of economics has its own text; the following list is highly selective. Microeconomics: C.E. FERGUSON, *Microeconomic Theory*, 2nd ed. (1969); K.J. COHEN and R.M. CYERT, *Theory of the Firm: Resource Allocation in a Market Economy* (1965). Macroeconomics: G. ACKLEY, *Macroeconomic Theory* (1961); T.F. DERNBURG and D.M. MCDOUGALL, *Macroeconomics*, 3rd ed. (1968). Development economics: C.P. KINDLEBERGER, *Economic Development*, 2nd ed. (1965); B.H. HIGGINS, *Economic Development*, 2nd ed. (1968). Public finance: R.A. MUSGRAVE, *The Theory of Public Finance* (1959). Monetary economics: A.G. HART and P.B. KENEN, *Money, Debt and Economic Activity*, 3rd ed. (1961); L.V. CHANDLER, *The Economics of Money and Banking*, 5th ed. (1969). International economics: C.P. KINDLEBERGER, *International Economics*, 4th ed. (1968). Labour economics: A.M. CARTTER and F.R. MARSHALL, *Labour Economics: Wages, Employment, and Trade Unionism* (1966); L.C. HUNTER and D.J. ROBERTSON, *Economics of Wages and Labour* (1969). Industrial organization: E.H. CHAMBERLAIN (ed.), *Monopoly and Competition and Their Regulation* (1954); J.S. BAIN, *Industrial Organization* (1959). Agricultural economics: T.W. SCHULTZ. *The Economic Organization of Agriculture* (1953), and *Transforming Traditional Agriculture* (1964). Growth economics: R.G.D. ALLEN, *Macro-Economic Theory: A Mathematical Treatment* (1967), a fairly difficult book; most texts on macroeconomics and economic development devote a chapter or two to growth theory. Mathematical economics: A.C. CHIANG, *Fundamental Methods of Mathematical Economics* (1967). Econometrics: A.A. WALTERS, *An Introduction to Econometrics* (1968); C. CHRIST, *Econometric Models and Methods* (1966).

(M.Bg.)

Economic Systems

An economic system may be defined as the sum total of institutions and patterns of behaviour that organize economic activity in society. This is the sense in which one speaks, for example, of a private-enterprise system or of a centrally planned socialist system. The production and distribution of goods in a modern industrial economy is a process in which millions of people must cooperate. For the sake of greater efficiency their functions are specialized so that none of them makes a whole product, and they are all mutually dependent. Working together, they are able to create a social product that is many times greater than the aggregate result of their isolated individual efforts would be. But the interdependence of millions of people in this gigantic process of cooperation raises a number of problems, which different economic systems solve in different ways.

The first problem is to determine what goods are to be produced, and in what quantities. Although productive

The basic economic problems

resources can be combined in an infinite variety of ways, technology and manpower set ultimate limits to the total size of the final output. Particularly important is the division of the national product into investment goods and consumption goods. A greater allotment to investment will permit increased future consumption but at the cost of a reduction of present consumption.

A second problem is that many thousands of interdependent production processes must be properly balanced if the economy is to function smoothly. The complexity of this is staggering. Let one integral part of this cooperative process of social production lag behind, and potentially all other parts will be affected.

A third problem of every economy is the way in which the common product is to be distributed. This has much to do with reward for effort. Since human beings are not simply creatures of instinct or of habit, they must be induced to perform the right functions in the right place and at the right time, whether they be farmers, industrial workers, or managers. To the extent that society relies on self-interest rather than coercion or inspiration, the promise of reward for effort becomes decisive.

Modern economic systems can be distinguished by the ways in which they solve these problems. Essentially, the differences among systems have to do with three main characteristics: (1) the degree of public or private ownership of the means of production, (2) the relative emphasis on planning, and (3) the prevalence of public or private decision making. These characteristics are interrelated. Where the ends of the economy are determined by the government rather than by the individual, central planning will tend to prevail together with public ownership of the material means of production (capital and land); where the demand decisions of consumers guide production, a market process must exist.

These distinctions must not be exaggerated, however. The centrally planned system can include a substantial amount of individual freedom and will have to make at least some use of market forces, while the private enterprise system will not only contain a large public sector but will likely seek to influence the working of the private sector through government fiscal and monetary policies. Nevertheless, it is helpful to approach comparative economic analysis through a separate study of the principles of the centrally planned and the competitive private enterprise economies. Only after the essential features of these two main types have been fully understood can the study of the existing systems of different countries become fruitful; only then can judgments be made as to whether their aims and policies are compatible and their economic structures internally consistent.

ARCHETYPAL ECONOMIC SYSTEMS

The private enterprise economy. *Its theoretical model.* The theoretical model of the private enterprise economy is built upon private ownership and control of the means of production, free choice of occupation, and consumers' sovereignty. Private enterprises, or firms, translate consumer demand for commodities and services into a demand for labour, capital, and other factors of production. The consuming units receive their income from the sale of productive services to the firms, and spend most of it on consumption. Thus money flows from the firms to the people and back to the firms in a continuous circle. This money flow is maintained even if a part of the national income is saved, that is, not spent on consumption, provided that the savings are borrowed by firms and spent on investment goods or borrowed by other consumers to be spent on additional consumption. Should the saved portion of the national income remain unspent, however, the circular flow of money would be interrupted, with serious consequences in terms of inadequate demand, reduced production, and unemployment.

Price and cost relationships decide what each firm produces in the various markets of the economy. Where demand has not been fully met, product prices will exceed production cost per unit, and the resulting profits will then call forth increased production. If supply exceeds demand, losses will result in reduced production. The market processes of the private enterprise economy involve a series of price relations that tend to lead to a state of equilibrium at full employment. In this state all resources will be employed wherever they are paid the highest price that equates supply and demand. This is so because they shift from one employment to another in response to consumer demand as expressed through the market.

Technical change and prices

The working of the price mechanism can be shown by assuming a disturbance to a given state of equilibrium resulting from a change in the techniques of production. The substitution of a newly invented machine for labour —for example—increases productivity, lowers costs, lowers the price of the product (under competitive conditions), and initially reduces the demand for labour. Lower prices permit the purchase of larger quantities of the commodity; alternatively, other commodities may be bought in increased quantities. In either case the demand for labour will increase again, though it will not always be easy to shift workers from one industry to another. All changes in the demand for products and for productive factors alter the prices on the various markets and lead to a chain reaction of adjustments until a new equilibrium is reached.

Thus self-interest, working through competitive markets, theoretically enables the private-enterprise economy to solve the problems of the social economy without the aid of a central plan. The government's sphere is limited to functions that private enterprise cannot perform, for example, the satisfaction of collective wants for which a market price cannot be charged (such as police protection or flood control). But once the government does not tell the citizen what to consume, where to work, or what to produce, the private-enterprise system offers a degree of economic freedom that cannot be matched by a centrally planned economy.

Three major criticisms. The private-enterprise system is usually subjected to three principal criticisms. The first applies both to theory and to practice; the other two address themselves only to practical applications of the system.

(1) Since private incomes are determined by prices of the factors of production and by individual ownership of these factors, the national income of the private-enterprise economy is very unequally distributed. Furthermore, individual incomes cannot be equalized to any large degree by tax and other government policies because the working of the system depends on private ownership of the material means of production and on wage differentials. The system claims the advantages of consumer's sovereignty, but since the market counts the dollars of the rich and the poor alike, the system does not maximize the satisfaction of wants for the economy as a whole.

(2) The market mechanism of the private-enterprise economy rests on competition. The growth of monopoly opens the way to exploitation, suppresses private initiative, and may damage, through price rigging, the working of a system that must rely on free price movements. Again, it is difficult for the government to rid the market of monopolistic impediments since these are, to a large extent, the outgrowth of modern mass-production techniques.

(3) Experience has shown that the private-enterprise economy does not always maintain high levels of employment. Productive resources may be wasted in depression and stagnation, and the national income may remain far below its potential maximum. Fortunately, it is generally agreed that high employment can be maintained through the application of modern fiscal and monetary policies, which have nothing in common with central planning or with the nationalization of major industries. The government merely sees to it that demand deficiencies in the private sector of the economy are counterbalanced by an expansion of public spending so that sufficient aggregate demand permits the full use of productive resources.

The centrally planned economy. The most thoroughgoing form of the centrally planned economy ("authoritarian socialism" or "command economy") features the public ownership of, or control over, all productive resources, including labour, and the direction of the economy by command from the centre according to an all-inclusive plan. But this extreme form of central planning is not likely ever to exist in real life, nor can it be assumed that any government would insist on distributing national income without allowing any choice on the part of consumers. Even the most autocratic and totalitarian governments have always placed considerable reliance on markets for consumer goods and have allowed some freedom in the choice of occupation.

The centrally planned economy should rather be seen as a system that seeks to combine public ownership of the *material* means of production (land and capital) and central planning, with as much personal freedom as this framework permits in terms of occupation and consumption. These concessions to economic freedom have important consequences. If workers are not simply commanded to perform the work that the plan requires, they must be induced to do so by wage differentials. This means that the centrally planned economy cannot achieve equal income distribution and that labour markets in the centrally planned and the private enterprise economies must be similar.

Limited consumer choice

Consumers in the centrally planned economy are permitted to choose among commodities and services; this makes it necessary to have consumer goods markets but does not require that production adjust itself automatically to changes in consumer demand. A central plan is not flexible enough to permit such automatic adjustments through price and cost signals.

The centrally planned economy meets the problems of a social economy as follows. The major goals of the economy are determined by the government and then translated into the details of the central plan. Since, at full employment, new investment competes with consumer goods production for the nation's resources, the plan must be all-inclusive. The details will depend on the ends to be achieved, the availability of productive resources, the techniques of production, and the need for proper balance. Balance might conceivably be achieved by input-output equations; in practice it is reached by approximation through trial and error.

The central plan does not rest on price and cost calculations and does not permit major independent decisions by decentralized enterprises. The plan sets output quotas that must be met on the basis of fixed allocations of productive resources. Managers who have technical know-how must obey commands issued by planners who lack knowledge. Indeed, the planners depend on information sent to them by men who may be tempted to influence the centre's decision in their favour by misrepresenting their enterprise's capacity, thus making it easier to fulfill the ensuing plan.

The central plan rests on material or physical balances, that is, on a system of relationships between, say, the total supply of steel needed for the fulfillment of production quotas in all steel-consuming industries and the total input of the various resources needed to reach the output quota for steel. Not all balances can be material balances. When dissimilar things (such as labour, fuel, and transportation services) must be added together, monetary balances are needed. Monetary balances are indispensable, furthermore, because workers are paid wages and the public makes its purchases with money.

The use of money in the centrally planned economy does not mean that the plan follows the guidance of market prices. All prices are set by the government and are but rarely changed. They serve for balancing and accounting purposes. Monetary transfers between enterprises help check plan fulfillment.

The chief criticism of the centrally planned economy is that it is not able to allocate its resources efficiently. Planners and managers cannot base their decisions on prices that reflect the real costs of goods and services. The authorities do, however, have the power to determine the rate of investment, and since the rate of investment is decisive for the rate of growth of the economy, the centrally planned system may be able to compensate for its relatively poor performance in factor allocation by its command over the rate of investment—at the cost of relatively low consumption.

Liberal socialist economies. *Consumer sovereignty.* Liberal socialists reject the authoritarian features of the centrally planned command economy. They insist on the guidance of production by the consumer; they want real consumer sovereignty, and not merely a choice among an assortment of commodities produced in accordance with the government's wishes. They also believe that consumer sovereignty makes more sense in a socialist than in a private-enterprise economy because public ownership of the material means of production will result in a more equal income distribution. Free choice of occupation, however, and the use of wage differentials admittedly exclude full equality.

Liberal socialists generally reject the command economy because of its overcentralization. They want managers of enterprises to make their own decisions on the basis of price signals. They argue that it is impossible to allocate productive resources efficiently unless this is done on the basis of a consistent and correct price structure that expresses the relative scarcities of factors and products. Liberal socialists, therefore, do not oppose all of the features of a private-enterprise economy.

Decisions on basis of price signals

Problems and difficulties. The main problem of liberal socialism stems from the fact that public ownership of the means of production eliminates genuine markets for investment goods, unless it is assumed that the managers will be permitted to compete freely for productive resources. To accomplish this it will be necessary to set million of prices correctly, so that they are consistent, correspond to the relative scarcities of products and factors, and equate prices and costs.

Another difficulty concerns the motivation of the managers. How are they to act, particularly if the decisions of the planning authorities happen to conflict with the wishes of the consumers? Some socialists have suggested that the managers should act very much as private businessmen would in a perfectly competitive market. They should produce just enough to equate prices and costs. To re-create the conditions of perfect competition in a private-enterprise economy, all accounting would have to be done *as if* the prices were independent of the manager's decisions.

The Polish socialist and economist Oskar Lange believed that "the real danger of socialism is that of the bureaucratization of economic life and not the impossibility of coping with the problem of allocation of resources." Bureaucratization will flourish, however, precisely because real markets will no longer exist for all transactions between public enterprises. As in the centrally planned economy, the problem is that the centre lacks the omniscience that would be required to set correct prices; and that the enterprise managers, who are on the spot and have the know-how, will be most reluctant to treat the prices that the planners are setting *as if* they were correct.

WESTERN-TYPE MARKET ECONOMIES

The private-enterprise system (or capitalism) is not very old. Emphasis on private enterprise developed in reaction to mercantilism, which involved, as Adam Smith said, the "attempt to direct private people in what manner they ought to employ their capitals." Adam Smith believed that every individual, intending only his own gain, is "led by an invisible hand to promote an end which was no part of his intention"—the interest of sociey. Smith was often wrongly accused of having advocated completely private enterprise with no government intervention. It is true, however, that under his influence the pendulum did swing from an emphasis on the most minute government regulation of trade to the ideal of minimum interference.

This attitude, too, proved untenable, for the market process alone is only an instrument and not an end. Left to itself, the market economy permitted the growth of monopoly, produced crises of mass unemployment, and favoured an income distribution that many did not consider socially "just."

The achievements of capitalism

Yet 19th-century capitalism could claim to be a highly productive economic system, as testified by Karl Marx who wrote that "the bourgeoisie, during its rule of scarce 100 years, has created more massive and more colossal productive forces than have all preceding generations together." On the whole, the system worked well enough to induce the countries of the West to return to it after World War I. Only prolonged stagnation and mass unemployment in the early 30s led to the complete reappraisal known as *Keynesian* economics.

John Maynard Keynes argued that while classical teaching "embodied some permanent truths of great significance . . . we cannot now accept them without much qualification." His main criticism was that classical economics taught that the market economy would always tend to a state of full-employment equilibrium and that savings would always be invested. Keynes argued that an increased desire to reduce spending and to hold liquid balances (for example, in expectation of falling prices) would raise interest rates, reduce investment spending, and create unemployment. He also considered it realistic to assume that wages and prices had lost much of their downward flexibility under monopolistic market conditions. If the market economy requires price variations as signals and if downward price movements are becoming increasingly sticky, only two possibilities remain. Either the system works on the basis of differentiated price changes upward, which means inflation, or (if the authorities try to stabilize the general price level by a tight monetary policy) there will be unemployment, since fixed commodity prices and rising costs will force business into contraction. Keynes held that vigorous government intervention was required to maintain the economy at a high level of employment.

In the leading countries of the West the governments now accepted the responsibility for maintaining high employment levels. Instead of balancing the budget each year, governments are willing to incur a (temporary) deficit either through increased spending or through lower taxes when necessary to maintain aggregate demand. These policies are complicated and the countries of the West have been using them in different "mixes" and with varying success, particularly so far as the simultaneous achievement of full employment and monetary stability is concerned. All of these fiscal and monetary policies have this in common, however, that they differ radically from Soviet-type central planning. Even socialist governments in the West appear to have lost much of their interest in the nationalization of major industries and in central planning, because the private-enterprise system seems to work reasonably well so long as aggregate demand is maintained without too much price inflation.

There are differences among Western economies in the rate of growth of the public sector, but this does not alter their basic similarity since both private and public demand are met chiefly by the output of private enterprise.

THE SOVIET-TYPE ECONOMIC SYSTEM

The main problem of the Soviet economy has lain in the complexities of the central plan. After Stalin's death the conviction grew that something was wrong with the organization of the system. Nobody suggested that the central plan itself be abolished, but it was said that the incentive system was faulty, that managers needed more independence, and that planners could not cope with the allocation problem. Reformers attempted to improve the relationship between centre and periphery through an incentive system that would make the managers support rather than obstruct the plan and spur production of the goods consumers really want.

According to Professor Yvsey Liberman of Kharkov, the system practically required the managers to misinform the planning authorities as to the true potential of their enterprises and to press for excessive supplies of materials and capital. Since the managers thus could not be trusted, they had to be kept under strict supervision; and the task of controlling tens of thousands of enterprises from the centre put too heavy a burden on the planning authorities, who necessarily lacked the technical knowledge that the managers possessed. Premier Kosygin agreed with Liberman when he admitted that "it is necessary to abolish excessive regulation of economic activity of enterprises" and asked for the establishment of "firm legislative guarantees for the expanding rights of enterprises."

The question of incentives

The solution proposed by Liberman was to revise incentives so that "what is profitable for society, as represented by the state, must be profitable for every collective of the enterprise and every member of that collective." The managers would still be required to fulfill production quotas. Furthermore, prices, investments, and "all the value, labour and major physical indices of rates and proportions in the sphere of production, distribution and consumption" would still be centrally determined. Nevertheless, "the petty tutelage of centralized planning over enterprises" would be replaced by reliance on "economic levers." The chief lever was a "profitability rate"—*i.e.*, enterprise profits expressed as a percentage of the monetary value of working capital—to be compared with a "profitability norm" that the authorities would establish as a benchmark for every branch of production. Liberman's emphasis on profits (in which the enterprise would share) was designed to prevent the managers from wasting capital by asking for larger allocations than they really needed. Their fear that good performance would only lead to stiffened norms was to be met by maintaining the norms for up to five years. The introduction of new products and techniques was to be encouraged by the setting of low profitability norms in high-priority industries.

The effort to reform the central-planning process through a decentralized use of economic levers logically required a system of prices that would mirror the relative scarcities of capital assets, productive resources, and commodities. Profit rates make no sense if they are figured on the bases of wrong prices and capital values. So far nobody has shown how the central authorities (whose limitations are now recognized where physical production planning is concerned) can correctly set and revise many millions of prices.

Another difficulty lay in the working out of the operational plans of the enterprises. It was not clear whether the managers were to compete for investment funds and would be permitted to use any input mix that promised to lower costs, nor, if they did, how their decisions would be made to conform to the quantity-assortment targets of the government.

The hybrid economy of Yugoslavia. The Yugoslav economy, as it evolved in the 1960s, seemed to demonstrate that a socialist system decidedly different from the Soviet model could be made to operate with a degree of decentralization that exceeded by far the Soviet concessions in that direction.

Under the Yugoslav system of workers' management, workers were authorized to select from their own ranks a workers' council and a managing board. The director of the enterprise, who is supposed to run it in accordance with the wishes of the workers, is not subject to them entirely; as a representative of the local community, he is entitled to set aside the workers' wishes if they are considered to be in violation of existing laws and regulations.

The impact of the workers' councils on the management of the enterprises has been considerable. Not only do the workers share in the enterprise's profits; they are able to raise wages at the expense of capital investment and, where there is monopolistic pricing, at the expense of the consumer. Furthermore, a firm may refuse to hire new workers who would be difficult to lay off and whose pay would reduce the profit shares of those already employed.

In Yugoslavia a series of reforms has reduced the national economic plan to the establishment of broad production targets. However, it is more than a mere projection of past trends; the government has the power to influence decisively the rate of capital accumulation and economic growth. The enterprises are not given production quotas or fixed allocations of materials. They formulate their own operational plans according to price signals and profit considerations.

The federal government in Belgrade does not allocate investment funds on the basis of purely market criteria but with political considerations in mind. Furthermore, the prices that govern the actions of the enterprises are in many instances either controlled or set by the government to induce the enterprises to meet the targets of the plan. Since such policies are also used in Western economies, for example in France's experiment with "soft" or "indicative" planning, the difference between Yugoslavia and some Western economies has in many respects become a matter of degree rather than of principle.

Neither Yugoslav nor French planning relies solely on the market economy. Yet, once the functioning of the economy has been largely entrusted to markets and the profit motive, misallocations and inconsistencies may develop if credit and price controls prevent the market forces from working properly. But at the same time, because they have no central plan, the market-planned economies may also suffer from inflation and unemployment.

MIXED ECONOMIES IN DEVELOPING COUNTRIES

If market forces alone were always sufficient, every country with a potential for economic growth would now have developed that potential. Obviously something more is required. The process of growth in less-developed countries must be stimulated by some form of government action that creates the conditions for growth. This does not mean, however, that a positive correlation exists between government action and the rate of growth. The reverse may be true. Results may be better, once the main barriers to development have been overcome, if the government succeeds in bringing market forces and private initiative into play.

Central planning will exceed the capacity of the bureaucracy, for such planning requires a strong, competent, and incorrupt administration, which is just what no backward country possesses. In the absence of such administration it may be better that governments should avoid planning rather than pretend to plan.

Nevertheless, some government initiative is indispensable if existing barriers to development are to be removed. The main barriers are, as a rule, a degree of poverty that does not permit sufficient private or corporate saving, the absence of monetary incentives, and inadequate government services. A clarification of the policies by which economic development can be fostered is urgent. A mixed system does not lend itself to a simple program. If the policies are not to be those of a Soviet-type economy, then whatever increased economic activity the government may undertake must have as its aim the creation and maintenance of a strong private economic sector.

Three main problems

The government will have to solve three main problems: (1) It will have to set investment priorities with respect to education, health, transportation, electric power, and many other parts of the neglected "infrastructure" of the economy. The allocation of very limited resources of capital and administrative talent among these competing demands will be difficult.

(2) The government, therefore, must try to keep its activities limited and leave as much as possible to the private sector. For this purpose it must create market conditions that permit private enterprise to work efficiently. Monetary incentives must be strengthened and investment funds directed into the most productive channels. A correct price and cost structure must be developed to serve this allocation process.

(3) The government must seek to attain the highest level of aggregate investment compatible with monetary stability and must abstain from using an excessive share of the investment funds for the public sector.

The administrations of developing countries are exposed to many temptations to do the wrong thing. For example, a government may try to use excess agricultural labour in the production of capital goods without realizing that major obstacles must be overcome before this absorption of "disguised unemployment" in rural areas can be achieved: consumption in agriculture must be kept from increasing after the labour force has been reduced; the released labour force must be supplied with capital goods; the labourers must be willing to respond to monetary incentives; and administrative or private initiative is needed to organize these changes. Furthermore, "disguised unemployment" must not be confused with the mass industrial unemployment of developed economies and therefore should not be attacked through deficit spending or by embarking on big industrial projects requiring large amounts of capital. Increased spending may do little to move workers from agriculture to industry and, in failing to raise productivity, may only create price inflation. Furthermore, an emphasis on capital-consuming investment projects violates the basic principle of factor allocation. Employment-creating projects in overpopulated poor countries should be prevailingly labour consuming; poor countries are ill-advised when they try to imitate the rich countries by show-case industrialization.

WILL ECONOMIC SYSTEMS CONVERGE?

A casual observer of more recent changes in Western and in Soviet-type economies may be tempted to conclude that they are converging. In Western economies there is a continuing growth of the public sector and increasing concern with the maintenance of a high level of economic activity at reasonably stable prices. In Soviet-type economies there is a trend toward decentralization and the use of profit incentives.

Economic systems must all deal with the same fundamental problems, and it is not surprising that they have many features in common. Modern industrial processes, which are essentially the same everywhere, also account for similarities in otherwise different economic systems. As time goes by and wealth increases there are strong pressures for more freedom for consumers in centrally planned economies and for more collective consumption in the private-enterprise economies. If these similarities constitute convergence then convergence exists. However, if convergence is taken to mean a coalescing of market forces and central planning, with a uniform system emerging that combines what was good and rejects what was bad in the original opposites, one must say that convergence is not occurring. Attempts to decentralize planning on the one hand, or to control the market on the basis of arbitrarily set prices on the other, have yet to show much success; they may in fact, lead to undesirable mixed systems in which the production processes are no longer properly guided either by central command or by private decisions.

BIBLIOGRAPHY

General texts: This article follows the arguments as presented in G.N. HALM, *Economic Systems: A Comparative Analysis*, 3rd ed. (1968). Other selected texts include: G. GROSSMAN, *Economic Systems* (1967), the most compact introduction to the subject; W.N. LOUCKS, *Comparative Economic Systems*, 8th ed. (1969), with an institutional orientation; A.R. OXENFELDT and V. HOLUBNYCHY, *Economic Systems in Action*, 3rd ed. (1965), a realistic comparison of the economies of the United States, the Soviet Union, and France; C. LANDAUER, *Contemporary Economic Systems: A Comparative Analysis* (1964), a liberal socialist approach.

Readings volumes: W.A. LEEMAN (ed.), *Capitalism, Market Socialism, and Central Planning: Readings in Comparative Economic Systems* (1963); M.I. GOLDMAN (ed.), *Comparative Economic Systems: A Reader* (1964); M. BORNSTEIN (ed.), *Comparative Economic Systems: Models and Cases*, rev. ed. (1969).

Central planning: The literature on central planning is as

enormous as the literature on the Soviet economy. For the special purpose of a comparison of different economic systems see: M.H. DOBB, *Soviet Economic Development Since 1917* (1948); P.M. SWEEZY, *Socialism* (1949); H. WELLISZ, *The Economies of the Soviet Bloc: A Study of Decision Making and Resource Allocation* (1964); F.D. HOLZMAN (ed.), *Readings on the Soviet Economy* (1962); M. BORNSTEIN and D.R. FUSFELD (eds.), *The Soviet Economy: A Book of Readings*, rev. ed. (1966).

Liberman discussion: On the Soviet side see M.E. SHARPE (ed.), *Planning, Profit, and Incentives in the USSR*, vol. 1, *The Liberman Discussion: A New Phase in Soviet Economic Thought*, vol. 2, *Reform of Soviet Economic Management* (1966). On the Western side see G.N. HALM, "Mises, Lange, Liberman: Allocation and Motivation in the Socialist Economy," in *Weltwirtschaftliches Archiv*, vol. 100, no. 1 (1968).

Liberal socialism: The most important discussion of the liberal socialist economy may be found in O.R. LANGE and F.M. TAYLOR, *On the Economic Theory of Socialism* (1938; reprinted 1969); and H.D. DICKINSON, *Economics of Socialism* (1939).

The Yugoslav economy: G. MACESICH, *Yugoslavia: The Theory and Practice of Development Planning* (1964); and S. PEJOVICH, *The Market-Planned Economy of Yugoslavia* (1966).

Economic policies in Western economies: Full employment and other modern economic policies in the developed countries of the West are discussed in all modern texts on economics.

Mixed economies in developing countries: While the number of books and articles on economic development is very large, the basic problem of planning and market has been rather consistently ignored. For an exception see H.G. JOHNSON, "Planning and the Market in Economic Development," *Pakistan Economic Journal*, 8:44–55 (1958), reprinted in M. BORNSTEIN (ed.), *Comparative Economic Systems*.

(G.N.H.)

Economic Systems, Primitive

Primitive economic systems are those that apply in so-called primitive societies. Although anthropologists have not quite succeeded in removing the word primitive from their vocabulary, they endeavour to do so because of the value connotations of the word. Some of the economies labelled as primitive represent a high degree of human adaptation, make use of extraordinary skill, and are highly complex in areas of life that in an industrialized society tend to be simplified. If anthropologists have had difficulty in getting rid of the word primitive, the general public has had even greater difficulty, and hence the phrase remains as the most common and easily recognizable in use.

One of the reasons for this is that most other labels similarly lack accuracy. Primitive economies have been called subsistence economies because in them the household produces a very high proportion of its own needs; nevertheless, there is an important component of exchange and often of trade. They have also been described as the economies of "simple societies" or economies with a "simple technological base." But it is often difficult to think of the kinship and religious aspects of Australian aboriginal society (which affect the economy) except as complex. The technology is of course simple by comparison with industrial society, yet components of it (such as highly involved methods of removing poisons from foodstuffs or the production of ornaments) can be highly complex. Perhaps the one universal characteristic (if some peasant societies are excluded) is that the peoples concerned were, at least until the advent of the West and its communications, nonliterate. Thus, "the economies of nonliterate cultures" would perhaps be the most accurate way of defining the topic.

The study of such systems is primarily the subject matter of economic anthropology, the branch of social and cultural anthropology that focusses on the nature of economizing (including the implications of such economizing for institutions and cultural values) and that endeavours to offer interpretations systematically and compare the results both temporally and cross-culturally. Some anthropologists attempt to apply general, abstract principles, mostly derived from economics, to the vast variety of human experience. Others deny the universal validity of such propositions and work through the classification of ideal types of institutions. The probability is that both approaches have their uses.

Much hinges on the application of the idea of economizing to nonliterate cultures. All peoples economize, in the sense in which the term is used in economics. That is, whenever they act they must choose to use limited resources for purposes that are chosen among competing ends. (Ends are competing in the sense that one cannot do everything at once and hence must choose what to do now and what to leave for another time.) It must be remembered that time is a resource: hence conversation and ritual are economic acts. There is nothing in this definition to imply that in some sense the economic is equated with the material: copyright, and its primitive equivalent, the right to perform an initiation song, is a resource; services are as important as material goods; the ideas of a culture are among the tools it has to use; iron in the ground can hardly be counted as a resource if people do not know it is there or have no ideas about its use.

Compared to modernized societies, nonliterate societies have a high proportion of events in which choice is severely constrained—constrained by such matters as ritual, lack of resources, restricted ideas and goals. They are not wholly and inflexibly ruled by custom, however, for myth, ritual, and custom can be manipulated to enable individuals and groups to gain their objectives; nonliterate societies can be flexible, changing, adjusting, even if the change is usually slow.

"Primitive money"

Such societies have sometimes been called nonmonetary, although there are frequently objects identified as "primitive money," such as ornaments of high value (shell-bead necklaces in New Guinea) or durable, countable objects (cowrie shells in Africa). Where money begins and nonmoney ends is difficult to determine, and from a theoretical point of view it may be that the question is meaningless. What is important to determine is what fulfills the functions of money. Ideally, money is a liquid medium of exchange; it consists of things (objects or symbols) that can be transferred freely from one party to another for value received. Its degree of liquidity depends on various restrictions, such as administrative rules (exchange regulations), physical difficulties of transfer (houses), lack of durability (some foods), lack of divisibility for small transactions (heirlooms), customary restrictions upon exchange use, lack of confidence, and so on. Nonliterate societies have many items that enter into exchange, and the current form of questioning about them is to ask: to what degree and for what purpose are they liquid? It thus turns out that in New Guinea the display ornaments once thought of as "money" are in fact limited to certain types of ceremonial exchange, whereas some relatively durable foodstuffs (yams) and tobacco are much more liquid. But no item approaches the universal liquidity represented by the abstract entries in a bank balance in modern industrialized societies.

Production. Why do men produce? First let it be said that the variety of nonliterate societies is considerable: it is much greater in many respects than the variety of Western and Oriental societies since Christ. Thus, any general, short statement concerning primitive economic systems must do violence to the facts as they apply to some particular primitive economies.

Reward or motivation. Under some conditions, such as those of desert bushmen, Australian aborigines, or Eskimos, men live so close to a perilous nature that they must adapt to it, almost become one with it, to survive. Their actions, their will to live, are dominated by their environment. But this is never the whole story and in most instances is but a tiny part of it. The goals, values, and ambitions toward which people orient their production are seldom wholly material. They include status, prestige, and the respect of one's fellow men. They include ritual and religious satisfaction, often in the form

of security against supernatural manipulation. They also have included territorial aggrandizement.

Reward for the fulfillment of production goals can thus come about in several ways. The simplest is production for direct consumption. In nonliterate societies, however, the kin connections used for production go far beyond the small nuclear family—that is, parents and their children. Since an isolated individual would have a difficult if not impossible time living on his own, everyone is part of a family; and the principle of mutual dependence, or reciprocity, becomes the element binding everyone together in production and exchange units. Its extension and reinforcement, indeed, become a major motivation in the minds of the ambitious, and a prime one for those whose very survival depends upon cooperation with others.

Thus, although production is usually much more direct and less "roundabout" than in a modern complex economy, it does not follow that incentives and rewards are simple, material, and direct. If the Western observer examines only one given transaction—for example, the growing of yams in a Melanesian garden or the construction of a canoe—he may be misled by the apparent absence of immediate material wages given to those who work. Only more complex analysis will reveal that the production is to assist kinsfolk or affinal (*i.e.*, marital) relatives in complex ceremonial transactions: the rewards are strictly long term, balanced through networks of mutual dependence, with nonmaterial considerations often predominating.

Division of labour and differentiation of social roles. In many material production tasks, the work of nonliterate societies is relatively unspecialized from family to family. All families can turn their hands to agriculture, domestic work, house building, and fishing. The division of labour tends to be based instead on sex or age categories, though within such categories, of course, skills emerge that can be called upon for special occasions. The production of nonmaterial services, on the other hand, can be crucial in the differentiation of social roles. A shaman, orator, lineage leader, or navigator may have knowledge and skills as important as wealth, resources, and power in giving him specific roles in the division of labour.

The influence of social structure

Both social structure and individual decision making and choice are important in analyzing production. The social structure helps to determine the normative goals of economic activity and embodies the organizational principles. And it influences means as well as ends: a proposed production goal, for instance, may be discarded because it would interfere with the social structure, or, alternatively, a production goal may be adopted or modified to take advantage of methods embodied in the social structure. On the other hand, although the social structure is a charter for action, it is a charter flexible enough to be manipulated in the interests of self and family. A Fijian producer, for example, has a wide choice of possible helpers: members of his household, of households belonging to neighbouring patrilineages, of lineage organizations linked to other households in his village, or lineages linked to women who by marriage have entered his family. Each one of these choices may be justified by the ethos of the social structure. The producer is faced with such choices when he is building a house or a canoe, organizing a fishing expedition, or putting together land, labour, and materials for agriculture. Whether he be African, Asian, an Indian of the Americas, or an islander of Oceania, he is in this sense a coordinating entrepreneur faced with the need to decide between alternatives.

In reference to specialization and division of labour, it should also be noted that between communities there may develop geographical specialties (such as production of ornaments or pottery) leading to trade. Although such geographical specialization may sometimes be found to be historically rooted in some comparative advantage—in the sense that one community has an advantage over another or that trade is more economical than unspecialized production—once a division of labour of this kind becomes institutionalized and customary, it may persist long after the conditions of advantage have changed or disappeared.

Exchange, distribution, and wealth. It should be evident from the discussion about that production, whether of material goods or nonmaterial services, has as its goal the satisfaction not only of subsistence needs but also of those requirements necessary to lead a "good life" as perceived in the community, including the maintenance and reinforcement of the social organization. This principle can be seen most clearly in the study of exchanges and transactions that take place in the course of relations between families, lineal groups, clans, and other units of a social structure. Such relations are, in nonliterate society, frequently, though not completely, expressed in a context of ritual, ceremonial, and conventional codes.

Important examples of such economic exchanges occur during so-called rites of passage accompanying birth, the attainment of adult status, marriage, and death, since on such occasions the familial and lineage links of the main parties require considerable emphasis. On some such occasions, exchanges are relatively simple, but on others they may involve scores or hundreds of participants and preparations lasting months. Some such occasions may have a predominantly religious or political character. It is not normally true to say that such exchanges are "wasteful," for usually the exchanges constitute a major production incentive and hence bring forth a supply of goods that would otherwise be lacking; they often have at their centre durable items; they are the occasion for capital investment; and they are responsible for redistribution of goods through a considerable social network, thus in a certain sense replacing trade.

The central place of gift and ceremonial exchange in nonliterate economies has given rise to a literature that has analyzed the socially compelling consequences of gift giving and the various obligations involved. To give a gift puts the recipient in a position of debt, a debt that may be a requirement for specific repayment or, more symbolically, an emotive and social obligation. This principle is at the heart not only of the economy but of the social system, and its consequences have implications even for modern society (as at Christmas, in marriage, and in business men's clubs).

Even in those societies in which trade and markets have developed, transactions are hardly ever impersonal. Trade partners tend to be selected and symbolized through a process of gift giving, and in market situations each successful transaction tends to reinforce subsequent bonds. Normal supply-and-demand influences on prices are often overruled by personal considerations, and regularity, security, and maintenance of the relationship are more to be valued than taking a maximum profit.

The anthropologist's notion of value, in short, tends to be ethically and philosophically oriented; whereas the economist's use of the term value is ethically neutral, denoting the degree of utility or demand associated with a given object or goal. Insofar as anthropologists do identify values in the economist's sense, their awareness of the wide range of goals that govern action in nonliterate society suggests to them that limited material indicators of standards of living are misleading. Such indicators fail to acknowledge that a society may be prepared to limit material growth in order to gain, for example, increased religious satisfaction or satisfaction derived from kinship relations. Unless such noneconomic values are taken into consideration, one might seriously underestimate or misunderstand the satisfactions, performance, and directions characterizing a particular society. The idea of a "profile of culture" is intended to take such cultural matters into account.

Such considerations lead one to examine more closely the nature of wealth. Westerners are accustomed to thinking of wealth in monetary terms as consisting of accumulated units of purchasing power or of capital. Although nonliterate societies vary in their degree of acquisitiveness and in their conceptions of which kinds of ac-

quisitiveness are ethical and which are unethical, even in those societies classed as materially acquisitive (such as in Melanesia) individuals do not appear very successful at accumulation. Variations in individual property or material holdings are not great, and differences in life style from one person to another are very small.

Where there is lack of continuous accumulation, there is often a lack of durable capital objects that can be passed on from generation to generation. Although durable capital items are never wholly lacking, their ability to call forth production may be restricted, and hence their capital function is also restricted. There are, however, other reasons for limiting accumulation; most nonliterate societies somehow recognize that the accumulation of material wealth carries with it power and political weight, which can upset the traditional or conventional balance of forces. There thus exist conventions and mechanisms to control and check such accumulation. These include symbolic redistribution or even destruction of accumulated material wealth (potlatch being the most famous example) and the use of sorcery and witchcraft to bring unduly ambitious men to heel.

Sometimes, although material accumulations may be frowned upon, there may be respected means of accumulating control over other resources, such as personal services and obligations. If an ambitious man obtains material objects, such as foods, ornaments, pottery, and mats, beyond his immediate requirements, he may escape the threat of sanctions by redistributing the goods as ceremonial or personal gifts. Each time he does this, however, he creates counterobligations. He can thus accumulate wealth, not in the form of a material stock but in the form of obligations to him. The ambitious entrepreneur can call on such obligations as he needs them, gaining prestige, the reputation for wealth, social influence, and economic and political control as he does so. The wealthy man under such conditions is a man without a bank or a warehouse. Wealth consists in the accumulation of obligations that imply powers of control.

Property. No human society exists without property and ownership in some form, and in no case can property rights be conceived of if one considers only a single individual. For property rights are an expression of recognized social relations. Theoretically, property emerged with the first objects of material culture. If a man using a tool prevents others from using it, he exercises or creates a property condition. If he shares it, he defines the property conditions in a different manner. In sharing it, he and others define the terms of sharing and the class of persons included in or excluded from sharing (such as those within or outside the family or kin group).

Such considerations affect all material objects. In traditional nonliterate societies, the relation of a person or group to the land is of course paramount; and whether the society is nomadic, pastoral, or agricultural, a prime element in social relations consists in the norms of interaction, with land as the physical referent. The patterns of land use, and the rights and obligations defining who has a say in the use and product of the land are a result of complex historical and ecological adjustments.

Mythological sanctions for property

Existing rights and practices frequently are justified by reference to history as expressed in myths, tales of origin, accounts of warfare, and the like. Such charters tend to be highly manipulated oral documents, with a strong component of religious and ritual content. Typically, they infer or explain the position of the defined social groups, such as lineages. The family or individual relates his property position to the general social charter and to the more recent specific history of interactions that can be related to it. The specificity of the family or individual position varies according to whether the need to define landed property is specific or general: thus hunting bands may not need to define individually exclusive rights within the total group rights, whereas agriculturalists and shifting cultivators must.

Observation of land use, including the way in which it is justified by reference to right and to historical transactions, usually reveals a pattern much more complex than that suggested solely by the formal social structure or mythology. The interplay of lineal and affinal relatives, the ability to gain access to land through a number of kinship channels, the holding of such land over long periods and the gradual acquisition of rights of disposition, the use or relinquishment of land as an aspect of wider reciprocal exchanges, and the acquisition of political power-carrying rights of general land administration are all factors that can lead to a much more complex crisscrossing of interests in land than might otherwise be deduced.

In addition to rights and obligations with respect to material objects and land, there are others that are rather like copyright, or property in ideas or skills. Knowledge is a significantly scarce resource in any society, and in nonliterate societies it has special manifestations. The right to use symbols (as with totem poles), to make use of a ritual song, to use a tale or a dream as a charter, to exercise a skilled craft such as that of manufacturer of shell ornaments are each examples. In very many instances, such rights (and the obligations that go with them) can be inherited and can even be traded or exchanged, though usually through ritualized forms.

Transformation. There is an extensive literature on the transformation and adaptation of economics under conditions both of traditional patterns of change and of the massive pressures of modern capital, trade, and political and religious action. From all the studies, however, there has as yet been relatively little successful theoretical synthesis or practical application. The best laid theories, when translated into applied projects of change or development, have more often than not been failures. In large measure this can be traced to the fact that change is complex, whereas the tools available to induce change tend to be simplistic, partially applied, out of tune with the values and ambitions of the people, and open to political or moral suspicion.

The difficulties of deliberately inducing change can be so great that many frustrated reformers have tended to find the explanation in the irrationality of traditional society and the dead hand of "custom." This is now normally recognized to be a misunderstanding. Traditional society is rational within its goals, and many reformers do not understand the basis of those goals and the rationality that flows from them and are hence misled. Also, within its own terms, traditional society can be highly adaptable. But if one does not understand this, it is easy to agree with the early analysts, who up to the late 1950s (with echoes still today) tended to feel that progress was only possible through the destruction of traditional values and social organization, opening up the way for individualism and materialism to take charge of events.

Today there are enough examples of successful change, expansion, and modernization based upon a non-Western core of values and social organization to indicate that a policy of destruction is more likely to hinder change (by creating oppositions) than to speed it, and that modernization has many forms. It is true that to date the most successful examples can be drawn from such countries as Japan, Taiwan, and India, which can hardly be described as traditionally nonliterate. Nevertheless, the growing cities, agricultural complexes, and industries of Nigeria, Zaire, Fiji, and other countries working from a traditionally nonliterate base, are well on the way to success as measured in terms of industrialized societies. And it is now clear that their modern form will be far from a mirror image of Detroit, the Midlands, Moscow, or the Ruhr. Adaptation and variety are likely to survive the creation of the computerized world village.

BIBLIOGRAPHY. Two recent reviews of the general issues are M. NASH, *Primitive and Peasant Economic Systems* (1966); and C.S. BELSHAW, *Traditional Exchange and Modern Markets* (1965). Some of the more important specialist articles on this subject have been collected in G. DALTON (ed.), *Tribal and Peasant Economies* (1967); and E.E. LeCLAIR and H.K. SCHNEIDER (eds.), *Economic Anthropology*, including a good bibliography (1968). On gift exchange and money, classic references are M. MAUSS, *The Gift: Forms and Func-*

tions of Exchange in Archaic Societies (1954; orig. pub. in French, 1925); P.M. BLAU, *Exchange and Power in Social Life* (1964); and P. EINZIG, *Primitive Money* (1949). General statements on change and development are contained in C.S. BELSHAW, *The Conditions of Social Performance* (1970); E.E. HAGEN, *On the Theory of Social Change* (1962); B.F. HOSELITZ, *Sociological Aspects of Economic Growth* (1960); and J.H. KUNKEL, *Society and Economic Growth* (1970). Monographs dealing intensively with particular economic systems, or their change, include the following: F.G. BAILEY, *Caste and the Economic Frontier: A Village in Highland Orissa* (1957); C.S. BELSHAW, *Under the Ivi Tree: Society and Economic Growth in Rural Fiji* (1964); P. BOHANNON and G. DALTON (eds.), *Markets in Africa* (1962); H.F. CODERE, *Fighting with Property: A Study of Kwakiutl Potlatching and Warfare, 1792–1930* (1950); R.W. FIRTH, *Primitive Polynesian Economy*, 2nd ed. (1965); G.M. FOSTER, *Primitive Mexican Economy* (1942); BRONISLAW MALINOWSKI, *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea* (1922, issued in pa. 1961); M. NASH, *The Golden Road to Modernity: Village Life in Contemporary Burma* (1965); K. POLANYI and A. ROTSTEIN, *Dahomey and the Slave Trade* (1966); L.J. POSPISIL, *Kapauku Papuan Economy* (1963); and R.F. SALISBURY, *From Stone to Steel: Economic Consequences of a Technological Change in New Guinea* (1962).

(C.S.B.)

Ecosystem

To understand the meaning of the word ecosystem and to appreciate the role this word plays in communicating man's increasing concern about his environment, it is well to consider the holistic concept that underlies the term. Holism, in this connotation, is based on the theory that living components (organisms, including man) and nonliving components (the physical environment) function together as a whole according to well-defined physical and biological laws. Also fundamental to such a holistic concept is an important generalization variously known as the theory of integrative levels or the theory of hierarchical controls. By this is meant that as components are added together to create larger functional units, additional attributes come into focus, attributes not present or not evident from the behaviour of the components functioning separately. When hydrogen is combined with oxygen in a certain manner, for example, water, which is distinct from both of its components, is formed. Similarly, when trees evolve together, they form forests, with an entirely new set of attributes. To understand fully water or forests, therefore, knowledge about the functional wholes as well as the parts is required. The importance of the ecosystem concept is the emphasis it places on the overriding necessity for understanding the large units of nature in which man, his domestic organisms, and his energy-demanding machines are interdependent components, along with other essential components that maintain the total life-support system.

The underlying holistic concept

The word ecosystem is a contraction of ecological system. Ecology, in turn, is derived from the Greek word meaning "house." Persons all over the world, seizing on the root meaning of ecology, have broadened its scope from former rather narrow academic confines and use it to refer to the study of the totality of man and environment, or to the whole "environmental house," as it were. The word system has been defined as regularly interacting and interdependent components forming a unified whole. An anthropocentric, or human centred, definition of ecosystem, therefore, could be: man as a part of, not apart from, a life-support system composed of the air, water, minerals, soil, plants, animals, and micro-organisms, all of which function together and maintain the whole (Figure 1). A more formal definition is: any unit —including all the organisms (biological factors)—interacting with the environment (physical factors) so that a flow of energy within a system leads to a clearly defined trophic (nutrient-requiring) structure, to biotic diversity, and to an exchange of materials between living and nonliving sectors.

The size of an ecosystem

There is no size limit implied in the definition of an ecosystem. It may be a square kilometre of jungle, a square metre of desert, a pond, a woodlot, a city, a farm,

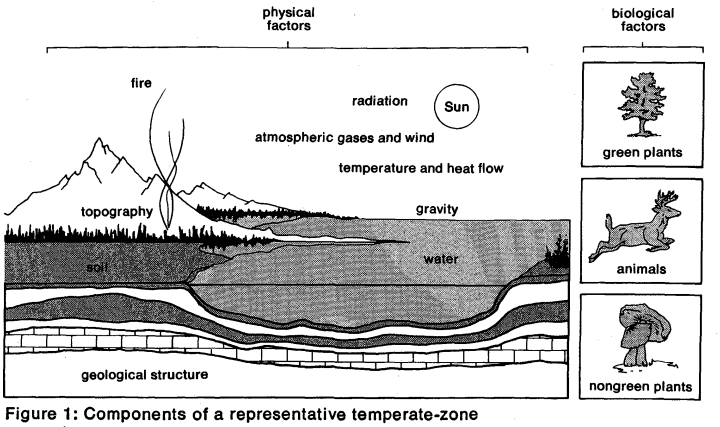


Figure 1: Components of a representative temperate-zone ecosystem.
From *Biological Science: An Inquiry into Life*, 2nd ed. (1968); Harcourt Brace Jovanovich, Inc., New York; by permission of the Biological Sciences Curriculum Study

or a closed container of small organisms (*e.g.*, an aquarium or a vivarium). The largest ecosystem is the biosphere, the entire world of life, and its associated geosphere, the inanimate earth (see BIOSPHERE); because of its magnitude, this ecosystem is often referred to as the ecosphere. Only in man-made ecosystems are the boundaries clear; natural ecosystems blend together at overlap areas called ecotones, as at the edge of a forest or at the seashore.

Historical review of the ecosystem concept. Although the term ecosystem was first proposed by a British ecologist, Arthur George Tansley, in 1935, the concept is much older. Allusions to the idea of the unity of organisms and environment—as well as the oneness of man and nature—can be found far back in written history and often as a basic part of many religions. Anthropologists and geographers, long concerned with the impact of man on his environment, early debated the extent to which man's continuing trouble with deteriorated environments stemmed from the fact that human culture tends to develop independently of the natural environment. A Vermont writer, George Perkins Marsh, produced a classical treatise on this theme in 1864, in which he analyzed the causes of the decline of ancient civilizations and forecast a similar doom for modern ones unless man takes what would now be called an ecosystematic view of man and nature. In the late 1800s essays began to appear on the unity of nature, interestingly enough in a roughly parallel manner in German, English, and Russian. Thus a German, Karl Mobius, in 1877, wrote about the community of organisms in an oyster reef as a biocoenosis. In 1887 an American biologist, Stephen Alfred Forbes, wrote a significant essay on "The Lake as a Microcosm." A pioneer Russian ecologist, V.V. Dokuchaev, and his disciple, G.F. Morozov, placed great emphasis on the concept of biocoenosis, a term later expanded to geobiocoenosis, or biogeocoenosis, which can be considered a synonym of the word ecosystem.

No one has expressed the relevance of the ecosystem concept to man better than did an American ecologist, Aldo Leopold, in his essays on the land ethic. In 1933 he wrote: "The Golden Rule tries to integrate the individual to society; democracy to integrate social organization to the individual. There is as yet no ethic dealing with man's relation to the land," which is "still strictly economic, entailing privileges but not obligations." Leopold argued that without the last, the little progress that has been made with the other two ethics will surely be lost.

Although the idea of the ecosystem was accepted for a long time, before the 1960s it was not in any way applied to man's affairs. Up to this time, applied science consisted largely of managing components as more or less independent units. There was forest management, wildlife management, water management, agriculture, soil conservation, and pest control, but no real planning and management of the whole environmental complex; that is to say, mankind had not yet practiced true ecosystem management. Controlled management of the human

population, together with the resources and the life-support system on which it depends as a single, integrated unit, now has become the greatest and certainly the most difficult challenge ever faced by human society. During the closing decades of the 20th century, mankind shall witness a historic revolution in the way people look at their environment for the very simple reason that for the first time in his short history man is faced with ultimate rather than merely local limitations.

The components of the ecosystem. An ecosystem has two components that are usually partially separated in space and time: (1) an autotrophic, or self-nourishing, component—largely green plants—in which light energy is used to build up complex organic substances from simpler inorganic ones; and (2) a heterotrophic, or "other nourishing," component, in which complex organic substances are utilized, rearranged, and ultimately decomposed. Natural ecosystems, therefore, consist of an upper green belt, which receives incoming solar energy and which overlaps with a lower brown belt, in which organic matter accumulates and decomposes in soils and sediments. The vegetation canopy comprises the autotrophic layer in terrestrial and shallow-water ecosystems, and the soil represents the heterotrophic layer, in which the necessities of life are regenerated in forms usable by green plants. Similarly, in deepwater ecosystems, such as the ocean, the layer of floating algae in the upper, lighted zone is the green belt; and the lower dark zone, with its sediments, is the brown belt.

The green and brown belts of natural ecosystems

It is convenient for purposes of analysis to recognize two layers, six structural components, and six processes as comprising the ecosystem according to the following outline:

A. Layers
  1. autotrophic layer (green belt): the buildup of complex substances predominates
  2. heterotrophic layer (brown belt): the decomposition of complex materials predominates

B. Components
  1. inorganic substances involved in material cycles
  2. organic compounds that link biotic and abiotic sectors
  3. climatic regime (*e.g.*, temperature, rainfall)
  4. producers: chiefly green plants that manufacture food from simple substances and light energy
  5. consumers: animals that ingest other organisms or particulate organic matter
     a. primary consumers: animals that eat plants
     b. secondary consumers: animals that eat other animals
  6. decomposers: chiefly micro-organisms that break down complex compounds with release of products that are recycled by plants or that affect other biotic components

C. Processes
  1. energy flow
  2. food chains (trophic relationships)
  3. diversity patterns in time and space
  4. water and mineral (nutrient) cycles
  5. development and evolution
  6. control (cybernetics)

As with any classification, such subdivisions are arbitrary, but they provide a useful beginning for analysis. Often the components are grouped under abiotic, or physical factors (1–3) and biotic, or biological, factors (4–6). In the holistic viewpoint all components are operationally inseparable; in practice different methods are required to delimit structure on the one hand and to measure rates of function on the other. The ultimate goal of study at any level of organization, however, is to understand the relationships between structure and function. In this article the emphasis is on major functions; the reader is referred to textbooks and review articles (see below *Bibliography*) for details on ecosystem components.

Energy flow and nutrient cycling. The two overriding processes that determine the nature and productivity of an ecosystem relate to energy flow and nutrient cycling. Figure 2 shows in a diagrammatic manner how the one-way flow of energy (clear channels) drives the movement of vital materials (shaded channels) in circular pathways

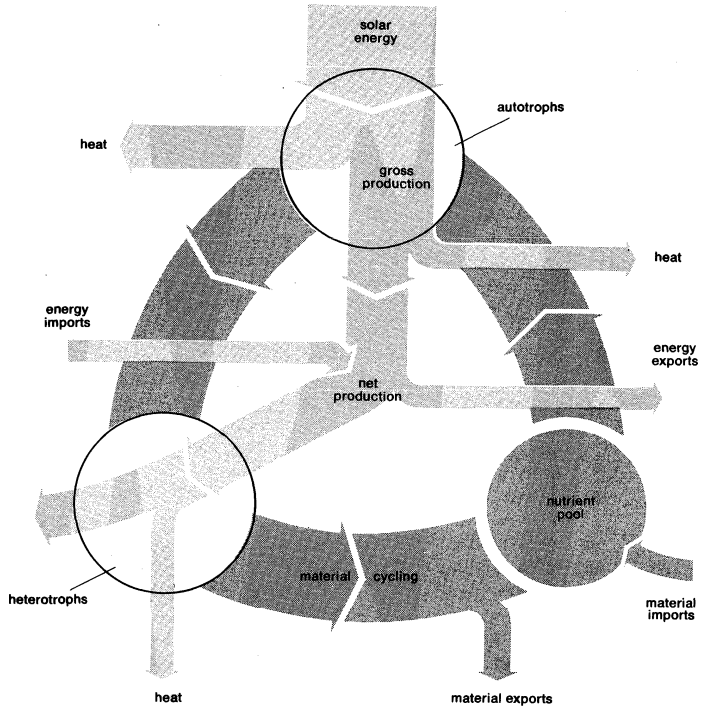


Figure 2: Relationship of energy flow to ecosystem productivity (see text).

(except for importation and exportation of certain matter). The ultimate source of energy that runs the ecosystem is solar radiation. Some of this radiation is converted into organic matter by the autotrophs; it is this primary production that provides the entire food supply of the entire earth community, including man. The sequence of transfer of food energy from one group of organisms to another is known as the food chain, or, if complex, as a food web (Figure 3). At each transfer a large part of the energy (as kilocalories) is lost to the system as heat (the respiration, or R-flows), in agreement with the second law of thermodynamics. In general, 1 to 5 percent of solar energy is fixed by the autotrophic layer (organisms requiring only carbon and nitrogen to synthesize all food requirements); 10–20 percent is transferred from one heterotrophic layer (requiring complex organic compounds) to another. These efficiencies can be increased somewhat, at least for short periods of time, by introducing into the system additional energy in the form of organic matter, cultivation, irrigation, and fertilization, as is done in agriculture; but experience shows that any increase in production or production-transfer efficiency above that which has evolved in nature results in additional thermodynamic disorder (*e.g.*, pollution, disruption of nutrient cycles), which must be taken into account before it can be concluded that a given strategy really has a net benefit in production.

Food chains and food webs

A forking pattern of food chains is shown in Figure 3, a model that relates more directly to the two-layered structure of ecosystems as described earlier. When living plants are consumed by herbivorous animals (grazers), which in turn are eaten by predators, the sequence can be designated conveniently as the grazing food chain. If, on the other hand, plants are not consumed until they die and fall into the heterotrophic stratum, in which decomposing micro-organisms play a major role, the situation is described as the organic detritus food chain. In mature ecosystems, such as forests, less than 10 percent of net primary production is consumed by grazers; the rest goes into the detritus food chain. In what can be termed youthful systems (see below *Ecosystem development*), such as most agricultural systems or fertilized fishponds, 50 percent or more of net primary production flows along the grazing pathway. Experience with pastures (grass–cow–man food chains) has shown that if more than about 40 percent of the annual net primary

production is grazed future productivity may be jeopardized; apparently, a large part of each year's production must be left in the system if moisture retention and nutrient-cycling capacity are to be maintained. In most natural ecosystems there exist feedback mechanisms that prevent overgrazing; unfortunately, man's economic systems, which are based on the short-term profit motive, tend to encourage overgrazing (and overuse of resources generally), since it is often possible to obtain increased yields for several years by overuse. The Table summarizes a nine-year study in which pastures were grazed at three different levels. It can be seen that it paid to overgraze (*i.e.*, to use more than 50 percent of yearly net production) for about five years in the sense that a larger amount of beef was produced. After nine years, however, the overgrazed pasture was so damaged that its yield was but a fraction of the pastures grazed in a moderate manner.

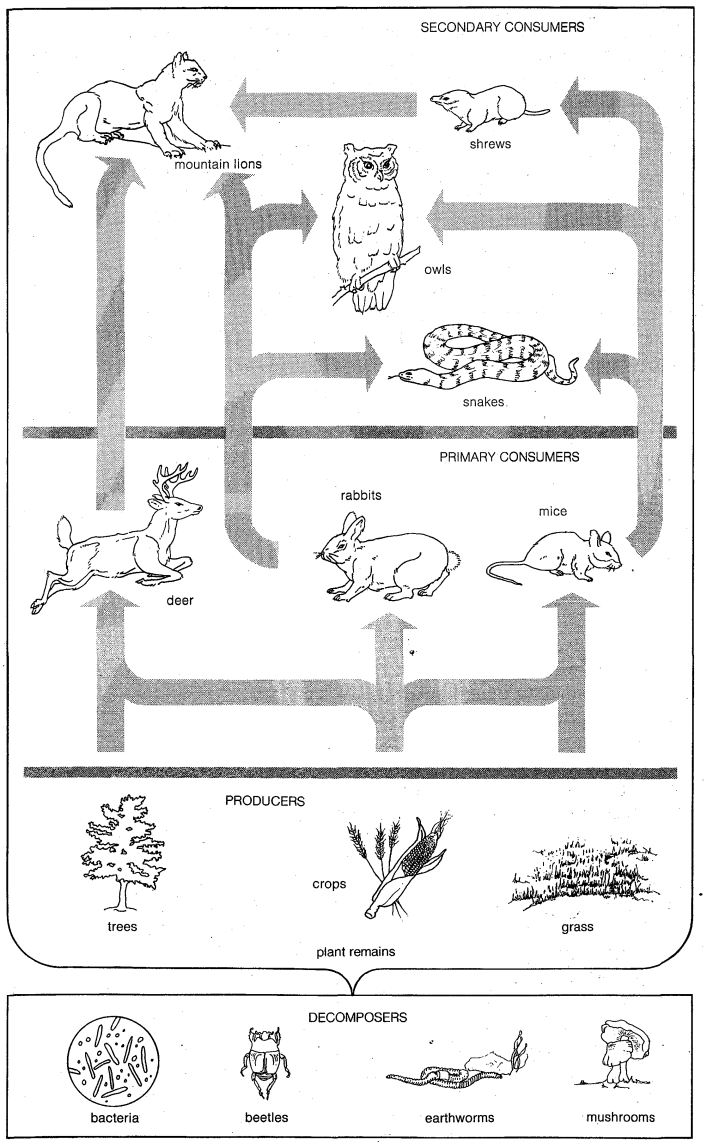


Figure 3: A representative food web comprised of several interacting food chains.
From *Biological Science: An Inquiry into Life*, 2nd ed. (1968); Harcourt Brace Jovanovich, Inc., New York; by permission of the Biological Sciences Curriculum Study

The value of analytical models

The bulk of incoming solar energy is dissipated in the work of driving the water and other physical cycles, in powering weather systems, and in maintaining a gradient of temperature conditions to which various kinds of organisms are adapted. From the global standpoint, one of the most important cycles is the water, or hydrologic, cycle shown in Figure 4, in terms of a picture model and a systems model derived from it. The advantage of the systems model is that it shows the steps that require solar energy to drive the cycle and the steps that release energy for useful work in operating the biological components of the ecosystem. Such a diagram also is the first step in translating energy flows and amounts into computer language for simulating and testing total ecosystem behaviour.

Grazing Intensity Model*

production efficiency	intensity of grazing		
	light (1 cow per 21 acres)	moderate (1 cow per 15 acres)	heavy (1 cow per 9 acres)
Forage utilization (percent of vegetation grazed)	37	46	63
Condition of range (end of 9-year period)	improved	unchanged	deteriorated
Annual beef production (lbs per acre)			
First 5 years	15	18	32
Next 3 years	15	15	18
Ninth year	20	17	9

*A nine-year study of a South Dakota range.

Until recently, man tended to take for granted this immense amount of "free" work of nature and scarcely bothered to measure the energy drive or the flux rate of water movement between different compartments of the ecosystem. As industrial power flows are greatly increased by releasing energy from fossil fuels and atomic sources, it becomes clear that stability of vital natural cycles and energy flows can be adversely affected; for example, the thermodynamic-disorder consequences of rapid burning of fuels, namely the release of CO_2 and dust into the atmosphere, may now already be affecting the heat balance of the Earth. Carbon dioxide tends to act as a heat trap, warming the Earth, whereas dust tends to reflect incoming energy, thus cooling the Earth. Even small changes either way can influence the amount of water available as liquid, which in turn could greatly alter sea levels and climates.

In contrast to the water cycle, the cycling of nitrogen is largely controlled by the biological community rather than by strictly physical forces. A key role is played by specialized micro-organisms, many of which are sensitive to poisoning by man-made chemicals and other wastes. The study of biological nitrogen fixation and recycling is of such vital importance to man that intensive and worldwide interdisciplinary ecological studies have been launched, as, for example, those supported by the International Biological Program.

Systems ecology. If man is to cope adequately with the ecosystem level of organization in the sense of developing a new applied science of ecosystem management, the models, as shown in Figures 1–4, must be formalized in terms of mathematics. The formalized, or mathematical model, approach to populations, communities, and ecosystems is known as systems ecology; it has become a major science in its own right for two reasons: (1) extremely powerful tools are now available in terms of mathematical theory, cybernetics, and electronic data processing; (2) formal simplification of complex ecosystems provides the best hope for solutions to man's environmental problems, which can no longer be trusted to trial and error or to one problem-one solution procedures that have been employed in the past. Although the formal models thus far are rather crude and as yet have limited predictive value, the following eight points have emerged as a result of modelling the whole of nature:

1. Since available energy declines with each step in a food chain, a system can support more herbivores than carnivores. For this reason, fewer meat eaters than vegetarians, including man, are supported by a given food base. Materials, however, often become concentrated

with each step in a food chain; failure to anticipate the possible biological concentration of pollutants, such as DDT or long-lived radioisotopes, has caused serious problems in man's environment.

2. It has become increasingly evident that high biological productivity (in terms of energy yield per unit area) in both natural and agricultural ecosystems is usually achieved with the aid of energy subsidies from outside the producing system. Energy subsidies take the form of wind and rain in a rain forest; tidal energy in an estuary; or the fuel, animal, or human work energy used in the cultivation of a crop. Comparisons of the productivity of different systems should take into consideration the complete budget—not just solar input.

Energy subsidies

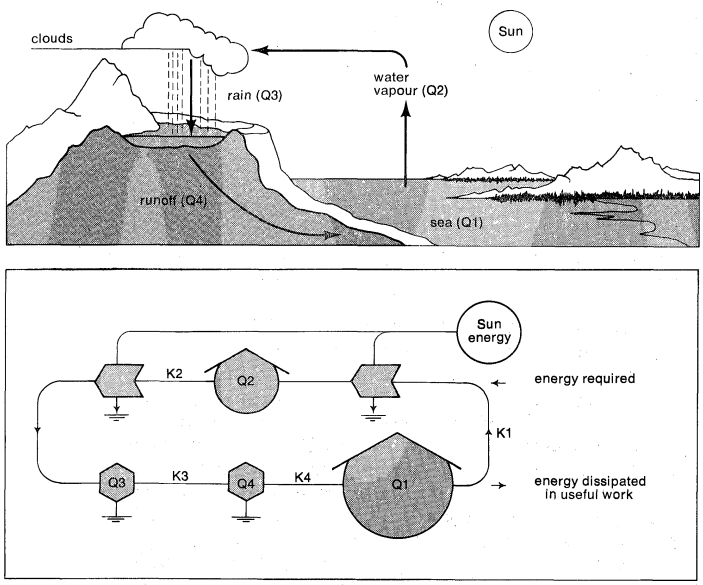


Figure 4: Conceptual models of the water cycle. The symbols are mathematical shorthand for energy movement.

3. It has become evident that both harvest and pollution are stresses that reduce the energy available for self-maintenance of an ecosystem. Man should be aware that he will have to pay the costs of added maintenance, or "disorder pumpout." Attempts to force too much productivity, or yield, from the landscape—the aim of the so-called green revolution—are dangerous because very serious ecological backlashes can result from the following: pollution caused by heavy use of fertilizers and insecticides and the high consumption of fossil fuels; unstable or oscillating conditions created by one-crop systems; genetic vulnerability to disease of highly bred plants because their self-protection mechanisms have been "selected out" in favour of high yield; and social disorder created by rapid shift of rural people to cities that are not prepared to house or employ them. The tragedy of the last point is that industrialized agriculture can result in increased food per acre, but it can also widen the gap between rich and poor, so that increasing numbers of people are increasingly unable to buy the food produced.

4. Although production and decomposition are generally thought of as being balanced in the biosphere as a whole, this balance has never been exact; rather, it has fluctuated from time to time in geological history. Throughout evolution, production has slightly exceeded decomposition, so that an atmosphere with a high oxygen content has replaced the original (reducing) atmosphere of the Earth. Man, of course, now tends to reverse this trend by increasing decomposition (burning of fuels) at the expense of production.

5. The diversity of living things is apparently directly correlated with stability, and perhaps inversely correlated with productivity, at least in many situations. Most ecosystems contain a larger number of species than would seem necessary to maintain basic energy flows and cycling; models indicate that variety may be a necessity in the evolution of natural systems. Thus, it is possible that the preservation of diversity in the ecosystem is important.

6. At the population level, it is now clear that the growth form of the human population does not conform to the simple model, in which density levels off below carrying capacity, or the peak ability to support persons. Since there will always be a time lag in the effects of crowding, pollution, and overexploitation of resources, growth will not automatically level off, as do populations of yeasts in a confined vessel. Instead, it is probable that the human population will exhaust some vital resource unless man can anticipate the effects of overpopulation and reduce growth rates before the deleterious effects of crowding are actually felt. Intelligent behaviour would seem to be the most effective control mechanism in view of the fact that the extent of such restraints as disease and famine have been greatly reduced—at least for the time being.

7. Some of the most important developments in ecology have been in the area of biogeochemical cycling and recycling. Since recycling of water and minerals will eventually become a major goal of human society, such pathways in nature are of great interest. At least four major ones occur, which vary in importance in different kinds of ecosystems: (1) recycling via microbial decomposition of detritus; (2) recycling via animal excretion; (3) direct recycling from plant back to plant via micro-organisms, such as the fungal mycorrhizae associated with roots; and (4) autolysis, or chemical recycling, in which no organisms are involved. The third pathway seems to be especially important in the humid tropics, which suggests that tropical agriculture might be redesigned to include food plants with mycorrhizae.

Recycling

8. The principles inherent in limiting factor analysis and in human ecology can be combined to formulate the following tentative overview: In an industrialized society, energy (*e.g.*, power, food) is not likely to be limiting, but the pollution consequences of the use of energy and of the exploitation of resources are limiting. Thus, pollution can be considered the limiting factor for industrialized man—which may be fortunate since pollution is a visible signal to man.

Ecosystem development. Principles concerned with the development of ecosystems are most relevant, in view of man's present situation, and for that reason are summarized here. In older textbooks this general subject is usually called "ecological succession."

In broad view, ecosystems develop through a rapid growth stage that leads to some kind of maturity or steady state (climax), usually a fluctuating steady state. The early successional growth stage is characterized by a high production to respiration (P/R) ratio, high yields (net production), short food chains, low diversity, small size of organism, open nutrient cycles, and a lack of stability. In contrast, mature stages have a high biomass (total mass of living matter) to respiration (B/R) ratio, complex food webs, low yields, high diversity, and stability. In other words, major energy flow shifts from production to maintenance (respiration) as an ecosystem reaches maturity.

The general relevance of the development sequence which has been described above to land-use planning can be emphasized by general contrast of young and mature ecosystems: young ecosystems display production, growth, and quantity, while mature ecosystems display protection (of production), stability, and quality.

Young versus old ecosystems

Since all of these characteristics are desirable in the aggregate, two possible solutions to the dilemma can be suggested. A compromise can be obtained so as to provide moderate quality and moderate yield on all the landscape, or the landscape can be compartmentalized so as to maintain simultaneously highly productive and predominantly protective types as separate units subjected to different management strategies. If ecosystem development theory is valid and applicable to land-use planning (total zoning), then the so-called multiple-use strategy, about which much has been written and said, will work only through one or both of these cases be-

cause in most cases projected multiple uses conflict with one another.

The general failure of land-use planning and zoning, as presently carried out, appears to result from, first, a failure to recognize conflicts between different uses that can not be resolved by continual compromise and, second, a failure of the economic system to place a monetary value on that part of the environment that functions as the life-support apparatus.

BIBLIOGRAPHY. EUGENE P. ODUM, *Fundamentals of Ecology*, 3rd ed. (1971), is a comprehensive college textbook and reference, designed also for the citizen, educator, and political leader; see also the same author's "The Strategy of Ecosystem Development," *Science*, 164:262–270 (1969). Three paperbacks that emphasize the ecosystem approach to ecology are: EUGENE P. ODUM, *Ecology* (1963); EDWARD J. KORMONDY, *Concepts of Ecology* (1969); and ROBERT H. WHITTAKER, *Communities and Ecosystems* (1970). HOWARD T. ODUM, *Environment, Power, and Society* (1971), is a semipopular introduction to systems ecology, with emphasis on technological, political, and economic solutions to man's environmental problems. *Ecosystem Structure and Function* (1971), contains the proceedings of the 31st Biology Colloquium, in which five authors deal with ecosystem concepts. ALDO LEOPOLD, *A Sand County Almanac, and Sketches Here and There* (1949), is an environmental classic, with near poetic essays on "The Land Ethic," "Wilderness," and "Conservation Esthetic," all imbued with the place of man in ecosystems.

(E.P.O.)

Ecuador

Ecuador, a republic situated in northwestern South America, is, after Uruguay and Guyana, the smallest state in South America. It is crossed by the Equator, from which it derives its name. Its possession of the Archipiélago de Colón, or the Galápagos Islands, which have an area of 3,075 square miles (7,964 square kilometres), brings its total land area to 109,483 square miles (283,561 square kilometres), of which continental Ecuador amounts to 101,429 square miles (275,597 square kilometres). The population, early in the 1970s, was about 6,000,000.

Ecuador is bounded by Colombia to the north, Peru to the south and east, and the Pacific Ocean to the west. The border with Peru, established by the Rio de Janeiro Protocol of 1942, has never been accepted by Ecuador, which claims a larger portion of the Amazon Basin. The country is traversed from north to south by the Andes mountains. The capital city is Quito.

Although Ecuador is rich in natural resources, rapid and sustained economic growth has not yet begun. The subsoil, especially in the east, is virtually unexplored. The economy is basically agricultural, with bananas and coffee constituting the two most important products. The exploitation of petroleum deposits has begun in the northeast, and copper is produced from mines in Azuay Province. It is hoped that these mineral resources will help to stimulate the economic development of the country.

(For associated physical features, see ANDES MOUNTAIN RANGES; GALAPAGOS ISLANDS; for historical aspects, see ECUADOR, HISTORY OF.)

THE LANDSCAPE

Physiography. *Relief.* Three parallel chains, or cordilleras, of the Andes Mountains, the eastern, central, and western—all running approximately north to south—divide the country into three main regions: the coastal region; the sierra, or mountain, region; and the east.

The three regions

The coastal region is composed of lowlands extending from the Pacific Ocean to the western edge of the mountains and rising from sea level to an altitude of 1,650 feet. In the centre of these lowlands—the location of the Guayas and Manabí provinces—a small mountain range of volcanic origin rises to 2,600 feet. It is variously known as the Colonche, Chongón, Paján, or Puca Mountains. The range is surrounded by a slightly undulating plain that stretches northward to the Colombian border. In the south, the lowlands were primarily formed by the silt deposits of the rivers that empty into the Gulf of Guayaquil. Puná, in the Gulf of Guayaquil, is the major coastal island.

The western and central ranges of the Andes bordering the sierra region constitute the country's highest and most continuous mountain chains. Many peaks are volcanic or snow covered; these include Cayambe, 18,996 feet (5,790 metres); Antisana, 18,717 feet (5,705 metres); Cotopaxi—the world's highest active volcano—19,360 feet (5,897 metres); Carihuairazo, 16,752 feet (5,106 metres); Chimborazo, 20,556 feet (6,267 metres); Tungurahua, 16,420 feet (5,005 metres); Altar, 17,277 feet (5,266 metres); and Sangay, 17,158 feet (5,230 metres). The two ranges are connected at intervals by transversal mountain chains of volcanic origin, much as individual stairs connect a staircase. Between the transverse mountains are large, isolated valleys or basins, called *hoyas*, that are named for the main river running through them.

The eastern region begins with the eastern spur of the central range, which extends to the border with Peru. This region is crossed by the eastern, and least important, cordillera of the Andes, which is also comprised of three sections—the Cordillera de Galeras, which includes the northern mountains; the Sierra de Cutucú, which borders the Upano Valley and includes the central peaks; and the Cordillera del Cóndor to the south, which borders the Zamora Valley. Beyond this eastern cordillera, again to the east, is the Amazon River Basin.

Drainage. As the annual thawing of snow occurs in the mountains, numerous rivers arise, pass through the *hoyas* of the Sierra, and flow either west to the Pacific coast or east to the Amazon River.

The river system

The coastal region is primarily drained by the Río Guayas Basin. Formed by the juncture of the Daule and Babahoyo rivers and their affluents, the Río Guayas is navigable for the greater part of its course. The rivers that also flow to the ocean include the Santiago and Cayapas and their tributaries; the Esmeraldas; the Jubones; and the Santa Rosa.

In the sierra region the rivers, torrential in their upper courses, become calmer in the plains areas but remain, nonetheless, unnavigable.

The majority of Ecuador's rivers, found in the eastern region, carry the greatest volume of water and are the most navigable of the nation's waterways. The most important is the Napo River, which receives the Coca and Aguarico rivers as well as other large tributaries as it takes its course towards Peru, where it joins the Amazon River (*q.v.*).

Climate. Despite the fact that Ecuador lies on the Equator, the climate is modified by the Andes Mountains and by a branch of the cold Peru or Humboldt Current that flows along the coast.

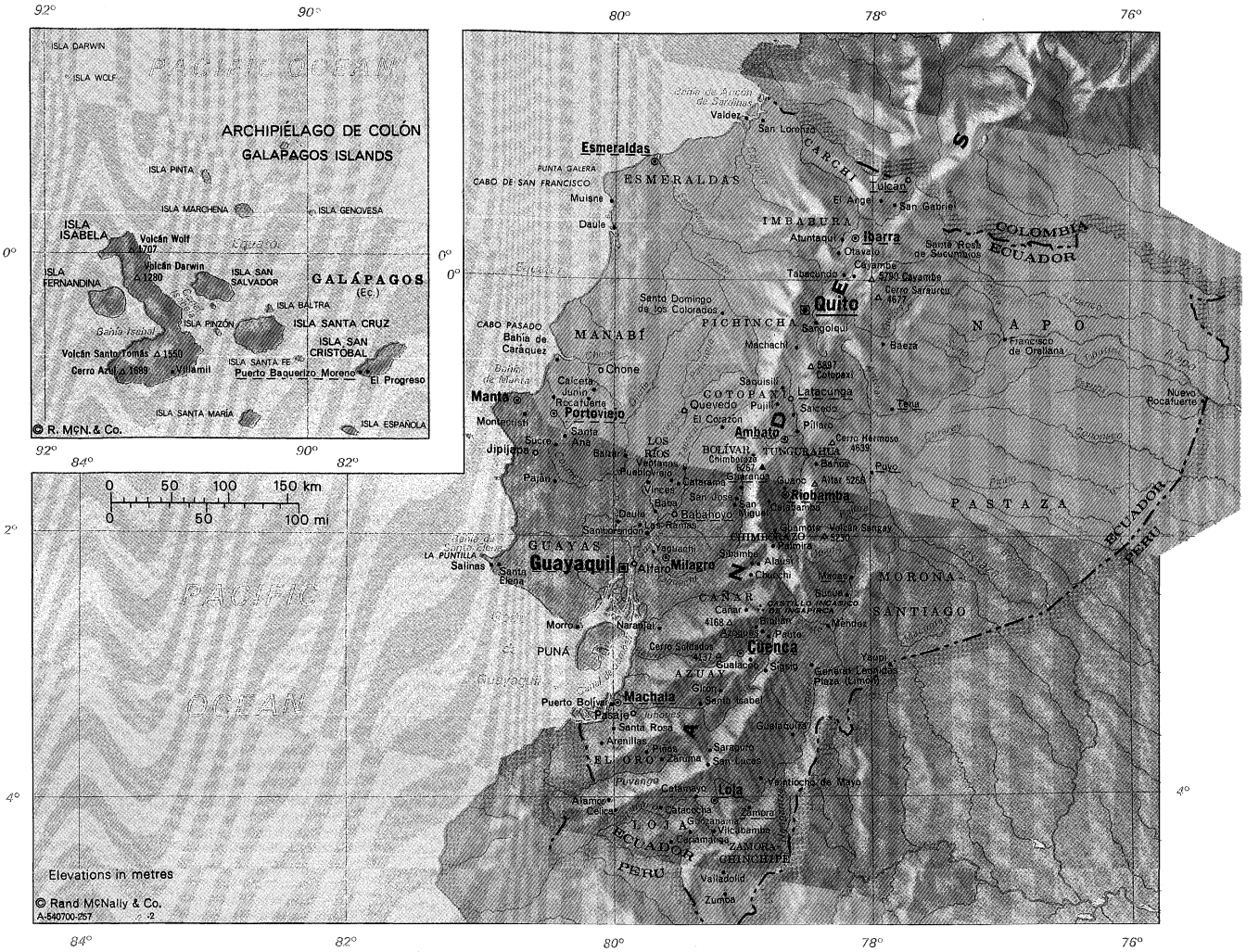
The coastal region is warm; temperatures rise slightly away from the sea and average between 73° and 77° F (23° and 25° C). Rain falls from January to May. Affected by the Peru Current, annual precipitation varies between 16 inches and 78 inches.

In the sierra region, the climate varies with altitude and is modified by winds. Climatic conditions run the gamut from glacial cold at the perpetual snow line to tropical temperatures in certain *hoyas*. Among inhabitants of Vilcabamba in Loja Province, famous for its exceptionally healthy climate in which the temperature varies between 63° and 68° F (17° and 20° C), it is not unusual to find individuals of 120 years or more. Using 79° F (26° C) as a median temperature at sea level, it may be said that the temperature decreases by 1.8° F (1° C) with every 660-foot increase in the altitude.

The east is the warmest and most humid region of Ecuador. Median temperatures range between 73° and 81° F (23° and 27° C), rainfall can reach 236 inches annually, while humidity as high as 90 percent may be sustained for several months.

Tropical rain forest areas

Vegetation and animal life. The wet lowlands of the east and the northern and southern corners of the coastal region are covered with tropical rain forest, containing various trees, lianas (climbing vines rooted in the ground), and many epiphytes (plants that grow nonparasitically on other plants, deriving their nutrients from



ECUADOR

MAP INDEX

Political subdivisions

Azuay..... 3·00s 79·00w
Bolívar..... 1·15s 79·05w
Cañar..... 2·30s 79·00w
Carchi..... 0·45n 78·00w
Chimborazo..... 2·00s 78·40w
Cotopaxi..... 0·55s 78·55w
El Oro..... 3·30s 79·50w
Esmeraldas..... 0·40n 79·30w
Galápagos..... 1·00n 90·00w
Guayas..... 2·00s 80·00w
Imbabura..... 0·22n 78·25w
Loja..... 4·10s 79·30w
Los Ríos..... 1·30s 79·25w
Manabí..... 0·40s 80·05w
Morona-Santiago..... 2·30s 78·00w
Napo..... 0·30s 77·00w
Pastaza..... 2·00s 77·00w
Pichincha..... 0·10s 78·40w
Tungurahua..... 1·15s 78·35w
Zamora-Chinchipe..... 4·15s 78·50w

Cities and towns

Alamor..... 4·02s 80·02w
Alausí..... 2·12s 78·50w
Alfaro..... 2·12s 79·50w
Ambato..... 1·15s 78·37w
Arenillas..... 3·33s 80·04w
Atuntaqui..... 0·20n 78·13w
Azogues..... 2·44s 78·50w
Baba..... 1·47s 79·40w
Babahoyo..... 1·49s 79·31w
Baeza..... 0·27s 77·53w
Bahía de Caráquez..... 0·36s 80·25w
Baizar..... 1·22s 79·54w
Baños..... 1·24s 78·25w
Biblián..... 2·42s 78·52w
Cajabamba..... 1·42s 78·45w
Calceta..... 0·51s 80·10w
Cañar..... 2·33s 78·56w
Cariamanga..... 4·20s 79·35w
Catacocha..... 4·04s 79·38w
Catamayo..... 3·59s 79·21w
Catarama..... 1·35s 79·28w
Cayambe..... 0·03n 78·08w
Celica..... 4·07s 79·59w
Chone..... 0·41s 80·06w
Chunchi..... 2·17s 78·55w
Cuenca..... 2·53s 78·59w
Daule..... 0·24n 80·00w
Daule..... 1·50s 79·56w
El Angel..... 0·37n 77·56w
El Corazón..... 1·12s 79·06w
El Progreso..... 0·54s 89·33w
Esmeraldas..... 0·59n 79·42w
Francisco de Orellana..... 0·28s 76·58w
General Leonidas Plaza (Limón)..... 2·58s 78·25w
Girón..... 3·10s 79·08w
Gonzanamá..... 4·15s 79·27w
Gualaceo..... 2·54s 78·47w
Gualaquiza..... 3·24s 78·33w
Guamote..... 1·56s 78·43w
Guano..... 1·35s 78·38w
Guaranda..... 1·36s 79·00w
Guayaquil..... 2·10s 79·50w
Ibarra..... 0·21n 78·07w
Jipijapa..... 1·20s 80·35w
Junín..... 0·56s 80·13w
Las Ramas..... 1·50s 79·48w
Latacunga..... 0·56s 78·37w
Limón, see General Leonidas Plaza
Loja..... 4·00s 79·13w
Macas..... 2·19s 78·07w
Machachi..... 0·30s 78·34w
Machala..... 3·16s 79·58w
Manta..... 0·57s 80·44w
Méndez..... 2·43s 78·19w
Milagro..... 2·07s 79·36w
Montecristi..... 1·03s 80·40w
Morro..... 2·39s 80·19w
Muisne..... 0·36n 80·02w
Naranjal..... 2·42s 79·37w
Nuevo Rocafuerte..... 0·56s 75·24w
Otavalo..... 0·14n 78·16w
Paján..... 1·34s 80·25w
Palmira..... 2·05s 78·43w
Pasaje..... 3·20s 79·49w
Paute..... 2·47s 79·50w
Pillaro..... 1·10s 78·32w
Piñas..... 3·42s 79·42w
Portoviejo..... 1·03s 80·27w
Pueblovejo..... 1·34s 79·30w
Puerto Baquerizo Moreno..... 0·54s 89·36w
Puerto Bolívar..... 3·16s 79·59w
Pujilí..... 0·57s 78·41w
Puyo..... 1·28s 77·59w
Quevedo..... 1·02s 79·29w
Quito..... 0·13s 78·30w
Riobamba..... 1·40s 78·38w
Rocafuerte..... 0·55s 80·28w
Salcedo..... 1·02s 78·34w
Salinas..... 2·13s 80·58w
Samborondón..... 1·57s 79·44w
San Gabriel..... 0·36n 77·49w
Sangolquí..... 0·19s 78·27w
San José..... 1·42s 79·01w
San Lorenzo..... 1·17n 78·50w
San Lucas..... 3·45s 79·15w
San Miguel..... 1·44s 79·01w
Santa Ana..... 1·13s 80·23w
Santa Elena..... 2·14s 80·51w
Santa Isabel..... 3·21s 79·19w
Santa Rosa..... 3·27s 79·58w
Santa Rosa de Sucumbíos..... 0·22n 77·10w
Santo Domingo de los Colorados..... 0·15s 79·09w
Saquisilí..... 0·51s 78·40w
Saraguro..... 3·36s 79·13w
Sibambe..... 2·12s 78·55w
Sígsig..... 3·01s 78·45w
Sucre..... 1·16s 80·26w
Sucúa..... 2·28s 78·10w
Tabacundo..... 0·03n 78·12w
Tena..... 0·59s 77·49w
Tulcán..... 0·48n 77·43w
Valdez..... 1·15n 79·00w
Valladolid..... 4·33s 79·08w
Veintiocho de Mayo..... 3·50s 78·52w
Ventanas..... 1·23s 79·25w
Vilcabamba..... 4·18s 79·15w
Villamil..... 0·56s 91·01w
Vinces..... 1·32s 79·45w
Yaguachi..... 2·07s 79·41w
Yaupi..... 2·59s 77·50w
Zamora..... 4·04s 78·58w
Zapotillo..... 4·25s 80·31w
Zaruma..... 3·41s 79·37w
Zumba..... 4·52s 79·09w

Physical features and points of interest

Aguarico, *river*.. 0·59s 75·11w
Altar, *volcano*... 1·41s 78·24w
Ancón de Sardinas, Bahía de, *bay*.. 1·30n 79·00w
Andes, *mountains*..... 1·00s 78·50w
Antisana, *river*.. 0·59s 78·01w
Azul, Cerro, *mountain*..... 0·54s 91·21w
Balao, *river*..... 2·54s 79·50w
Baltra, Isla, *island*..... 0·26s 90·16w
Bobonaza, *river*. 2·36s 76·38w
Carihuairazo, *volcano*..... 1·24s 78·45w
Castillo Incasico de Ingapirca, *ruins*..... 2·34s 78·50w
Catamayo, *river*. 4·18s 80·09w
Cayambe, *volcano*..... 0·02n 77·59w
Cayapas, *river*... 1·13n 79·03w
Chimborazo, *volcano*..... 1·28s 78·48w
Chone, *river*..... 0·35s 80·25w
Coca, *river*..... 0·29s 76·58w
Colimes, *river*... 1·32s 80·00w
Colón, Archipiélago de (Galápagos Islands)..... 0·30s 90·30w
Conambo, *river*. 2·07s 76·03w
Condor, Cordillera del, *mountains*..... 4·00s 78·30w

MAP INDEX (continued)

Cononaco, *river*. 1·32s 75·35w
Cotopaxi, *volcano*..... 0·40s 78·26w
Curaray, *river*... 2·20s 74·05w
Darwin, Isla, *island*..... 1·39n 92·00w
Darwin, Volcán, *volcano*..... 0·10s 91·18w
Daule, *river*..... 2·10s 79·52w
Esmeraldas, *river*..... 0·58n 79·38w
Española, Isla, *island*..... 1·25s 89·42w
Fernandina, Isla, *island*..... 0·25s 91·30w
Galápagos Islands, see Colón, Archipiélago de
Galera, Punta, *point*..... 0·49n 80·03w
Genovesa, Isla, *island*..... 0·20n 89·58w
Guayaquil, Golfo de, *gulf*..... 3·00s 80·30w
Guayas, *river*.... 2·36s 79·52w
Hermoso, Cerro, *mountain*..... 1·10s 78·12w
Isabel, Bahía, *bay*..... 0·38s 91·27w
Isabela, Canal, *channel*..... 0·20s 90·55w
Isabela, Isla, *island*..... 0·30s 91·06w
Jambelí, Canal de, *channel*.... 3·00s 80·00w
Jubones, *river*... 3·13s 79·57w
La Puntilla, *point*..... 2·11s 81·01w
Macuma, *river*.. 2·44s 77·24w
Manta, Bahía de, *bay*..... 0·54s 80·44w
Marchena, Isla, *island*..... 0·21n 90·29w
Nangaritza, *river*..... 3·45s 78·38w
Napo, *river*..... 3·20s 72·40w
Pacific Ocean... 3·00s 83·00w
Palenque, *river*.. 1·02s 79·28w
Palora, *river*.... 1·51s 77·49w
Pasado, Cabo, *cape*..... 0·22s 80·30w
Pastaza, *river*... 2·32s 76·40w
Paute, *river*..... 2·40s 78·16w
Pindo, *river*..... 2·07s 76·03w
Pinta, Isla, *island*..... 0·35n 90·44w
Pinzón, Isla, *island*..... 0·36s 90·40w
Puná, *island*.... 2·50s 80·08w
Putumayo, *river*. 3·07s 67·58w
Puyango, *river*.. 3·52s 80·12w
Quijos, *river*.... 0·14s 77·43w
San Cristóbal, Isla, *island*..... 0·50s 89·26w
San Francisco, Cabo de, *cape*.. 0·40n 80·05w
Sangay, Volcán, *volcano*..... 2·00s 78·20w
San Juan, *river*.. 1·18n 78·40w
San Salvador, Isla, *island*..... 0·14s 90·45w
Santa Cruz, Isla, *island*..... 0·38s 90·23w
Santa Elena, Bahía de, *bay*.. 2·06s 80·53w
Santa Fe, Isla, *island*..... 0·49s 90·04w
Santa María, Isla, *island*..... 1·17s 90·26w
Santo Tomás, Volcán, *volcano*..... 0·48s 91·07w
Saraurcu, Cerro, *mountain*..... 0·06s 77·55w
Soldados, Cerro, *mountain*..... 2·54s 79·10w
Suno, *river*..... 0·42s 77·08w
Tiputini, *river*... 0·47s 75·32w
Toachi, *river*.... 0·08n 79·18w
Upano, *river*.... 2·45s 78·12w
Vinces, *river*.... 1·39s 79·47w
Wolf, Isla, *island*..... 1·23n 91·49w
Wolf, Volcán, *volcano*..... 0·02n 91·20w
Yaguachi, *river*.. 2·02s 79·44w
Yasuní, *river*.... 0·56s 76·23w
Zamora, *river*... 2·59s 78·13w
Zapotal, *river*.... 1·40s 79·28w

the air and the rain). This forest thickens as it approaches the zone of maximum rainfall, which occurs between 4,000 and 5,000 feet above sea level. In the Guayas Valley, the forest contains balsa, which is exploited for its light lumber; in the eastern forest the cinchona trees were a valuable source of quinine before synthetic equivalents reduced demand for it. The trees of the coastal region are fast disappearing because of exploitation, while the east remains untouched.

In the coastal region between Esmeraldas and the Gulf of Guayaquil, where the climate is affected by the Peru Current, the rain forest gives way to deciduous and semi-deciduous woodland. There, scattered palms produce the ivory-coloured nuts used in making buttons, while the hat carludovica (*Carludovica palmata*) furnishes the fibre used for Panama hats. Areas of swampy coast and the river floodplains are covered by thick mangrove forest.

In the sierra region, the vegetation of the *hoyas* resembles the rain forest. Above the valleys, the *ceja de la montaña* ("eyebrow of the forest") consists of dense moss-covered woodland that at 10,000 to 11,000 feet gives way to tall, strawlike grasses known as *elichu*. In the highest regions, only lichens grow where there is no snow cover.

In the rain forest live a wide variety of monkeys, as well as such carnivorous mammals as the jaguars, ocelots, foxes, weasels, otters, skunks, raccoons, coatis (members of the raccoon family), and kinkajou (a tree-dwelling nocturnal animal with a prehensile tail). Hoofed mammals include tapir, deer, and peccary, an animal related to the pig. There are many species of rodents and bats.

Birdlife

About 1,500 species of birds have been identified; different species of birds are associated with the various types of vegetation. Among many types of North American birds that migrate to Ecuador for the winter are the Virginia rail, the kingbird, the barn swallow, and the scarlet tanager.

The fish population is similar to that of the Amazon River although in the west the electric eel and the piranha are not found. All major groups of reptiles are represented, including dangerous snakes; two types of marine turtle are hunted for their meat and shells.

Numerous insects include butterflies, moths, grasshoppers, bees, and ants. A giant beetle grows to five inches in length; some other beetles emit phosphorescent light.

Settlement patterns. *Traditional regions.* The traditional regions of Ecuador coincide with the country's three physical regions of the coast, the mountains, and the Amazon Basin. The highest proportion of peoples of Indian descent lives in the isolated mountain basins. The east is thinly inhabited by indigenous peoples.

Rural settlement. The coastal region is the site of banana plantations; its other tropical products include cacao, tobacco, yucca (a fibre-bearing plant of the lily family), and fruit. Most of the population lives in grass or cane houses with roofs of palm leaves or zinc. The sierra is the granary and cattle centre of Ecuador. Grains, beans, and peas are grown in small patches that, from the air, give the land the appearance of a giant, multicoloured chessboard. The peasants occupy houses of adobe (brick made of sun-dried earth and straw) roofed with straw. A small amount of settlement has taken place in the east, where grain is grown on land reclaimed from the rain forest.

The land tenure system

The rural peoples lead lives characterized by poverty and hard labour. Some of them own small patches of land, or *minifundios*, the products of which are insufficient even for subsistence. Others work as tenant farmers, or *huasipongueros*, on the haciendas (large estates). In return for their labour, they receive a house and a small plot of land on which they may cultivate wheat and raise domestic animals. The large landowners generally live in the cities and only occasionally visit the haciendas. An agrarian reform program has been launched to attempt to improve conditions through the agency of the Ecuadorian Institute of Agrarian Reform (IERAC).

Urban settlement. Quito (population 530,000), the national capital, known for its distinctive architecture, is one of the most beautiful cities in Latin America. Guayaquil (population 795,000) is the principal port and national commercial centre. Cuenca (population 77,000) is known for its traditional culture and for its growing industry. Other notable urban centres include Riobamba, Ambato, Loja, and Ibarra.

The main cities are increasing their populations at an accelerated rate. Quito houses most of the country's government officials; state salaries are not high, and most civil servants live from day to day. In Guayaquil, however, the majority is engaged in commerce, and economic opportunities are greater. Social and cultural life is intensified in the main urban areas. On weekends people often go to the countryside, or attend such sporting events as football (soccer) matches, bullfights, or horse races.

THE PEOPLE

Composition. Ecuador has long been mistakenly represented as either a predominantly Indian country or as a land of headhunters. In effect its population is ethnically mixed, with mestizos—persons of mixed Indian and European (white) origin—forming the majority.

The Indian population

The Indians, whose numbers are decreasing, form the second largest ethnic group. They are divided into a number of groups, including Otavalos in Imbabura; Salasacas in Tungurahua; and the descendants of the Puruhá in Chimborazo, who represent the largest and the poorest Indian group. Other Indian groups are the Cañari in Cañar and Azuay provinces; the Saraguros in Loja; the Cayapas in Manabí; and the Colorados in Pichincha Province.

The white population, third in order of size, is largely descended from Spaniards who arrived in Ecuador from the 16th to the early 19th centuries. Other white emigrants travelled from North America and Europe during the 20th century. The blacks were brought from Africa as slaves during the colonial era and are mainly found

in the northern coastal province of Esmeraldas. There is also a small black community in the Chota Valley of Imbabura Province, as well as a few families living in Catamayo. Mulattos, or *zambos*—people of mixed black and white parentage—are found living throughout the coastal region.

The Jívaro, who form the main Indian group in the eastern region, consist of small tribes or family units; they are gradually being assimilated into Ecuadorian society. They originally engaged in headhunting, a practice now extinct. Shrunken heads sold to tourists today are generally imitations fashioned by talented artisans from goat or other animal skins. Other eastern Indians include the Záparos, Cofanes, and Aucas.

Linguistic patterns

Spanish, the official national language, is spoken by nearly all the population. Quechua, an Indian language, is also spoken within Indian groups, while the Jívaros and Colorados use dialects of their own.

The overwhelming majority of the people are Roman Catholic; Protestants comprise less than 1 percent of the population.

Demography. The total population of Ecuador was estimated at 6,384,000 inhabitants in 1971. In quantitative order, the population is composed of mestizos, Indians (except for Jívaros), whites, blacks, mulattos, and the Jívaro Indians. About 51 percent of the people live in the sierra region, 46 percent in the coastal region, and 2 percent in the eastern region (the remainder in the Galápagos).

The annual birthrate is 45 per 1,000 inhabitants, and the infant mortality rate is 86 per 1,000. The annual population growth rate of 3.4 percent is not appreciably affected by immigration or emigration rates, which are low and variable.

Ecuador, Area and Population

	area*		population	
	sq mi	sq km	1962 census	1971 estimate
Regions				
Galápagos Islands (Archipiélago de Colón)				
Island territory†				
Archipiélago de Colón	3,075	7,964	2,000	4,000
Coastal				
Provinces				
El Oro	3,053	7,908	161,000	251,000
Esmeraldas	5,680	14,710	125,000	185,000
Guayas	7,368	19,082	979,000	1,464,000
Los Ríos	3,076	7,968	250,000	371,000
Manabí	7,602	19,688	613,000	849,000
Eastern‡				
Provinces				
Morona-Santiago	‡	‡	26,000	43,000
Napo	‡	‡	24,000	40,000
Pastaza	‡	‡	14,000	23,000
Zamora-Chinchipe	‡	‡	11,000	19,000
Sierra				
Provinces				
Azuay	3,211	8,316	275,000	320,000
Bolívar	1,288	3,336	132,000	184,000
Cañar	1,614	4,180	113,000	138,000
Carchi	1,581	4,094	95,000	123,000
Chimborazo	2,322	6,014	277,000	242,000
Cotopaxi	2,241	5,804	155,000	384,000
Imbabura	3,458	8,956	174,000	218,000
Loja	4,445	11,512	285,000	389,000
Pichincha	5,976	15,478	588,000	876,000
Tungurahua	1,366	3,538	179,000	263,000
Total Ecuador	109,483	283,561	4,476,000§	6,384,000‖

*Area is subject to boundary dispute with Peru. †The Territorio Insular del Archipiélago de Colón is a separate administrative entity. Though it does not have the same autonomy as a province, it is considered a first-order subdivision. ‡The total area of eastern region is 135,013 sq km; no breakdown is available. §Excludes nomadic Indian tribes which were not enumerated. ‖Figures do not add to total given because of rounding.
Source: Official government figures.

THE ECONOMY

While Ecuador is a country of tremendous economic potential, it has never experienced a period of sustained development. The agricultural possibilities of El Oro Province and the sierra have been well-developed, but the mineral wealth of the mountains remains untouched. The eastern region, with its rich, virgin land, is now being settled, but there are thousands of acres of undeveloped jungle that represent an unknown reserve potential. The majority of Ecuadorean land remains uninhabited, and technological methods of agriculture are used only on the banana plantations of the coastal region.

Natural resources. *Mineral resources.* Since the birth of the republic almost 150 years ago, no comprehensive scientific exploration of Ecuador's mineral wealth has been undertaken. The extent and depth of deposits of white clay, marble, coal, and sulfur remain unknown because they have been examined only superficially if at all. In the late 1960s, however, the government at last showed interest in its mineral wealth and began to contract for its exploitation. Oil, found in northwestern Oriente Province, is being extracted by the Texaco-Gulf Oil Corporation, which constructed a pipeline to Esmeraldas. Copper mines at Chaucha are operated by a Japanese company. Deposits of silver and molybdenum have also been located.

Oil and copper resources

Forest resources. Pines are cultivated in the sierra region, where there are also natural forests of romerillo (herb yielding a yellow dye), walnut, and cedar. Large eucalyptus forests occur in all three regions.

Power resources. The Andes Mountains present vast possibilities for hydroelectric development. The present small hydroelectric plants are inadequate for industry, or even for domestic use. At the instigation of the Ecuadorian Centre of Economic Reconversion, a Canadian firm carried out a study along the Río San Pablo, an affluent of the Paute River, and subsequently proposed the installation of a hydroelectric station with a capacity of 1,050,000 kilowatts. The Ecuadorian Institute of Electricity (Instituto Ecuatoriano de Electrificación), a government agency, is responsible for the development of power resources.

Sources of national income. *Agriculture, forestry, and fishing.* Agriculture employs about 56 percent of the working population. The chief products are bananas and coffee; other products include barley, cocoa, cottonseed, corn, potatoes, rice, wheat, kidney beans, chickpeas, lima beans, and fruit.

The forests supply the domestic needs of the construction and cabinetmaking industries. Fishing provides work and a subsistence diet for small coastal settlements. The tuna, sardine, and shrimp industries in Guayaquil and Manta produce over 26,500 tons annually.

Stock raising is extensive in the sierra and coastal regions, and is growing in importance in the eastern region. In the early 1970s, Ecuador's livestock resources amounted to about 2,400,000 cattle, 1,900,000 sheep, 240,000 horses, and 110,000 mules.

Mining and quarrying. The mining and quarrying industries are still nascent. Annual production includes about 770 pounds of gold, 3 tons of silver, 40,000 tons of salt, 150 tons of sulfur, and small quantities of zinc and marble. The production of crude oil—about 300,000 tons annually—and of copper increased considerably in the 1970s.

Industry. The major industrial centres in Ecuador are Quito, Guayaquil, and Cuenca; the other cities have experienced only slight industrial development. The main industrial products of the country include textiles, shoes, tires, processed foods and beverages, galvanized pipes, vehicle frames, pharmaceutical and chemical products, cement, and oil derivatives.

Financial services and foreign trade. *Financial services.* The Central Bank and the National Bank of Promotion, both state controlled, have branches in all of the provincial capitals. The former is the government depository and controls the monetary system, while the latter handles agricultural and industrial credit. Private or commercial banks are both foreign and domestic. The bank supervisory board is a technical organization that monitors all banking activities.

Government and private banking

Foreign trade. Exports include bananas (47 percent of the total), coffee (25 percent) and cacao (11 percent); these products are sold primarily to the United States,

Japan, West Germany, Belgium, and Luxembourg. Imports include motor vehicles, food and chemical products, consumer goods, and machines and primary materials for industry. They are mainly bought from the United States, Japan, West Germany, the United Kingdom, Belgium, The Netherlands, and Luxembourg. In 1970 all exchange restrictions on international transactions except those relating to foreign loans and investments were ended. Tax allowances were provided for exports, while surcharges on imports were abolished. In the same year Ecuador began the exportation of specified products to Bolivia, Chile, Colombia, and Peru on a duty-free basis.

Management of the economy. The government regulates national production through various organizations, such as the Planning and Coordination Council, which deals with economic and social developmental planning; the Central Bank; and the National Developmental Centre. The private economic sector includes a large part of agriculture and stock raising, as well as business and industry.

All of the provincial capitals have their own commercial, industrial, and agricultural associations, or *cámaras;* they are composed of persons associated with various economic sectors, and protect the interests of their membership. Each association is autonomous and is locally influential.

The Ministry of Finance collects taxes on income and commercial transactions through the General Director of Revenues; taxes on rural and urban property are collected through the municipalities. The Ecuadorian taxation system was being reformed in the 1970s in order to attempt to ensure a more complete tax collection.

Transportation. Because of difficult topography and lack of economic development, Ecuador has only a partially developed road system. The farmer, or *campesino,* who wishes to take his products to market frequently must resort to horse or mule transport, or must even backpack to reach motorized transportation. These circumstances reduce the profits earned and constitute an obstacle to production in many areas. The problem is most acute in the southern sierra and in the eastern regions.

Road system

There are about 13,000 miles of roads of which 1,800 miles are paved. The Pan-American Highway crosses the country from north to south through the sierra region. Other important roads connect the port of Guayaquil with Quito, Cuenca, Loja, and the port of Salinas. Machala and Puerto Bolívar are linked to Huaquillas and Cuenca, and there is a narrow road from Pallatanga to Bucay (situated 54 miles east of Guayaquil on the Guayaquil–Cuenca Road). In the east, the main roads lie within the Andes region and radiate from Ambato, Cuenca, and Loja.

Railroads total about 700 miles. The principal lines connect Quito with Guayaquil and Cuenca through the railway junction at Sibambe. There is also a line from Quito north along the mountains and across the coastal plain to the port of San Lorenzo. Service is slow and poor, and the railways are used more for the transport of freight than for passengers.

Guayaquil, with its modern facilities at Puerto Nuevo, is the country's chief port. Other ports of importance are San Lorenzo, Esmeraldas, Manta, and Puerto Bolívar. All have a port captaincy (an administrative district under the command of a captain) as well as customs and immigrations offices; all are linked with the interior by road or railway.

Most of the major cities have airports. There are several national airlines, and the main international airline is *Ecuatoriana de Aviación.* Nine international airlines stop at Quito and Guayaquil, but the airports at Cuenca, Loja, and Riobamba lack the necessary facilities for larger aircraft.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. *Administration.* Ecuador is a sovereign democratic republic. The president and vice president are elected for four-year terms by popular, direct, and secret voting. The State Council of 19 members has a consultant function, and the chief executive is aided by an 11-member cabinet. Legislative power is vested in the national congress, which is made up of two houses—the senate, with two representatives from each province, and one from the Archipiélago de Colón, as well as 15 "functional" senators elected by special economic or cultural groups—and the House of Deputies, whose members are chosen from each province according to population distribution.

The chief executive appoints as his representatives the provincial governors, political chiefs in each of Ecuador's 19 provinces, which are divided into *cantones* (regions), which in turn are divided into parishes. The chief executive also appoints the political chief in each *cantón,* and a political lieutenant in each parochial district. Every province also has a provincial council presided over by the provincial prefect.

Each *cantón* constitutes a municipality, the governing of which is in the hands of a *cantón* council. The council of each province capital is headed by a mayor.

Politics. Literate, native-born, or naturalized Ecuadorians over 18 years of age have the right to vote in all elections by direct, secret ballot. Political groups include on the right of the spectrum, the Conservative Party, the Social Christian Movement, and the Nationalist Revolutionary Action Party of Ecuador. Parties on the left include the Liberals, the Socialists, and the Communists. A large part of Ecuadorian politics, however, revolves around the caudillos, or political leaders. The Concentration of Popular Forces Party, for example, was led by Asaad Bucaram. The Velasquista Party centred on José María Velasco Ibarra, five-time president of the republic. These two parties drew members from nearly all of the other political parties.

Political parties and groups

Justice. The constitution provides for free access to justice for all. The Supreme Court, with its seat in Quito, has national jurisdiction. Eight superior courts are located in the provincial capitals, and there are also provincial tribunals, *cantonales* or regional courts, and criminal and transportation courts. Special tribunals handle cases concerning the press and radio, and the protection of minors. Constitutionally, no judgement may be appealed more than three times.

The armed forces. The armed forces consist of the national army, the air force, and the navy. The army consists of six divisions and some small operational forces; the navy has a small number of frigates, gunboats, patrol vessels, and other craft; the air force has about 60 aircraft. The National Civil Police have the responsibility of maintaining internal security and order. With the exception of Guayaquil, which has an autonomous transportation commission, the police also control traffic. The rural police, a division of the national civil police, exist only in the small towns and the interior. The police force throughout the country is below strength because of a lack of funds.

Social conditions. *Education.* Education is free and is obligatory at the primary level for all children ages six to 12. All state schools are nondenominational, but private schools—which also receive state funds—are divided into nondenominational and religious schools.

In the early 1970s there were about 7,000 primary schools, about 360 high schools, and almost 240 technical schools, of which about 40 were teacher-training schools.

There are 14 institutions of higher learning. They include the Central University of Ecuador at Quito, the University of Guayaquil, the University of Cuenca, the University of Loja, the Technical University of Manabi, the National Polytechnic School of Quito, and the Polytechnic School of the Litoral at Guayaquil. Church-affiliated universities include the Pontificate Catholic University at Quito, the Catholic University of Guayaquil, and the Catholic University of Cuenca.

Welfare services. All public and private employees are affiliated with the National Social Security Institute. In return for a monthly deduction from employees' salaries, the agency provides such services as medical and

hospital insurance coverage, state-run clinics and dispensaries, low-interest loans for surgery and mortgages, retirement pensions for civil and state employees, and pensions for widows and child dependents.

The Social Welfare Program, a division of the Ministry of Public Health, maintains public hospitals in all of the provincial capitals and in the principal *cantóns*. The Anti-Cancer Society and the Anti-Tuberculosis Society are also part of the ministry. Only a very small percentage of the national budget is, however, devoted to public health programs. As a result, research institutions lack funds, and health conditions are generally poor. There are a number of endemic diseases including goitre, typhoid fever, and tuberculosis.

Salaries are not related to a cost-of-living index, and low wage scales present a problem that daily becomes more acute. Minimum wages established by law are not always paid. In an attempt to provide minimum wage security, a work administration was created, but by the early 1970s it had not yet proved effective.

The national press frequently focusses on social problems. *Vistazo* Magazine from Guayaquil has denounced the imbalance between work performed and salaries paid. Some governmental employees, for example, may receive more than 50,000 sucres monthly, while others do not even receive the lawful minimum monthly wage of 750 sucres (25 sucres = $1 U.S.; 60 sucres = £1 sterling, on December 1, 1970).

Economic divisions

The traditional social divisions based on family lineage are being translated into economic terms, thus allowing for greater mobility. Alongside the opulent few, however, a large percentage of the population lives under almost subhuman conditions, for example in some city ghettos and in some Indian communities, especially in Chimborazo Province.

Cultural life. Despite an illiteracy rate of almost 33 percent and a poor economy, Ecuador nevertheless maintains a generally high cultural level. A tendency to erudition is evident, and music forms an integral part of Ecuadorian life. Many people play musical instruments, accompanying their own songs. The various art forms are distinctive for their treatment of the themes provided by the landscape or Ecuadorian folklore. National literature often focusses on the Andes or on Indian themes. Perhaps the best example of Ecuadorian art is the poem *Boletín y elegía de las mitas* ("Record and Elegy on Indian Forced Labour") by the late Cuencan poet, César Dávila Andrade.

Ecuadorian music is characterized by its sadness. Among folksong types, the *yarabí* evokes the immense solitude of the Andes and the pain of the oppressed Indian. Paradoxically, many tunes, through their sadness, lighten Ecuadorian fiestas and popular dances.

The most important cultural institution in the country is the House of Ecuadorian Culture (founded in 1944), the aim of which is to direct and preserve all aspects of national culture. It has 14 provincial branches. The National Library, the National History Archives, the Museum of Colonial Art, the Pedro Traversari Musical Instrument Museum, and the Pío Jaramillo Alvarado Museum all constitute subdivisions of this central institution. Other institutions include the Centre for the Promotion of Popular Art, as well as the Ecuadorian Institute of Artisans and Manual Artists, which has workshops for applied art, ceramics, painting, and weaving. Each of the major cities has French and North American institutes, as well as representative agencies from other countries.

Major newspapers include *El Comercio*, *El Tiempo*, and *Ultimas Noticias*, all published in Quito; *El Universo* and *El Telégrafo*, published in Guayaquil; *El Mercurio* and *El Tiempo* in Cuenca; *El Heraldo* in Ambato; *La Opinión del Sur* and *La Verdad* in Loja; and *Diario Manabita* in Manabí. The most widely circulated magazine in the country is *Vistazo*, published in Guayaquil.

There are more than 200 radio stations in the country. Major cultural stations include Radio Nacional del Ecuador (which is state-run); La Voz de los Andes; and Radiodifusora Casa de la Cultura in Quito. Commercial stations include Radio Atahualpa and Radio Quito in Quito; Atalaya and Cadena Radiodifusora Ecuatoriana in Guayaquil; and Radiotelevisora Ondas Azuayas in Cuenca. Each of the major cities has two or more television stations.

Prospects for the future. A common expression that is traditionally attributed to Baron Alexander von Humboldt, the German geographer and naturalist, who visited Ecuador in 1802, still remains relevant:

> The people are the strangest in the world. They live in poverty on mountains of gold; sleep tranquilly at the foot of volcanoes; and cheer themselves with sad music.

Despite the mineral wealth enclosed in the Andes Mountains, and despite a considerable potential for hydroelectric power, many Ecuadorians continue to live in poverty. The Andes Mountains themselves, despite their beauty and their influence in modifying the climate, continue to obstruct communication and development, and also symbolize the latent danger of earthquakes. As recently as 1949, Ambato and 50 other towns were destroyed by an earthquake.

Ecuador's future depends upon further development, both in industrialization and in the improvement of its agricultural economy. Disparate standards of living that parallel social divisions, past and present, are becoming increasingly apparent and represent a source of political tension that can only be lessened by the development of a trend towards the establishment of more equitable conditions.

BIBLIOGRAPHY

Topography: J. COLLIER, JR., and A. BUITRON, *The Awakening Valley* (1949), combines a comprehensive text with a variety of illustrations. L. LINKE, *Ecuador: Country of Contrasts*, 3rd ed. (1960), is also wide-ranging. For detail, E. WHYMPER, *Travels Amongst the Great Andes of the Equator* (1892), is still worth reading; for a geographical survey, see D.A. PRESTON, *Negro, Mestizo and Indian in an Andean Environment* (1965). The jungle is covered in H. MICHAUX, *Ecuador: journal de voyage* (1968; Eng. trans., *Ecuador: A Travel Journal*, 1970), a vivid personal account; H. ACEBES, *Jungle Explorer* (1956); J. BITSCH, *Jivaro* (1958; Eng. trans., *Across the River of Death: True Adventures in the Green Hell of the Amazon*, 1958); and V.W. VON HAGEN, *Ecuador the Unknown* (1940). M.M. SUMWALT, *Ecuador in Pictures* (1969), has chiefly illustrations. Writings on flora include T.H. GOODSPEED, *Plant Hunters in the Andes*, 2nd ed. rev. (1961); and R. SPRUCE, *Notes of a Botanist on the Amazon and Andes*, 2 vol. (1908, reprinted 1970).

Archaeology: B.J. MEGGERS, *Ecuador* (1966), is a wide archaeological survey, while the detailed W. BOLLAERT, *Antiquarian, Ethnological, and Other Researches in New Granada, Ecuador, Peru and Chile, with observations on the pre-Incarial, Incarial and Other Monuments of Peruvian Nations* (1860), is still interesting.

Peoples: B. FLORNOY, *Haut-Amazone* (1939; Eng. trans., *Jivaro: Among the Headshrinkers of the Amazon*, 1953); L.N. COTLOW, *Amazon Head-Hunters* (1953); E.E. WALLIS, *The Dayuma Story: Life Under Auca Spears* (1960); and R. BLOMBERG, *Vildar, en berättelse om aucaindianera i Ecuador* (1949; Eng. trans., *The Naked Aucas: An Account of the Indians of Ecuador*, 1956), provide accounts of various tribes, while P. ALLARD and P. LUZUY (photographers), *Faces of Bronze*, with text by P. LUZY and P.A. BOUTANG (1960), is a moving picture book. D.A. PRESTON, *Social Geography and Human Relations in an Andean Town: Pimampiro, Ecuador* (1966), is a picture of a more urban life. J. DOLINGER, *The Forbidden World of the Jaguar Princess* (1964), is a biography of Pamela Hawkins, a girl reared in the jungle, and *The Head with the Long Yellow Hair* (1958), describes an expedition. E. ELLIOT, *The Savage, My Kinsman* (1961) and *Through Gates of Splendor* (1957); and F. and M. DROWN, *Mission to the Head-Hunters* (1961), represent the missionary stance.

Constitution: G.I. BLANKSTEN, *Ecuador: Constitutions and Caudillos* (1964), analyzes modern political developments.

Economy: The ECONOMIST INTELLIGENCE UNIT, *Three-Monthly Economic Review of Peru, Bolivia and Ecuador*, is useful, while C.R. GIBSON, *Foreign Trade in the Economic Development of Small Nations: The Case of Ecuador* (1971); and R.J. WATKINS (ed.), *Expanding Ecuador's Exports: A Commodity-by-Commodity Study with Projections to 1973*, prepared by the Surveys and Research Corporation (1967), are also helpful in their fields.

(H.P.V.)

Ecuador, History of

In 1551, when Francisco Pizarro's lieutenant Sebastián de Belalcázar conquered what is now called Ecuador, he found an area that had been part of the Inca empire for about half a century. The last Inca emperor, the tragic Atahuallpa, was born near Quito.

In the mountainous Andean area of central Ecuador (the sierra), the Spaniards established a colony of large estates worked by Indian peons. People lived in semi-autonomous Indian villages or in Spanish and mestizo administrative and religious centres such as Quito, Ambato, and Cuenca. The making of rough textiles in primitive sweatshops was the only industry. Life was rural, hierarchical, and slow to change.

On the Pacific coastal plain (the tropical costa), there were fewer Indians to do the work, and the area was extremely unhealthy until the advent of modern medicine. As a result, the coast was neglected during the colonial period, although there was some shipbuilding and exporting of cacao from the port of Guayaquil. The small coastal population of mixed races, with plenty of vacant land and little coercion of labour, developed a very different culture from that of the sierra (see further LATIN AMERICA AND THE CARIBBEAN, COLONIAL).

On the eastern slopes between the Andes and the headwaters of the Amazon (the oriente), recalcitrant Indians and the equatorial climate prevented settlement, and the only Spaniards who attempted to live there in any numbers were missionaries. Later, this demographic vacuum was to cause Ecuador many problems.

The country's fourth major subdivision, the Galápagos Islands (in the Pacific Ocean about 580 miles [933 kilometres] west of the mainland), were little more than pirate nests during the colonial period; but they were to achieve world fame in the 19th century because Charles Darwin made there a major portion of his observations that led to his theories on evolution and the *Origin of Species*.

The people of Quito, the Ecuadorian capital, claim that it was the scene of the first Ecuadorian patriot uprising against Spanish rule (1809). Invading from Colombia in 1822, the armies of Simón Bolívar and Antonio José de Sucre came to the aid of Ecuadorian rebels, and on May 24 Sucre won the decisive Battle of Pichincha on a mountain slope near Quito, thus assuring Ecuadorian independence.

Early national history (1830–c. 1925). Ecuador's early history as a nation was a tormented one. For some eight years it formed, together with Colombia and Venezuela, the confederation of Gran Colombia. But in 1830, after a period of protracted regional rivalries, Ecuador seceded and became a separate, independent republic.

Liberal–Conservative hostilities. An increasing rivalry and ideological difference between the sierra and the costa usually focussed on the two leading cities—Quito, the capital, in the sierra, and Guayaquil, the country's principal port. Quito was the home of a landed aristocracy, whose positions of power during this early period were based on large, semifeudal estates worked by servile Indian labour; it was (and to some extent has remained) a conservative, clerical city, resistant to changes in the status quo. Guayaquil, on the other hand, by the 19th century had become a bustling, cosmopolitan port, controlled by a few wealthy merchants; these men and those around them were influenced by 19th-century liberalism; interested in trade, they favoured free enterprise and expanding markets; some were anticlerical. Their bourgeois attitudes conflicted sharply with the more aristocratic beliefs of the sierra elites. These early rivalries tended to be exacerbated by the nature of the two cities. The people of Guayaquil, the nation's breadwinner and the home of Ecuador's industry and trade, felt that a disproportionate part of the state's tax income was spent in Quito by government bureaucrats. Those in Quito complained that their exports had to pass through the monopolistic bottleneck of Guayaquil, which acted as a traditional middleman and, by adding to the price of sierra products, reduced their competitiveness in the world market.

Development of Guayaquil

Rivalry between Flores and Rocafuerte (1830–45). Ambitious generals and politicians have played on this Quito–Guayaquil rivalry since the foundation of the republic in 1830. During the period 1830–45, two generals from the wars of independence—Juan José Flores and Vicente Rocafuerte—struggled for power; Flores found much of his support in Quito, Rocafuerte in Guayaquil. Hostility was not constant, and for a few years the rivals agreed to alternate in the presidency. They were not simply personalist dictators; Rocafuerte in particular had a coherent ideology of government and did much to improve the educational institutions of the main cities. Both, however, were capable of obnoxious conduct in their efforts to retain or regain power, inviting foreign armies to enter the country and assassinating opponents and hostile journalists; Flores, on one occasion, even invited the Spaniards to return.

Breakdown of national government (1845–60). The rivalry between Flores and Rocafuerte was a struggle between two strong leaders. Between 1845 and 1860, however, the nation went through a period of chaos in which a series of squabbling, weak leaders (usually self-proclaimed liberals) fought for the presidency with little regard for the nation's integrity or development. This period reinforced the already close ties between the military and the national government.

The regime of García Moreno (1860–75). The next years (1860–75) saw one of Latin America's most extraordinary experiments in autocracy, during the presidency of Gabriel García Moreno. As a young man García Moreno, a religious intellectual and conservative, had witnessed the chaos in Ecuador and the selfish struggles of the various cliques. He had also seen the European revolutions of 1848 and had developed an abhorrence of liberalism and of uncontrolled violence. A careful analysis of Ecuadorian society led him to conclude that the young nation lacked unifying factors; it had no great tradition, suffered from regional resentments, and was sharply divided by class and between whites and Indians who did not even share a common language. García Moreno concluded that the only social "cement" was religion—the general adherence of the population to the Roman Catholic Church. He felt that in time, nationalism could be created and that more social cohesion would emerge as a result, but that meanwhile Ecuador needed a period of peace and strong government. When he became president, therefore, he based his regime on two factors—strong authoritarian personal rule and the Roman Catholic Church. All education and welfare, and the direction of much government policy, were turned over to clerics. Other religions were harshly discouraged. All opposition was ruthlessly suppressed, and some leading liberals spent many years in exile.

Although many aspects of García Moreno's regime were negative, it did mark the first period of genuine progress for Ecuador. Roads, schools, and hospitals were built. A start was made on a Quito–Guayaquil railroad, to tie together the costa and the sierra. García Moreno encouraged the planting of eucalyptus trees from Australia to combat erosion in the sierra, where the original ground cover had been cut down for fuel by the impoverished Indians. Other agricultural reforms slowly raised production. By the end of his regime a strong feeling of nationalism had been created among the urban classes.

In the 19th century, however, this authoritarian, clerical government seemed an anachronism, and liberal opposition grew both at home and abroad. When García Moreno was assassinated on the steps of the government palace in 1875, a liberal intellectual and pamphleteer Juan Montalvo proclaimed from exile, "My pen has killed him."

Shift to liberalism (1875–97). García Moreno's death, as he himself might have forecast, brought a period of near anarchy. Conservatives and Liberals struggled for power. But Ecuador had become part of the world market; the importance of the coast slowly increased, and the Liberals of that area more and more dominated the economy.

A new Liberal hero emerged from the lower classes as

the leader of the coastal reaction to sierra conservatism and clericalism. A man of great personal magnetism, Gen. Eloy Alfaro led a march against the sierra in 1895 and after a year became constitutional president. Eloy Alfaro served two terms as president (1897–1901 and 1906–11). Much of the administrative structure of the García Moreno era was dismantled. The anticlerical Liberals gradually removed the church from state education; they instituted civil marriage and burial, proclaimed freedom of religion, permitted divorce, and eased controls on the press. The church's tithe was abolished, and many of its large estates were confiscated by the state (some estates passing into the hands of Liberal leaders).

The leadership of Eloy Alfaro

In many ways, however, in spite of political manifestos to the contrary, the Liberals of this era shared the basic ideas of the previous period. They advanced García Moreno's road- and railroad-building programs; the Quito-Guayaquil railroad was completed in 1908 during Eloy Alfaro's second term. Moreover, central government did not lose its authoritarian caste; Alfaro, the Liberal strong man (caudillo), was as abitrary and ruthless as his Conservative predecessor. In the sierra and on the coast, power remained in the same hands. The problem of the great haciendas was not touched, and the change to liberalism presumably meant little to the great masses of impoverished Indians and peasants.

Eloy Alfaro's overthrow, like that of García Moreno, was brought about by his stubborn attempts to perpetuate himself in office. A coalition of Conservatives and dissident Liberals forced him and his clique from the presidency in August 1911, but when the next president died in office shortly thereafter, the aging and increasingly unpopular Alfaro returned from exile and tried to recapture his following. The leaders of the Liberals rejected him, and after some fighting he was arrested in Guayaquil. He and his lieutenants were sent to a model prison in Quito, built years before by García Moreno. There, on Jan. 28, 1912, a lynch mob broke in, dragged the prisoners through the streets, and burned the bodies.

Problems of the early 20th century. The Liberals remained in office, although the various factions quarrelled constantly. The real power in the nation continued to be the wealthy merchants and bankers of Guayaquil. During World War I and the short boom that followed it, this powerful clique further extended its influence and diversified its capital with a view to owning the agriculture of the coastal plain. Cacao was the dominant export crop, as in the colonial period, but sugar and rice became increasingly important.

A depression followed in the early 1920s. The price of food increased and exports in general declined. The sucre —the national unit of currency—fell rapidly in value. At the same time, the nation's cacao plantations became infected with a blight known as "witchbroom" disease, and production sagged. These crises brought urban discontent, the formation of trade unions in Guayaquil, riots, and massacres by the army. Hundreds died during riots and shootings in November 1922.

In 1925 the army stepped into this turbulent situation, claiming that it wished to restore national unity and blaming many of the country's problems on the merchant bankers of Guayaquil. Like most Latin American revolutions the one of 1925 brought little change in social and economic structures, but it did replace one governing clique with another.

Recent history (since c. 1925). The period between 1925 and 1948 was one of greater turbulence than Ecuador had ever known. Increasing involvement in the world market and in international politics meant that the nation could no longer escape entanglements and the consequences of world ideological conflicts. Yet during this crucial period, when the tempo of events was accelerating, Ecuador's internal disunity prevented the modernization of its social structure, land tenure system, education, and communications. Thus, the country was badly equipped to face the demands of the age.

Economic development and loss of territory in the 1940s. Ecuador was still suffering from the effects of the great depression when it became involved in World War II. It sided with the allies and allowed the United States to build military bases on its territory, but it played little direct part in the war. Under Pres. Carlos Arroyo del Río, Ecuador drew some benefit from the higher prices for raw materials caused by the war. Nazi infiltration never reached the proportions it did in other parts of South America, and the early years of the war were ones of relative prosperity and tranquility.

World War II

World War II had a serious secondary effect on the nation, however. Because of lack of capital and people, Ecuador by 1940 had not effectively settled its vast Amazonian territories. In July 1941, after long diplomatic bickering and a series of border incidents, the Peruvian army invaded, siezed much of the disputed Amazonian area, and devastated El Oro province (Ecuador had lost territory to each of its more powerful neighbours during its troubled history). The Ecuadorian forces, poorly trained and equipped, were easily defeated, and the national disgrace caused the overthrow of Carlos Arroyo del Río. The United States and the other major powers were too preoccupied with the war to allow such small conflicts to destroy allied unity or to disrupt the production of vital raw materials. A peace conference was hastily called in Rio de Janeiro in 1942, and Ecuador was forced to relinquish title to much of the Amazonian region to which it had claims. Subsequently, Ecuador made repeated attempts to reopen the question, claiming that the "Protocol of Rio" was imposed by force and that the new borders were therefore invalid. This constant Ecuadorian irredentism, although having a strong legal and moral base, nevertheless, caused the country much harm, because it was used repeatedly by demagogues and ultranationalists to distract attention and effort from urgent internal problems.

Domination of Velasco Ibarra (post-World War II). Politics and government after World War II presented a paradox. Except for one interlude of army government (1963–66), Ecuador enjoyed constitutional government and relatively free elections following the election to the presidency of Liberal Party leader Galo Plaza (1948). Yet the period was dominated by the presence of one of the last of Latin America's great caudillos—José María Velasco Ibarra.

By 1970 Velasco had been president of Ecuador five times but had completed only one of these terms. He seemed able to win any election, such was his popularity with the masses; but his terms of office were marked by sudden switches in policy, contradictory economic programs, wild personal outbursts, temporary suspensions of civil liberties, military interventions, and public strife. Some critics claimed that Velasco drew clandestine support from Communist groups; others saw him as little more than the puppet of powerful business groups in Guayaquil; but neither these nor other groups were able to control the erratic Velasco for long.

His political presence may have inhibited the development of coherent political parties and programs for changing the nation's antiquated social and economic structure. His appeal, personal and charismatic, cut across parties and ideologies. The traditional parties—the Liberals and the Conservatives—were thrown into disarray by Velasco's incursions, and the growth of newer parties such as the Socialists and the Social Christians was retarded. Opponents alleged that Velasco made economic progress impossible because constructive measures undertaken by the governments of such presidents as Galo Plaza (1948–52) and Camilo Ponce Enríquez (1956–60) were halted or even reversed by the chaotic Velasco interludes.

Ecuador in the 1970s. Of the many problems faced by Ecuadorians during the 1970s, the most important was how to achieve rapid and just economic development. Most of the population was still rural, although the cities were growing rapidly. One of the highest birthrates in the world ate into per capita gains in productivity and income. Illiteracy, malnutrition, and infant mortality remained huge problems. Many of the people, especially in the sierra, clung to the remnants of their pre-Columbian cultures and played no real part in the national econ-

omy either as consumers or as producers; and because they exerted little political pressure, their incorporation into the nation was talked about constantly but never achieved. Thus, the gap between rich and poor remained vast. The middle class was small and confined to the cities. Large haciendas still dominated the land tenure pattern in the Andean countryside. These estates—undercapitalized and largely unmechanized—were owned by Spanish-speaking white or mestizo city dwellers, while much of the work was performed by Indian sharecroppers or hired labour. Large land reform projects remained unimplemented.

The Ecuadorian economy in the 1970s still suffered from a chronic lack of developmental capital, an unskilled labour force, and, in spite of the opening of many new roads, poor communications. Per capita income was about $200 per year. To worsen the economic picture, the export economy was caught in a monocultural squeeze; the cacao industry of the 1920s had been replaced by banana plantations, but beginning in the early 1960s the country's banana export industry was challenged by a stagnating world market, combined with new competition from Colombia and the Caribbean. The predominance of one export crop also placed the welfare of the nation at the mercy of the world price of bananas. The government had tried to diversify industry to lessen its dependence on bananas, but successes were few. The most promising developments were large oil strikes in the 1960s near the Putumayo River along the Colombian border (see further ECUADOR: *The Economy*).

BIBLIOGRAPHY. G.I. BLANKSTEN, *Ecuador: Constitutions and Caudillos* (1951), a good discussion of Ecuadorian politics in the 1940s and an analysis of the role of José María Velasco Ibarra; L. LINKE, *Ecuador: Country of Contrasts*, 3rd ed. (1960), the best survey to date; M.C. NEEDLER, *Anatomy of a Coup d'Etat: Ecuador 1963* (1964), an account of the role of the military in Ecuadorian politics and of the breakdown of constitutional government in 1963; A. PAREJA DIEZCANSECO, *Historia del Ecuador*, 2nd ed., 2 vol. (1958), a good standard history; J. COLLIER, JR., and A. BUITRON, *The Awakening Valley* (1949), a beautifully illustrated study of the Indians of Otavalo; J.J. PARSONS, "Bananas in Ecuador: A New Chapter in the History of Tropical Agriculture," *Economic Geography*, 33:201–16 (1957); J.V.D. SAUNDERS, *The People of Ecuador: A Demographic Analysis* (1961); N.E. WHITTEN, JR., *Class, Kinship and Power in an Ecuadorian Town: The Negroes of San Lorenzo* (1965), a careful analysis of the social structure in a town near the Colombian border.

(M.J.MacL.)

Ecumenism

Ecumenism, a term of recent origin, was first given wide currency in connection with the Roman Catholic council known as the second Vatican Council (1962–65) and denotes in an admittedly vague manner the awakened conscience of the Christian Church towards its universal aspects as well as towards its rekindled sense of mission and service. Being etymologically a derivative of the Greek words *oikos* ("household") or *oikoumenē* ("inhabited earth") and historically of the ecumenical councils of the distant past and of the ecumenical movement of modern times, ecumenism represents a movement towards church unity that its advocates claim deserves to be studied, heeded, and encouraged.

NATURE AND SIGNIFICANCE

The quest for unity. The nature and the significance of the ecumenical movement in the 20th century lies in the longing of Christians for a greater and more effective role in the affairs of the modern world. Granted that unity and mission belong to the essence or nature of the church—as Christians for centuries have proclaimed—many modern Christians in increasing numbers now assert that they can no longer justify the duality of what they consider to be a beautiful faith on the one hand and an ugly and contradictory life—*i.e.*, a divided church—on the other. Church unity—or the quest for it—is considered by modern ecumenists to be the echo of the spirit calling for the reacceptance by Christians of the One Church. Participants involved in modern theological and ecumenical deliberations often speak out for the need to chart a converging course, which, through mergers and regroupings of churches and denominations, they hope might lead ultimately to unity.

The urge for unity is believed in part to be the result of the pent-up frustrations of Christians who view what they take to be the impotence and the failure of the church to exert an influential impact on social ethics and to recapture the heart and mind of the unbelieving man. The ecumenical dialogue may then be seen as the attempt of the church to re-engage modern believers in a search for that which constitutes the soul or essence of Christianity. Participants in this dialogue place their hope in the revival of the ecumenical Christian community, which they believe once stood and ideally would stand again as the fulfillment of the promised Kingdom of God, and in which a conscientious belief in God would be attendant with peace and good will.

The church as the ecumenical community. The definition of the church by St. Paul in his 1st-century-AD letters as *to sōma Christou* ("the Body of Christ") dominates modern ecumenical thinking, and his premise that the diversity of spiritual gifts or of ministries does not permit a schism in the body seems to force denominational doctrinal positions or confessional theology (based on statements of belief by particular church bodies; *e.g.*, The Augsburg Confession of the Lutheran churches) into a reappraisal of the theology held by the early church, which is viewed as having been undivided. On the basis of this approach the ecumenical movement may be credited with encouraging the churches towards an ecumenical theology even in respect to a recommended study of all "theologies"—*i.e.*, the doctrinal views of the various churches. The ecumenical movement, thus, can be viewed as being of parallel significance with modern theological studies and movements. The healing of the deep wounds of schism inflicted by persons or segments of ancient and medieval Christianity that exhibited divisive qualities may be realized, according to ecumenical enthusiasts, through a revitalized ecumenical theology or the restudy and re-examination—in the light of what is believed to be historical truth—of the true nature and significance of ancient theological and ecclesiological precepts and doctrines.

Diversity in the church

Schisms and heresies, brought about either by disciplinary, moralistic, or philosophical actions leading towards independence of particular segments of ancient Christianity, were seen as disruptive of the unity, rather than of the theology, of the church. Ecumenical adherents believe that through a united or ecumenical theology the unity of the church will be restored. They hold that the unity of the church cannot be supported by sophisms; it can survive only by direct and undisputed theological affirmation.

History counts innumerable conciliar efforts undertaken at one time or another by the ancient church, and it records many ecumenical or provincial councils called at intervals to deal with schismatic or heretical actions or with ecclesiastical "coup d'etat" situations (*i.e.*, illegal or uncustomary assumptions of ecclesiastical offices). Modern chroniclers, however, would find it difficult to count even one collective or partly collective effort of the church to heal or confront any divisive action and situation in the life of modern Christianity. Thus, the ecumenical movement is believed by its advocates to be the only agency in which unity-minded Christians may hope that a truly representative ecumenical council of the church universal may be convened in order to reorder Christian priorities, that such a council or synod will be the first and the last that will accomplish the recomposition and the resurrection of the One Church, and that it will claim all Christians—irrespective of national, racial, or cultural origins and backgrounds, though respecting fully their local disciplines and liturgical traditions.

HISTORY

Ecumenism's early origins. Christian ecumenism can be traced to various roots: to the command of Christ to his Apostles (*i.e.*, to preach the gospel to all nations and

Ecumenism in the New Testament

to baptize persons in the name of the Father, Son, and Holy Spirit, as it is described in all three Synoptic Gospels [Matthew, Mark, and Luke]); to the miracle of the Pentecost (Acts) in which the disciples received the gift of the Holy Spirit; to the founding of the primitive church (Acts); to the Apostolic Synod, also called the Council of Jerusalem (Acts); and to the first ecumenical decree to send forth missionaries (Acts).

During the 1st and 2nd centuries the church was composed of many local and area units with a wide variety of organizational forms: monarchical episcopates (jurisdictional units presided over by a single bishop with sole authority in ecclesiastical matters); colleges of presbyters (ecclesiastical units presided over by several priests with shared authority, with one of their number serving as president of the group); local units that were somewhat democratically organized; and other forms. During the 3rd century the organizational form of the church moved more and more in the direction of the episcopate. In addition to different organizational forms, there were variations in theological interpretation. Four Gospels (Matthew, Mark, Luke, and John) were utilized in varying degrees of importance in the various churches, according to certain theological viewpoints: Mark, with its practical bent, by the Roman Church; Matthew, with its Jewish orientation, by the church in Palestine; John, with its philosophical orientation, by the Greek Church; and Luke, with its universal appeal, by churches throughout the Greek and Syrian East and the Latin West. Despite the lack of uniformity in theological and organizational matters, the churches accepted the view that they were One Church in their faith in Christ, and the church was

Unity in diversity

viewed as having unity in its diversity, even in the face of the intense nationalistic desires within the Roman Empire that affected believers and nonbelievers alike. When, in the edict of Milan (AD 313), the emperor Constantine (d. 337) granted toleration to Christianity within the Roman Empire, a new era for the church began. The goal of unity, however, was marred by theological controversy, and Christianity, which, many had hoped, would be the "cement of the empire," was itself lacking in consensus theologically, and eventually organizationally. In spite of the conflicts, the first ecumenical Council of Nicaea (325) affirmed that the church was one, holy, catholic, and apostolic. This council and the ones that followed were convened in order to define, fortify, and preserve the unity of faith.

Ecumenical contacts between the Eastern and Western churches: 5th to 19th centuries. Ecumenism proper (recognition that the church is not one in organization or mission) entered history with the Byzantine (Eastern Roman) emperor Zeno's *Henotikon* (482), a document aimed at the reunion of the religious and political separatist movements of his time with the religious and political *oikoumenē*, *i.e.*, the Byzantine Church and Empire. During the reign of one of Zeno's successors, Justinian the Great (527–565), ecumenism lost the identity just described and acquired a new meaning: that of religious and political orthodoxy in which the true faith of the church coincides with the true interests of the state. In view of this change in emphasis and the rising influence of the papacy and Western nations, ecumenism—especially after the East–West Schism (beginning in 1054) and the Reformation (16th century)—meant little other than Roman Catholicism in its autocratic and temporal connotations. Because of its position within nations that rose to greater and greater power politically, economically, and culturally, Roman Catholicism held councils that it called ecumenical, though, in reality, they were Western. These councils—especially the fourth Lateran Council (1215), the second Council of Lyons (1274), the Ferrara-Florence Council (1438–45)—though they included some representation from the Eastern churches, indicate that ecumenism was not at that time viewed from the standpoint of the conscience, mission, and service of the church.

Ecumenism within Eastern Christianity, however, has been viewed as a matter of conscience directly related to the survival of the church itself. This is because of Eastern Christianity's long centuries of isolation following the Great Schism (1054) and the subjugation of the East by the Muslims during and after the 7th century. Captive under Muslim political overlords and isolated from Western Christianity, the Church of Constantinople (the traditional centre of Eastern Christianity) attempted to breach the thick walls of its forced isolation. It went as far as Lyons in 1274, Ferrara in 1438, and Florence in 1439 to seek to repair the damages wrought by the Crusades. The Church of Constantinople made approaches to a responsive sister church, the church of Rome, in hopes of restoring unity, but without success. The catastrophic fall of the Byzantine Empire in 1453 obliterated all such hopes in the hearts of the Christians of the East. Though the years of apparent hopeless mourning seemed to have reached an end when the Roman Church sent its Jesuits (monks belonging to the Society of Jesus) into Poland and the Ukraine during the 16th and 17th centuries, the hope was short-lived; it was soon obvious that the Jesuits were not liberators, but conquerors, who, through the program known as the Unia (Union with Rome), induced many Poles and Ukranians to leave the Orthodox Church. Eastern Christians viewed these actions by Roman Catholics as subjugating the feeble spiritual freedom that the Eastern Christians had upheld during more than a century of enslavement to the Muslim Turks (see also ROMAN CATHOLICISM, HISTORY OF; EASTERN RITE CHURCHES).

Ecumenism and the survival of the church

Thus, when the Church of Constantinople received the messenger and the message of the Protestant Reformation leaders in the 16th century—who attempted to reawaken "the dormant conscience of the church" in the hope that their revolt against what they viewed as the abuses of Rome could unite all those Christians who at one time or another were abused by the temporal authority of the pope—it appeared as though the hope for unity might become an actuality. An historic correspondence between Leipzig (Germany) and Constantinople—which began in 1555 between Luther's follower Philipp Melanchthon and Patriarch Joasaph II of Constantinople and reached its apex when the Lutheran Augsburg Confession of 1530 was sent to Joasaph's successor Patriarch Jeremias II by Jacob Andreas, the chancellor of the University of Tübingen—came to an inglorious end with the rejection of the Confession by Jeremias in 1576. An attempt to revive Lutheran–Orthodox relations by the Lutheran theologians Martin Crucius, Jacob Andreas, Lucius Osiander, and Stephen Gerlach, who met in Wittenberg in June 1580 to prepare a reply to Jeremias' thesis, did not fare any better. Jeremias responded in 1581 that it would be useless to continue such a futile correspondence.

Reformation ecumenical relations

Constantinople's ecumenical hopes, nevertheless, remained unaltered and alive. The third Western Church with which the Church of Constantinople came into a series of theological contacts was the Church of England. The Ecumenical Patriarchate (Constantinople) decided to attempt certain semiofficial contacts with the Anglican Church in the 17th century based upon the assumption that the sympathetic attitude displayed by Anglican chaplains in Constantinople, Hunt (1627–38), William Gotobet (1642) and Thomas Smith (1668–70), was an indication of their friendly concern. The first approaches were made by Cyril Lucaris, the alleged Calvinist patriarch of Constantinople, through two youthful ecclesiasts, Metrophanes Critopoulos and Nathaniel Conopius (early 17th century). Joseph Georgarinis, bishop of Samos, who travelled to Britain in 1676 to organize the first Greek Orthodox parish, entered into actual theological discussions with Anglican theologians. His return to Samos in 1685 in no way ended the mutually desired contacts between the two churches. Callinicus II, patriarch of Constantinople during three periods from 1688–1702, replying to an invitation to send students to England, obliged by sending a number of them. All were forced, however, to return, because of what they described as poor food and lodging. Relations were again resumed officially in 1701 when Neophytus, the Bishop of Philippopolis, was invited to be accorded an honorary degree

of doctor of divinity at Oxford. Following the ceremonies, Bishop Neophytus was received by Archbishop Tenison of Canterbury, who in turn introduced him to Dr. John Covel, master of Christ's College at Cambridge, for the purpose of discussing certain doctrines. The deliberations—the apparent aim of which was for the Anglicans to obtain an Orthodox confession of faith that accurately reflected Orthodox beliefs after the persistent confusion over Patriarch Lucaris' "confession," which was Protestant in tone—came to an end when a "synodical" letter sent to the lovers of the Greek Church in England, signed by Patriarch Dionysius IV, was delivered to the Anglican theologians in 1672. Talks on intercommunion (allowing members of both churches to receive the Lord's Supper in either church) that were encouraged by both churches soon ended because of the internal split in the life of the Anglican Church caused by the appearance of the nonjuror movement (a schism concerned with episcopal succession) in the early 18th century (see also ANGLICAN COMMUNION).

Relations between the Orthodox and Old Catholic churches

Another Western Church with which the Church of Constantinople entered into ecumenical relations through correspondence and legation was the Old Catholic Church, which came into existence (in Germany, Switzerland, and Holland) as the result of the decrees of the Roman Catholic Council known as the first Vatican Council (1870). Although the new church had eloquent and forceful exponents, it was lacking both in organization and self-confidence. Along with the Jansenist Church of Holland (a movement that broke with the Church of Rome), in the 18th century, it developed into what is known today in Europe as the Old Catholic Church and in Poland and the U.S. as the Polish National Catholic Church. The letters and the contacts exchanged between Bonn (the headquarters of the German Old Catholic Church) and Utrecht (headquarters of the Jansenist Old Catholic Church) on one part and Constantinople on the other ended with the feeling of ecumenical proximity and hopeful of a "close and intimate friendship" (see also OLD CATHOLIC CHURCHES).

Ecumenical contacts between the Church of Constantinople and the other Orthodox churches. Constantinople's enthusiasm for ecumenism was not always shared by the other Orthodox churches. Its failure to achieve some form of communion or communication between East and West resulted in a new effort or approach on the part of Constantinople—*i.e.*, promoting more frequent meetings and cooperation with the other Orthodox churches, which had no easy means of contact under the new rulers (Muslims) of the former Byzantine Empire. The endeavours of the Ecumenical Patriarchate to bring about Orthodox unity through "Pan-Orthodox synods" proved to be unsuccessful because of continuing political confrontations.

Five Pan-Orthodox synods were convened, and a sixth Pan-Orthodox synod was proposed. None, however, was truly pan-Orthodox, for full representation from Orthodoxy proved to be totally impossible. They were the Synod of Constantinople (1482–84), the Council of Jassy (1642), the Synod of Jerusalem (1672), the Synod of Moscow (1690), the Synod of Constantinople (1772), and the proposed Pro-Synod of Mount Athos (1922)—none of which lived up to expectations. Patriarch Meletios Metaxakis issued invitations to all the Orthodox churches to attend the Pro-Synod of Mount Athos, but because he was forced by the Turkish government to resign it never materialized.

Ecumenical conferences

Ecumenism in the 20th century. Ecumenism in the early part of the 20th century derived much of its impetus from three movements: international missionary conferences (Protestant), beginning with the Edinburgh Conference (1910) and led by a Scottish church leader J.H. Oldham and the International Missionary Council (1921) led by an American missionary leader John R. Mott; the Faith and Order Conferences (dealing with doctrines and practices) beginning with the conference at Lausanne, Switzerland, in 1927 and including both Protestant and Orthodox leaders; and the Life and Work Conferences (dealing with social and other practical problems) beginning with the Stockholm (Sweden) Conference in 1925. These three movements gradually were incorporated into the World Council of Churches (founded in 1948). The second and the third quarters of the 20th century are claimed by many leaders of the ecumenical movement to exemplify ecumenism because of the ecumenical role played by religiously associated organizations (some established in the 19th century) in promoting an evangelistic, nonconfessional understanding of Christianity. Among the more important of these organizations are: the British and Foreign Bible Society, established in England (1804); the Young Men's Christian Association (YMCA) of England, initiated in 1844—which ultimately led to the creation of the World's Alliance of YMCAs; the World Student Christian Federation (1895, Sweden); and the International Missionary Council (founded 1921), which was based upon actions of the International Missionary Conference (convened in Scotland, 1910). Giving new summation to the ecumenical idea that was assuming more and greater importance and influence were ecumenists such as the Swedish Lutheran archbishop, Nathan Söderblom; the Anglican archbishop William Temple and G.K.A. Bell (bishop of Chichester); the Greek Orthodox metropolitan (archbishop) Germanos of Thyateira; the Dutch leader of the World Council of Churches, W.A. Visser t'Hooft; and Hamilcar Alevizatos (professor of canon law, University of Athens), who are mainly responsible for the creation of the World Council of Churches in Amsterdam in 1948.

Modern Orthodox-initiated ecumenical attempts

Of no less significance is the ecumenical spirit that animated the Church of Constantinople in its ecumenical endeavours both before and after the fall of the Byzantine Empire. As noted, that church's many contacts with Lutherans as well as with Anglicans, Calvinists, and Old Catholics attest to the ecumenism of Eastern Christianity. An encyclical letter of Dorotheos (1920) stands out as the most ecumenical document of the early 1920s. The ecumenism that was in the minds and hearts of the Orthodox citizens of Muslim-dominated Byzantium regained much of its former quality and dynamism. As a result, the Church of Constantinople not only associated itself with the World Council of Churches from its inception but it also played the role of protagonist in convincing other Eastern European Orthodox churches to join the council, (as they did at the 3rd General Assembly of the World Council of Churches in New Delhi, November 1961) and in reopening communications with Rome in 1959, when Ecumenical Patriarch Athenagoras I sent his first emissary (Archbishop Iakovos) to Pope John XXIII at the Vatican.

Ecumenism in the 1960s owed much to the second Vatican Council, which, under the leadership of popes John XXIII and Paul VI and through the writings and initiative of the German theologian Augustin Cardinal Bea, promulgated—after thorough discussion and enlightening exchanges of ideas—a noteworthy decree on ecumenism that has helped to make possible many new contacts and visits among high-ranking ecclesiastical leaders, including the encounters between Patriarch Athenagoras and Pope Paul and the unprecedented journeys that they and other religious leaders have made in the cause of the ecumenical movement. The ecumenism of today could well be defined as the centrifugal movement of the conscience of the church universal, aiming at nothing less than the recapturing of the sense of its oneness, holiness, catholicity, and apostolicity, all of which constitute the true essence of Christian ecumenism.

The Pan-Orthodox Synod

The less than successful results of the six abortive Orthodox synods from 1484 to 1922 may indicate to many Orthodox Christians a need for a truly Pan-Orthodox Synod that they hope will be realized before 1980. Such a synod would define more clearly that which appears to need redefinition or reaffirmation in matters of church teachings and would also re-examine and redefine discipline and ecumenism. In the view of many Orthodox leaders, if the ecumenical movement is destined to serve the cause of unity, the Orthodox churches need to present themselves as one, not only in faith but, what is more important, in terms of attesting or witnessing to the Ortho-

dox faith through adequate policy and polity in today's ecumenical, ecclesiastical, political, and social affairs.

CURRENT AGENCIES AND EXPONENTS OF ECUMENISM

The World Council of Churches. Though by its nature it is consultative and conciliar, the World Council of Churches (since its inception in 1948) serves as the chief agent and exponent of ecumenism. By absorbing and enhancing the goals and activities of its forerunners—the Faith and Order Conference in 1948, the Life and Work Conference in 1948, and the International Missionary Council in 1961—the World Council of Churches has moved to become a theological catalyst and, subsequently, the Protestant and Orthodox organ of modern ecumenism. Its ecumenical activities include theological studies as well as commissions on world evangelism; through interfaith colloquiums, seminars, and meetings, directly or indirectly sponsored by the council; it promotes ecumenical consciousness and universal Christian one-mindedness.

Because of Visser t'Hooft, who developed and advanced the global dimensions of ecumenicism and broadened its role from that of a cooperative body to a potent ecumenical force, the World Council of Churches is no longer a loose confederation of churches as it was initially envisioned but is an active and dynamic expression of their concerns and of their anguish. The range of the council's interests begins with ecumenical praying and culminates in church mergers and unions (many already accomplished) that the council silently but effectively guided and inspired.

Protestant and Orthodox agencies in the U.S. The National Council of Churches of Christ in America (1950) is another ecumenical expression of Christian concern. Though it owes its creation to the cooperative initiative of 12 original interdenominational agencies, the council now comprises the large majority of the Protestant and Orthodox churches in the U.S. Through its multiple divisions, departments, and committees that study and analyze national and international issues, the National Council of Churches serves the ecumenical intent in a broad sense.

The Consultation on Church Union

The Consultation on Church Union, another American church unit, envisages the realization of ecumenicity through multiple national mergers. The movement, which was initiated in 1960, in the 1960s and early 1970s encompassed nine Protestant churches (Episcopal, Methodist, United Church of Christ, United Presbyterians Church in the U.S.A., Disciples of Christ, Presbyterian Church in U.S. [Southern], African Methodist Episcopal, African Methodist Episcopal Zion, and Christian Methodist Episcopal). Its goal and plan is to be united in a "Church of Christ Uniting."

Roman Catholic agencies. The Secretariat for the Promotion of Christian Unity is the Roman Catholic equivalent of a faith and order and of life and work commission. Conceived by Pope John XXIII in 1959 and organized by Augustin Cardinal Bea in 1961, it is the guiding ecumenical agency of the Roman Catholic Church. Its major significance lies in the fact that the second Vatican Council—contrary to the first Vatican Council—recognized for the first time in Roman Catholic history the existence not only of separated brethren but also of separated churches, of which it speaks with reverence and solicitude. The Secretariat, furthermore, accepts the theory of unity in diversity and insists not upon the return of the estranged but on the need of a dialogue that would eventually lead Christians and their churches to unity, not with Rome but with one another in Christ with Christian love (agape) and with understanding. The Secretariat seems to understand "unity" as a unity that binds Christians in the fundamentals of the faith but allows them liberty in what may be called the nonessentials.

Orthodox agencies. The Pan-Orthodox Conference is the ecumenical agency of the Orthodox churches acting collectively. Organized in 1961 through the initiative of Patriarch Athenagoras I, it adopted a lengthy and all-important agenda, starting with the definition of dogma and concluding with the detailed list of theological, sacramental, and socioreligious subjects to be discussed with Roman Catholics, Anglicans, Old Catholics, and Lutherans. In its three meetings the Pan-Orthodox Conference designated a theological commission as well as dialogue committees that would elaborate on the ecclesiological problems between Eastern and Western Christianity. Another committee is working on the Great Pan-Orthodox Synod, tentatively scheduled for 1973, which it is expected will set forth in precise terms the Orthodox standards, presuppositions, and positions for the eventuality of reunion with any Christian Church of the West or with any of the ancient Oriental Monophysite churches, *i.e.*, those that broke with Constantinople over the doctrine of the nature of Christ.

The Standing Conference of the Canonical Orthodox Bishops in the Americas (SCCOBA) may also be listed as an ecumenical agency for American Orthodoxy. Organized in 1960 through the initiative of Archbishop Iakovos, the exarch of the Ecumenical Patriarchate of Constantinople, which has jurisdiction over the Greek Orthodox Archdiocese of North and South America, the conference has succeeded through its Ecumenical Commission in representing a united Orthodox view in ecumenical meetings and in embarking upon important conferences and clarifying bilateral concerns with the Anglican, Roman Catholic, Lutheran, and Reformed churches in America.

THE NEW ECUMENISM

Ecumenism of underground churches

A new ecumenism that is in opposition to some of the traditional ecumenical concepts underlies the thinking of the underground churches that are now found especially among the younger generations of the established Christian churches and, most interestingly, in the indigenously emerging Christianity of the emancipated African and Asian nations. New ecumenical currents and countercurrents combine in this movement and have become verbalized in strong words of indignation against what its proponents consider to be an institutionalized and fossilized church—which, it is believed, needs to be liberalized and humanized. Various writers associated with this new ecumenism imply that the church, while it should rediscover its spiritual power, should divest itself of any and all glory and human weaknesses and prove itself to be what it should always be: the Holy Spirit in action. Such writers also speak of a secular ecumenism founded on the doctrine of recreation. In view of such comments a special document, prepared for and presented to the July 1970 meeting of the Lutheran World Federation in Evians-Les-Bains, France, introduced a tripartite ecumenism (pastoral, practical, and social) as an answer to the predicament of a stalled ecumenism.

The new ecumenism, though secular and unconventional in nature, is considered by its devotees, and by those who find merit in it, "as not peripheral to theological unity." If, however, it sanctions violence on behalf of love for the oppressed, it may become a divisive rather than a unifying force, which would impede the progress of the ecumenical movement. Such a new ecumenism, according to the estimate of those associated with what may now be called traditional ecumenism, might meet the ecumenical standards of such things as the Socialist-oriented Prague Peace Conference or the demands of militant and radical Christians but not the requirements of those Christians who continue to believe in the pre-eminence of faith, hope, and love expressed in I Corinthians that sums up the essence and mission of a united Christianity.

BIBLIOGRAPHY. WALTER M. ABBOTT (ed.), *The Documents of Vatican II* (1966), is a documentary account of the historic actions decided upon by Vatican II. GEORGE F. LEWIS, *Toward Anglican-Roman Catholic Unity* (1962), points to the confessional as both encouraging and discouraging Anglicans as well as Roman Catholics in their understanding of ecumenism. IAN HENDERSON, *Power Without Glory: A Study in Ecumenical Politics* (1967), is a good analysis of the many-sided (Protestant-Orthodox-Catholic) post-Vatican II ecumenism, focussing its attention on spiritual power as the principal factor that should underlie the pursuit of Church union. ARTHUR MICHAEL

RAMSEY, *God, Christ and the World* (1969), places emphasis on theology as the only way towards reuniting Christianity. GEOFFREY FRANCIS FISHER, *The Archbishop Speaks* (1958), reflects an understanding of ecumenism as an invitation to an objective re-evaluation of the theology of the particular Christian churches. N. GOODALL in his *The Ecumenical Movement* (1961), records the personal feelings and thoughts of a man committed to world evangelism as a means of achieving true ecumenism. JOHN COLEMAN BENNETT, *Foreign Policy in Christian Perspective* (1966), is an analysis of the Christian perspective of foreign policy by a theologian committed to religious and political or international ecumenism. S.C. NEILL, *Towards Church Union, 1937–1952* (1952), is a record and chart of theological mores affecting, and the steps that should be taken in gaining, a clearer perspective of the theological dimensions of modern ecumenism. DENO JOHN GEANAKOPLOS, *Emperor Michael Palaeologus and the West, 1258–1282* (1959), studies in depth the political overtones and motivations of the reunions negotiated and agreed upon in Lyon and Florence-Ferrara. R. ROUSE and S.C. NEILL, (eds.), *A History of the Ecumenical Movement 1517–1948* (1954), stands out as the most scholarly work in the series of ecumenical literature to be consulted in understanding both the division and the imperative of bringing about a united Christendom. ERNST BENZ, *Geist und Leben der Ostkirche* (1957; Eng. trans., *The Eastern Orthodox Church*, 1963), gives an objective analysis of the openness of the Eastern Church, vis-a-vis the ecumenical orientation of modern times. STEVEN RUNCIMAN, *The Great Church in Captivity* (1968), reveals the abandonment of the Eastern center of Christianity, resulting from the westernization and alleged self-righteousness of Christianity, pointing to the need for a new evaluation of unity, both as a demand of God and of history.

(I.)

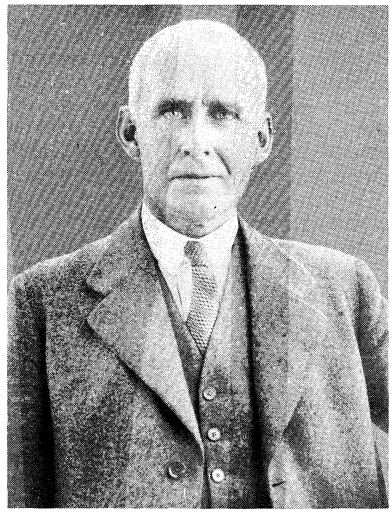
Eddington, Sir Arthur Stanley

Astronomer, physicist, and mathematician, Sir Arthur Stanley Eddington was a pioneer in the fields of relativity, cosmology, and the dynamics and internal constitution of the stars. He was also a notable expositor of new scientific ideas and their philosophic significance.

He was born on December 28, 1882, the son of the headmaster of Stramongate School, an old Quaker foundation in Kendal near Lake Windermere in the northwest of England. His father was a gifted and highly educated man who died of typhoid in 1884. The widow took her daughter and small son to Weston-super-Mare in Somerset, where young Eddington grew up and received his schooling. He entered Owens College, Manchester, in October 1898, and Trinity College, Cambridge, in October 1902. There he won every mathematical honour, as well as Senior Wrangler (1904), Smith's prize, and a Trinity College fellowship (1907). In 1913 he received the Plumian Professorship of Astronomy at Cambridge and the following year became also the director of its observatory.

From 1906 to 1913 Eddington was chief assistant at the Royal Observatory at Greenwich, where he gained practical experience in the use of astronomical instruments. He made observations on the island of Malta to establish its longitude, led an eclipse expedition to Brazil, and investigated the distribution and motions of the stars. He broke new ground with a paper on the dynamics of a globular stellar system. In *Stellar Movements and the Structure of the Universe* (1914) he summarized his mathematically elegant investigations, putting forward the thesis that the spiral nebulae, cloudy structures seen in the telescope, were galaxies like the Milky Way.

By courtesy of the University of Chicago: photograph, Yerkes Observatory, Williams Bay, Wis.



Eddington.

During World War I he declared himself a pacifist. This arose out of his strongly held Quaker beliefs. His religious faith also found expression in his popular writings on the philosophy of science. In *Science and the Unseen World* (1929) he declared that the world's meaning could not be discovered from science but must be sought through apprehension of spiritual reality. He expressed this belief in other philosophical books: *The Nature of the Physical World* (1928), *New Pathways of Science* (1935), and *The Philosophy of Physical Science* (1939).

During these years he carried on important studies in astrophysics and relativity, in addition to teaching and lecturing. In 1919 he led an expedition to Príncipe Island (West Africa) that provided the first confirmation of Einstein's theory that gravity will bend the path of light when it passes near a massive star. During the total eclipse of the sun, it was found that the positions of stars seen just beyond the eclipsed solar disk were, as the general theory of relativity had predicted, slightly displaced away from the centre of the solar disk. Eddington was the first expositor of relativity in the English language. His *Report on the Relativity Theory of Gravitation* (1918), written for the Physical Society, followed by *Space, Time and Gravitation* (1920) and his great treatise *The Mathematical Theory of Relativity* (1923)—the latter considered by Einstein the finest presentation of the subject in any language—made Eddington a leader in the field of relativity physics. His own contribution was chiefly a brilliant modification of affine (non-Euclidean) geometry, leading to a geometry of the cosmos. Later, when the Belgian astronomer Georges Lemaître produced the hypothesis of the expanding universe, Eddington pursued the subject in his own researches; these were placed before the general reader in his little book *The Expanding Universe* (1933). Another book, *Relativity Theory of Protons and Electrons* (1936), dealt with quantum theory. He gave many popular lectures on relativity, leading the English physicist Sir Joseph John Thomson to remark that Eddington had persuaded multitudes of people that they understood what relativity meant.

Relativity theory

His philosophical ideas led him to believe that through a unification of quantum theory and general relativity it would be possible to calculate the values of universal constants—in particular involving the number 137 (the fine-structure constant), the ratio of the mass of the proton to that of the electron, and the number of atoms in the universe. This was an attempt, never completed, at a vast synthesis of the known facts of the physical universe; it was published posthumously as *Fundamental Theory* (1946), edited by Sir Edmund Taylor Whittaker, a book that is incomprehensible to most readers and perplexing in many places to all but which represents a continuing challenge to some.

Eddington received many honours, including honorary degrees from 12 universities. He was president of the Royal Astronomical Society (1921–23), the Physical Society (1930–32), the Mathematical Association (1932), and the International Astronomical Union (1938–44). He was knighted in 1930 and received the Order of Merit in 1938. Meetings of the Royal Astronomical Society were often enlivened by dramatic clashes between Eddington and Sir James Hopwood Jeans or Edward Arthur Milne over the validity of scientific assumptions and mathematical procedures. Eddington was an enthusiastic participant in most forms of athletics, confining himself in later years to cycling, swimming, and golf. He died on November 22, 1944.

Contributions to astrophysics

Eddington's greatest contributions were in the field of astrophysics, where he did pioneer work on stellar structure and radiation pressure, subatomic sources of stellar energy, stellar diameters, the dynamics of pulsating stars, the relation between stellar mass and luminosity, white dwarf stars, diffuse matter in interstellar space, and so-called forbidden spectral lines. His work in astrophysics is represented by the classic *Internal Constitution of the Stars* (1925) and in the public lectures published as *Stars and Atoms* (1927). In his well-written popular books he also set forth his scientific epistemology, which he called "selective subjectivism" and "structuralism"—*i.e.*, the interplay of physical observations and geometry. He believed that a great part of physics simply reflected the interpretation that the scientist imposes on his data. The better part of his philosophy, however, was not his metaphysics but his "structure" logic. His theoretical work in physics had a stimulating effect on the thought and research of others, and many lines of scientific investigation were opened as a result of his work.

BIBLIOGRAPHY. A biography by A. VIBERT DOUGLAS, *The Life of Arthur Stanley Eddington* (1956), contains a list of his books and scientific papers. A popular discussion of Eddington's work is contained in J.R. NEWMAN, *Science and Sensibility*, vol. 1 (1961). His place in the history of physics is described in EDMUND T. WHITTAKER, *From Euclid to Eddington: A Study of Conceptions of the External World* (1949). For his philosophical ideas, see HERBERT DINGLE, *The Sources of Eddington's Philosophy* (1954). Technical discussions of his work are C.W. KILMISTER, *Men of Physics: Sir Arthur Eddington* (1966); and NOEL B. SLATER, *The Development and Meaning of Eddington's "Fundamental Theory"* (1957).

(A.V.D.)

Edentata

Edentata, an order of mammals, includes 31 living species distributed among the armadillos, true anteaters, and tree sloths, as well as eight extinct families of ground sloths and armadillo-like animals. The living families and six of the extinct families comprise the suborder Xenarthra. A second suborder, Palaeanodonta, consists of two extinct families. The entire evolutionary history of the edentates is restricted to the Western Hemisphere and the majority of the living species occur today in South America.

GENERAL FEATURES

Edentata means lacking teeth, though in reality only the true anteaters are toothless. The majority of edentates have simple, peglike cheek teeth that lack enamel; canine-like teeth do occur in some forms. Certain armadillos may have as many as 100 teeth. Edentates possess specialized traits, such as reduced dentition, a long sticky tongue, powerful, clawed, forefeet, associated with their insect diets. They also have primitive traits, such as the possession of five toes on the hindfeet, a simple uterus, and small, uncomplicated brain, which place them close to the primitive stock that gave rise to the infraclass Eutheria, which includes all placental mammals.

A persistent myth holds that any kind of armadillo, in defense against enemies, can roll itself into an impregnable armour-plated ball. In fact, only the three-banded armadillo (*Tolypeutes tricinctus*) is able to perform this feat. Most armadillos contort themselves into a reasonable facsimile of a ball, but some only slightly increase the convexity of the dorsal shell (carapace) while flattening against the ground. Individuals of some species flee to burrows or attempt to burrow right on the spot but may freeze and feign death when flight and rapid excavation are somehow prevented. Misconceptions about other edentates also exist. The toothless, weak-jawed anteaters seem to be sucking in ants and termites through a small tube, when, in fact, it is the sticky product of the greatly enlarged submaxillary glands covering the lengthy extensible tongues that facilitates the capture of these insects. The slow-moving arboreal sloths, typically depicted as moving upside down along a branch, really spend only about 10 percent of their time moving at all.

The armadillos and anteaters are important ecologically, helping to control many kinds of abundant insects, especially ants and termites. Armadillos feed on some carrion and thus play a small part in cleaning up the environment. Although some armadillos may become agricultural pests, and the armour of several species is made into purses and other curios, edentates are generally of little economic importance. Generally, the greatest interest in edentates lies in their bizarre appearance.

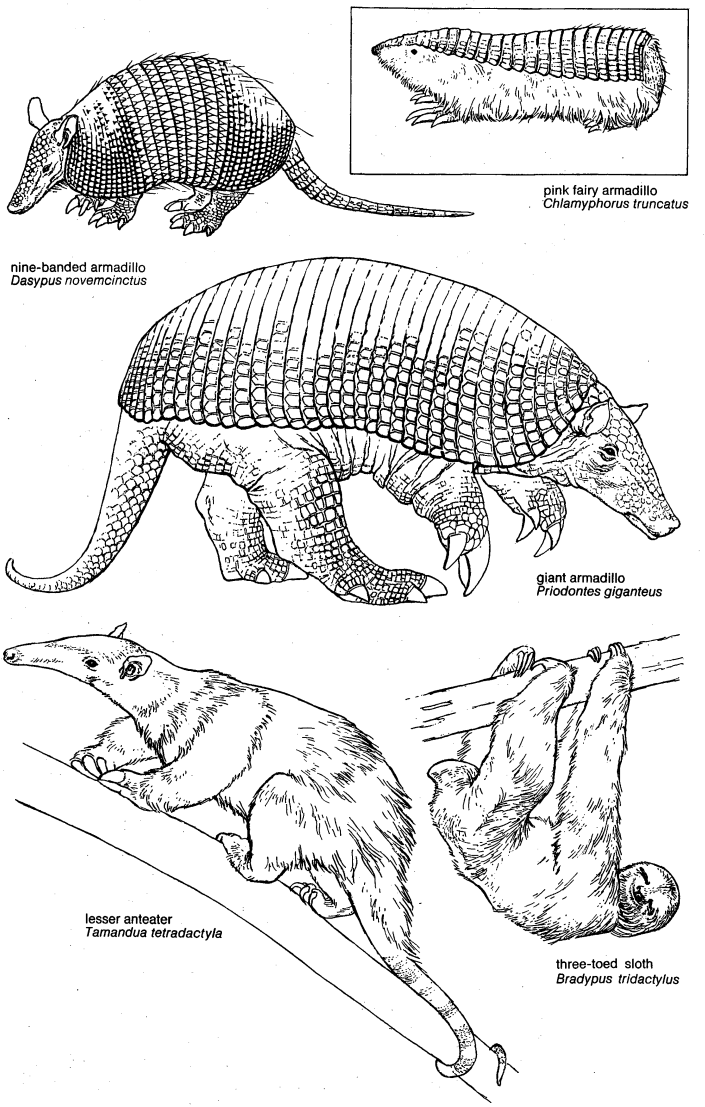


Figure 1: Edentate body plans.
Drawing by J.C. Barberis

Living edentates range in size from the tiny pink fairy armadillo or "lesser pichiciego" (*Chlamyphorus truncatus*) of Argentina, measuring about 16 centimetres (6.5 inches) in length and weighing little more than 100 grams (a few ounces), and the slightly larger two-toed or silky anteater (*Cyclopes didactylus*), just over 37 centimetres (15 inches) long and weighing about 325 grams (12 ounces), to the 60-kilogram (125-pound) giant armadillo (*Priodontes giganteus*), nearly 1.5 metres (5 feet) long, and the giant anteater (*Myrmecophaga tridactyla*), which weighs up to 25 kilograms (55 pounds) and may be over 2 metres (6 feet) in length. Among extinct forms, an Oligocene armadillo, *Prozaedyus proximus*, had a skull under 8 centimetres (about 3 inches) in length and was only somewhat larger than the smallest of the living armadillos. The Pleistocene ground sloth, *Megatherium americanum*, was 6 metres (20 feet) long and was larger than a modern elephant; a Pleistocene glyptodon (*Doedicurus clavicaudatus*) was more than 4 metres (13 feet) long and 1.5 metres (5 feet) high. Both flourished in South America about 500,000 years ago.

Anteaters, tree sloths, and armadillos are strikingly dif-

ferent in appearance, but within each of these families, there is only a modest diversity of structure. The toothless anteaters vary from strictly arboreal to terrestrial forms that are characterized by progressively more elongate tubular mouths. The arboreal species have only two (*Cyclopes*) or three (*Tamandua*) clawed toes on the forefeet, whereas the terrestrial giant anteater has four. In addition, the arboreal species have prehensile tails. At a casual glance, the two genera of herbivorous (plant-eating) tree sloths (*Bradypus, Choloepus*) differ from each other in general appearance only in the number of toes on the forefeet (two in *Choloepus* and three in *Bradypus*) and in pelage coloration. The armadillos (Dasypodidae) are distinguishable on the basis of relative size and on the distribution, flexibility, and pattern of the bony shields that form their body armour.

The anteaters, essentially a tropical family, occur from northeastern Argentina and northern Uruguay to the extreme northern part of the state of San Luis Potosí, Mexico. Tree sloths are restricted to forests from northern Rio Grande do Sul, Brazil, to southeastern Honduras. Armadillos are found from the northern part of the province of Santa Cruz, Argentina, to south-central Kansas. Most living edentates are fairly common in their preferred habitats, although several species have disappeared from parts of their range largely because of habitat destruction and excessive hunting by man. These endangered species include the giant anteater, the three-toed sloth (*Bradypus torquatus*), the three-banded armadillo, the giant armadillo, the pink fairy armadillo, and Burmeister's armadillo (*Burmeisteria retusa*).

NATURAL HISTORY

Behaviour. Anteaters eat termites, ants, and occasionally other insects. They may forage in trees (two-toed), on the ground (giant *Myrmecophaga*), or in both situations, as does the lesser anteater (*Tamandua tetradactyla*). Their prey is usually located by smell, the sense of which is better developed in these animals than are those of eyesight and hearing. With the long, powerful foreclaws, the anteaters rip open anthills and termite mounds, into which they then insert their long sticky tongues. The tongue, covered with prey and debris, is then drawn into the tubular mouth. When frightened, the giant anteater usually flees at a slow clumsy gallop, but if forced it will fight, using its claws to rip the enemy. *Cyclopes and Tamandua* also use their foreclaws for defense, anchoring themselves to a branch with their prehensile tails. When alarmed, these two emit shrill calls.

Anteaters travel solitarily or as mother and young and do not seem to have permanent resting places. The giant anteater chooses a secluded spot such as a hollow log or the abandoned burrow of some other animal and then curls up to rest, with its long bushy tail covering its head and body. The arboreal anteaters find shelter in trees. The two-toed and lesser anteaters are mainly nocturnal, but the giant anteater is both nocturnal and diurnal. In areas remote from civilization, the giant anteater is most active during the day, whereas in heavily populated regions it is primarily nocturnal.

Food habits of sloths

Tree sloths are strictly herbivorous, but the absence of enamel on their teeth suggests that the group perhaps was once insectivorous. The lack of enamel could pose serious problems relating to tooth wear for animals eating such abrasive food as vegetation. The difficulty, however, is alleviated by continuous growth of the teeth, which have an outer covering of hard cement. As is typical of herbivores, tree sloths have large multichambered stomachs in which the slowly digesting plant material steeps. Constantly filled, the stomach and its contents constitute about 30 percent of the sloth's body weight. A meal may take up to a week to digest. Every two to eight days, the sloth descends to the foot of its home tree and passes a large amount of excrement.

The diet of the two-toed sloth consists of a variety of leaves, stems, and fruits. The three-toed sloth, on the other hand, eats only the leaves of the cecropia tree. The sloth locates its food by touch and smell and then hooks a branch with its curved claws and pulls it to the mouth. Although colour vision is present, the eyesight and hearing of sloths are not very well developed, and they orient themselves mainly by touch.

In the wild, the hair of the sloth's body is inhabited by algae, which flourish in the jungle humidity. A variety of mites, beetles, and small moths live as commensals (harmless companions) in the hair, feeding on the algae.

Activity. Tree sloths are normally solitary creatures and spend the days and most of the nights sleeping. The little time that remains is spent eating, resting, and moving. This "laziness" contributes greatly to their survival, because their diurnal sleeping posture (hanging from a branch with feet bunched together and head tucked in on the chest) and their coloration give them the appearance of a bunch of dead leaves, rendering them virtually invisible. At night, when they are active, their slow movements do not attract the attention of nocturnal enemies. In addition, sloths have a thick skin and the ability to tolerate strong poisons and severe injuries.

When angered or otherwise disturbed, the three-toed sloth utters a high-pitched cry that sounds like "ai," whereas the two-toed sloth snorts and hisses. Normally, sloths are docile and in the presence of danger remain still, relying on their protective coloration to conceal them. If molested, however, they bite savagely and strike out furiously with the sharp foreclaws.

Burrowing by armadillos

Armadillos are primarily nocturnal, spending the daylight hours in their burrows, in expanded nest chambers lined with grasses and leaves. The burrow is also used for the escape from enemies and as a food trap. When using a burrow, in an emergency, some species are able to flex their bony back plates and brace with the feet in such a manner as to become firmly wedged in the tunnel and virtually impossible to pull out. When in danger away from a burrow, armadillos will draw in their feet and lie so that the edges of their armour touch the ground, or will run, burrow rapidly, or even defend themselves with their claws. The three-banded armadillo can roll itself into a ball with feet inward and the head and tail protected by special shields. If captured, an armadillo usually reacts by "playing dead," either stiffening or relaxing, but in either case remaining perfectly still for a brief period. If this course of action does not result in release, the captive armadillo begins kicking vigorously, often facilitating escape. Several kinds of sounds are reported to have been made by fleeing or otherwise agitated armadillos. The peludos (three species of *Chaetophractus*) make snarling sounds, the mulita (*Dasypus hybridus*) repeatedly utters a guttural monosyllabic sound reminiscent of the rapid fluttering of a human tongue, and other species emit a series of grunts or buzzes when alarmed.

The act of burrowing involves the forelimbs, hindlegs, and tail. The heavy claws on the forefeet are used for scraping and loosening the dirt, pushing it toward the rear. The hindlimbs throw the dirt out behind while the tail elevates the rear end of the animal so that the hindlegs may kick more freely. In this manner, burrows from 0.6 to six metres (two to 20 feet) long, from 0.3 to 1.5 metres (one to five feet) underground and with as many as 12 entrances, may be constructed in a relatively short period. Occasionally, other small animals share these burrows. In North America, rabbits (*Sylvilagus*) and cotton rats (*Sigmodon*) have been found living in the burrows of the nine-banded armadillo.

Armadillos eat insects, mollusks, frogs, small reptiles, carrion, and some plant material and will burrow underneath an animal carcass to obtain maggots and other insects. An acute sense of smell enables some armadillos to detect grubs and other insects up to 12 centimetres below the surface of the earth. The peludo is reputed to make conical holes with its snout in the earth in search of food. It first inserts its snout and then rotates its body so that it literally drills the hole. It is able to hold its breath under such circumstances for up to four minutes. Armadillos have relatively good hearing and vision.

Locomotion. The giant anteater walks with a shuffling gait bearing the weight of its foreparts on the side of the fourth (outermost) digit on the front foot. When hard pressed, this species breaks into a slow, clumsy gallop. It

is also a good swimmer. The tamandua walks rather clumsily on the sides of its forefeet when on the ground, in a manner similar to that of the giant anteater. The tamandua is an agile climber, aided both by its foreclaws and hindclaws and by its prehensile tail. Two-toed anteaters are strictly arboreal and are thus very strong climbers. They also have a prehensile tail.

The claws of tree sloths, which are quite similar to those of their terrestrial forebears, have become adapted to a strictly arboreal life. On the ground, a tree sloth is nearly helpless and cannot move at all unless there is something to grasp, then it is able to drag itself along with its claws. Most of the time sloths simply hang upside down, but when they do move they use a slow hand over hand motion both on horizontal and upright branches. Interestingly enough, they are excellent swimmers.

Generally, armadillos walk on the tips of the claws of their front feet and on the soles of the hindfeet; occasionally, they rear up and walk bipedally on their hindlegs. They trot with a stiff-legged gait. At a gallop, some species can outdistance a man. Because of their heavy dermal armour, armadillos have a specific gravity so high that they would normally sink in water. To overcome this problem, they gulp quantities of air that give them the buoyancy needed for swimming. Under some circumstances, armadillos seemingly take advantage of their specific gravity and cross small streams simply by walking under water on the bottom.

Habitat. Edentates occur in a variety of habitats, in temperate, subtropical, and tropical environments. Giant anteaters are found in savannas and in wet forests, where they feed either at night or in the day, ripping the terrestrial mounds of termites and ants apart with their powerful foreclaws. Lesser and silky anteaters are both arboreal, the latter totally so. Lesser anteaters prefer tropical forests, parklands, and, to a lesser extent, savannas, in which they forage mostly in trees on ants, termites, and various other insects. They are predominantly nocturnal but may be active at twilight as well. Silky anteaters, which are nocturnal, are found in tall humid tropical forests usually high up in the trees, where they feed almost exclusively on termites. The herbivorous tree sloths occur only in dense tropical rain forests. Armadillos frequent grasslands, parklands, open and even thick forests. They may live largely in subterranean burrows (pink fairy armadillo) in grasslands or dig burrows for temporary shelter in forests (*Dasypus* species). The shape of the burrow entrance is often diagnostic of the cross-sectional shape of the owner; the nine-banded armadillo (*Dasypus novemcinctus*), for example, which has a high rounded carapace, digs a nearly circular entrance. The six-banded armadillo (*Euphractus sexcinctus*), which is compressed from top to bottom, digs a burrow with an entrance that is wider than it is high. The habitat preference of the three commonest armadillos in Uruguay illustrates the ecological diversity of this group. *Dasypus hybridus*, the smallest of the species, prefers grassland and most of its burrows are found there. The six-banded armadillo burrows near streams in open woodlands and grasslands where it occasionally overlaps with the former species and with the nine-banded, which most frequently makes its burrows in woodlands.

Reproduction. The gestation period of those edentates for which it is known ranges from 65 days in the hairy armadillo (*Chaetophractus villosus*) to 263 days in the two-toed sloth. The protracted gestation period of the sloth is attributed to delayed implantation. The very early embryo, consisting of but a few cells, lies free in the uterus without developing further for several months, before becoming imbedded in the uterine wall, at which time normal embryonic growth recommences. Anteaters and tree sloths normally bear single young. Most armadillos give birth to one or two young, although some species demonstrate polyembryony (the birth of from 2 to 12 identical embryos, which develop initially from a single fertilized ovum) and delayed implantation (attachment of the ovum to the uterine wall after a variable delay period). The nine-banded armadillo has both of these phenomena. The delayed implantation is seasonal, the stages occurring at fixed times of the year, related to environmental influences. In the Northern Hemisphere, young of this species are born in March and April. Breeding occurs between June and September, but, regardless of when breeding occurs, implantation is delayed until November and the postimplantation pregnancy is between November and March. Delayed implantation is presumably beneficial because birth does not take place until spring, when food is plentiful.

Little information is available on the birth and early life of many edentates. A recent study of the two-toed sloth (*Choloepus didactylus*) revealed that the young is born as the mother hangs upside down. The baby emerges head first and face upward, and, as soon as its forelimbs are free, it grasps the abdominal hair of the mother and pulls itself up on her chest. The mother may assist in birth by pulling on the young. Subsequently, the female breaks the umbilical cord with her teeth.

The young sloth's eyes and ears are open at birth. It suckles for about five months but does not leave the mother until the ninth month, when adult proportions, but not size, have been attained. Adult size is reached at from two and one half to three years. Young anteaters also cling to their mothers for several months to a year after birth. The giant anteater rides on its mother's back, the two-toed anteater on its mother's tail.

Armadillo young are born with soft leathery skins, which harden as they grow older. In the six-banded armadillo, the eyes do not open for several days after birth, but in the nine-banded armadillo, the young are born with their eyes open. Young armadillos are able to walk when only a few hours old. In the nine-banded armadillo, physical maturity is achieved at six months, but females do not have their first litter until they are two years old. In captivity, the giant anteater has lived 14 years, the two-toed sloth 19 years, the peludo 15½ years, and the three-banded armadillo 11 years.

FORM AND FUNCTION

External appearance. Anteaters (Myrmecophagidae) have elongate heads; tubular muzzles with no teeth; and long, sharp claws. Their toes vary in number from two to four on the forefeet and from four to five on the hindfeet, depending on the species. The thick coat of the giant anteater is gray with lateral white-bordered blackish stripes, and the hair is long and straight, especially on the characteristically stiff bushy tail. Tamanduas, or lesser anteaters, have short hair, which is usually cream-coloured to brownish, with a black vestlike area over the upper thorax. The entire underside of the tail, as well as the tip, is naked. The two-toed anteater (*Cyclopes*) has soft silky fur, usually golden yellow in colour. In this species, only the underside of the prehensile tail is naked. Males and females are of approximately equal size in *Cyclopes* and *Tamandua*, but males are much larger in *Myrmecophaga*. All anteaters have enormously developed salivary glands that supply the sticky saliva that covers their long extrusible tongues.

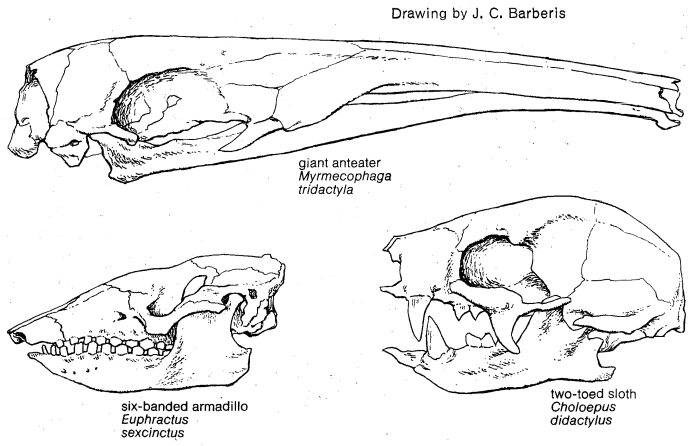


Figure 2: Edentate skulls.

The shaggy pelage of tree sloths grows from the belly toward the back on the body and toward instead of away from the body on the limbs (*i.e.*, downward, when the animal is suspended beneath a branch), thus facilitating the shedding of rain. Tree sloths have pale faces with darker eye rings. The body hair is yellowish to brownish, but the individual hairs are fluted and contain a growth of algae that gives a greenish cast to the entire animal, especially in the rainy season. In the drier conditions of captivity, the algae die and the animal loses the natural coloration. The legs are long, with the forelimbs longer than the hindlimbs, and are designed for suspension of the body rather than for "column-like" support.

Armadillos are characterized by a dorsal bony carapace made up of scapula (shoulder), dorsal, and pelvic shields, separated by a series of movable bands. A bony cephalic (head) shield is also present. The tail is typically armoured as well, except in the naked-tailed armadillos of the genus *Cabassous*. The body hair is greatly reduced, but some, such as the pink fairy armadillo, have dense fur on their sides and underparts. A sparse growth of hair also occurs among the bony plates of all species.

Specializations of internal anatomy. Female edentates have a common urogenital sinus (*i.e.*, both reproductive and excretory functions involve a single chamber) and a simple uterus (a single uterine cavity rather than a branching two-chambered organ). The placenta (the uterine connection between the mother and embryo) is discoid in shape. In other words, the actual areas between which gases, food, and nitrogenous wastes are exchanged are restricted to a single disc-shaped region. The placenta is also deciduous; *i.e.*, it is shed at the time of birth. Females have one or two pairs of mammary glands that are located just lateral to the "armpits," on the chest or on the abdomen. In males, the testes are abdominal and the length of the penis is associated with the degree of difficulty encountered in copulation. In tree sloths, where the partners hang abdomen to abdomen, it is quite small, whereas in armadillos, in which the carapace prevents the partners from getting very close to one another, the penis is long and pendulous.

In general, edentates have poor temperature control. Definitely related to the conservation and manipulation of heat are the extensive anastomoses (fusions) of the veins and arteries (termed retia mirabilia) in the limbs and tails of most species. These structures conserve heat, transferring heat from the arterial components carrying warm blood outward along the limbs to the veins lying adjacent for return to the trunk. At the same time, the appendages are cooled.

Tree sloths are heterothermic (*i.e.*, have imperfect control of body temperature) and under experimental conditions their body temperature, normally ranging between 28° and 35° C, may drop to 20° C. At this temperature, the animals become torpid. Temperature control is likewise relatively poor in armadillos. The fact that it builds a subterranean leaf-lined nest has aided the nine-banded armadillo in the northward expansion of its range into the United States. Nonetheless, many individuals of this species perish during severe winters.

EVOLUTION AND PALEONTOLOGY

The ancestry of the order Edentata is believed to be rooted in some unknown but probably generalized insectivore (order Insectivora). The oldest known members of the two presently recognized suborders of edentates were already too specialized for either group to have given rise to the other. Members of the extinct suborder Palaeanodonta, a group restricted to North America, by the time of their first appearance in the fossil record had already developed specializations that eliminate them as possible direct ancestors of the suborder Xenarthra, which flourished throughout the Tertiary in the then isolated South American landmass.

The palaeanodonts

The palaeanodonts, which lived from the late Paleocene until the beginning of the Oligocene, comprised two families, Metacheiromyidae and Epoicotheriidae. The most completely preserved form in the former family is *Metacheiromys*, an archaic mammal about 45 centimetres (18 inches) in length, reminding one somewhat of an armadillo lacking armour. This animal possessed an assemblage of clearly edentate traits, including the near absence of teeth (except for the enlarged bladelike canines), long compressed foreclaws, and a slight inward twist of the wrist, which is typical of virtually all members of the order. Other metacheiromyids did not show the strong development of the canines. *Palaeanodon*, known from the upper Paleocene and probably ancestral to *Metacheiromys*, had a series of four or five peglike cheek teeth that lacked enamel. The Epoicotheriidae, which had a skull that in outline is reminiscent of that of a small armadillo, were apparently small aberrant forms, highly specialized for a fossorial (digging) livelihood. When the palaeanodonts passed from the scene at the end of the Oligocene, they left no descendants.

The second suborder, Xenarthra, by far the more successful group of edentates, enjoyed an evolutionary history independent from that of the palaeanodonts because North and South America were separated throughout most of the Tertiary, not becoming rejoined until the Pliocene. Contemporaneous with the earliest palaeanodonts of North America was the first South American xenarthran, *Utaetus*, already recognizable as a primitive armadillo and thus sufficiently advanced to eliminate any possibility that known palaeanodonts might be directly ancestral to xenarthrans. The fact that primitive armadillos are the first known members of the suborder strongly suggests that the other xenarthran families are in some way derived from the Dasypodidae.

The Tertiary radiation of edentates in South America resulted in several more or less separate lineages, which are conveniently arranged in three infraorders. The first of these infraorders, Loricata, includes the armadillos (Dasypodidae) and their relatives, the extinct glyptodonts (Glyptodontidae) or so-called turtle armadillos, distinguished from the former especially by their possession of a rigid carapace. Armadillos flourished in South America in the Tertiary and were abundant by the Miocene. In the Pleistocene, some species had moved northward as far as what is today the central United States.

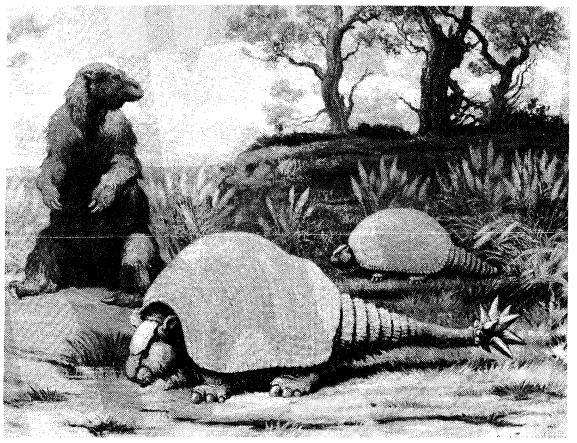
Fossil armadillos

Four distinctive lines developed in the Dasypodidae, comprising the subfamilies Dasypodinae, the Stegotheriinae, the Pampatheriinae, and the Chlamyphorinae (including only the living genera *Chlamyphorus* and *Burmeisteria*). The Dasypodinae includes most of the living genera of armadillos as well as *Utaetus*. The fossil members of this group varied in size from that of a small contemporary armadillo to animals as large as the South American Pliocene form *Macroeuphractus* that had a skull nearly 28 centimetres (11 inches) long. The head and body of the latter must have been almost two metres (6 feet) long. The fossil members of this subgroup all seemed to favour invertebrates in their diets, but they probably ate carrion and some plant material as well. This distinguished them from the extinct Stegotheriinae, a subfamily whose members were highly specialized for a strictly insectivorous diet. *Stegotherium* of the lower Miocene of South America had an elongate anteater-like skull with much reduced cylindrical cheek teeth.

The Pampatheriinae, known from the late Miocene to the Pleistocene, was a group characterized by great size. At least one form, *Pampatherium*, a nearly complete skeleton of which was recovered from a Pleistocene deposit in Texas, was nearly as large as a rhinoceros. This animal and its closest relatives were apparently adapted to an herbivorous diet. Both insect- and plant-eating groups are thus included among the true armadillos.

The genus *Peltephilus* and several other closely related genera of aberrant armadillo-like creatures have been placed in their own family, Peltephilidae, because, among other characteristics, they possessed at least one pair of hornlike protuberances on the muzzle. These animals, which appeared in the Oligocene and disappeared in the Pliocene, were apparently specialized for flesh eating.

The glyptodonts (Glyptodontidae) evolved from primitive armadillos, first appearing in the fossil record in the Eocene with the genera *Glyptatelus* and *Lomaphorelus* but remaining relatively uncommon until the beginning



(Left) Ground sloth (*Megatherium*) and (right) glyptodonts (*Doedicurus*). Painting from fossil skeletons by Charles R. Knight. By courtesy of Field Museum of Natural History, Chicago

of the Miocene. Subsequently, they experienced a radiation of their own and by the Pleistocene several kinds had spread into North America. Throughout their history, they demonstrated an increase in size culminating in the Pleistocene giant, *Doedicurus*. They became extinct before the end of that epoch.

The second infraorder, Pilosa (hairy) includes the sloths, both the living arboreal forms (Bradypodidae) and the fossil ground dwellers of the families Megatheriidae, Megalonychidae, Mylodontidae, and Orophodontidae. There are no fossils of tree sloths, but they appear to be structurally related to the ground sloth families Megatheriidae and Megalonychidae. The oldest fossils of ground sloths belong to the Megatheriidae and come from Oligocene deposits in South America. By the early Miocene, primitive members of this group were rather common. One form, *Hapalops*, although typical of that time, was of relatively small size for a ground sloth. It had a slender body about 1.3 metres (4.5 feet) in length and had five toes with developed claws on both its forefeet and hindfeet. *Hapalops* had a moderately elongated skull with slender premaxillae (bones of the upper jaw), and a process resembling a spout that protruded from the lower jaw and may have supported horny plates for cropping. This animal had five upper teeth and four lower teeth in each jaw, the first upper pair of which resembled canines. The small Pleistocene ground sloth *Nothrotherium* apparently was a descendent of *Hapalops* and the two comprise a branch of the Megatheriidae, which throughout its history was given to small size. This stock differed greatly from the main stem of the family, which had no specialized "canines" and had the two inner toes reduced, represented by the Pleistocene *Megatherium*, a massive ground sloth larger than an elephant.

The megalonychids, although related to the megatheriids, were distinguished from them by their definite canine-like teeth and lack of the spoutlike process on the lower jaw. The megalonychids ranged in size from one island-dwelling animal no bigger than a domestic cat to the mainland ox-sized *Megalonyx*. Both the Megalonychidae and the Megatheriidae reached southern North America by the Pleistocene, but they became extinct on the Continent at the end of the epoch. Some megalonychids invaded the West Indies, however, and experienced a minor radiation in Hispaniola, Puerto Rico, and Cuba. One or more species may have flourished there to within a few hundred years of the present.

Distinct from the megatheriid–megalonychid line of ground sloths were the Mylodontidae, which first appeared in the Oligocene; from the Pliocene until their extinction at the end of the Pleistocene, they were the dominant South American group of ground sloths. The mylodonts differed from the other ground sloths in their partially developed upper canines, reduced first toe of the hindfoot and rounded ossicles (bony tubercles) in their skin. The mylodonts followed the trend of other ground sloth families toward great size and awkwardness, which in this group culminated in the Pleistocene giant, *Glossotherium*, which was the size of a large ox. As with other xenarthran groups in the Pleistocene, the mylodonts extended as far north as the southwestern United States.

The other family of ground sloths, the Orophodontidae, is known only from the Oligocene of South America. This briefly-lived side group is represented by *Orophodon* and *Octodontotherium*, both of which were rather small sloths that do not seem to have left any descendants.

The final infraorder, Vermilingua, comprising the anteaters (Myrmecophagidae), is known no earlier than the Miocene. The fossil myrmecophagids, as represented by the Pliocene form *Palaeomyrmedon*, were already completely toothless and presumably well adapted to an insectivorous diet. The Myrmecophagidae first appeared in South America in the early Miocene and apparently only fairly recently have spread to Central America.

CLASSIFICATION

Distinguishing taxonomic features. The degree of development of certain sets of structural characteristics taken in concert defines the Edentata and its subgroups. None of these features is universally distributed throughout the order. The most important are as follows: (1) in addition to the normal zygapophyses (protuberances), the occurrence of supplementary articular surfaces called xenarthrous articulations between the arches of sequential posterior trunk vertebrae; (2) development of certain projections of the scapula bone, or shoulder blade, called the acromion and coracoid processes; (3) articulation of the ischium bone of the pelvis with the proximal caudal (tail) vertebrae, forming an elongate sacrum; (4) length and massiveness of limb bones; (5) fusion of distal limb bones; (6) hyperdevelopment of claws; (7) development of forefeet and hindfeet; (8) a skull characterized by a small brain housed in a long tubelike brain case, typically with a reduced premaxilla (*i.e.*, elongate palate), often an incomplete zygomatic arch, and a reduction in number of teeth and their enamel covering.

Annotated classification. The classification presented here combines those proposed by the two paleontologists R. Hoffsteter and Alfred S. Romer. The features given below are only those that distinguish each group from the others. Groups indicated by a dagger (†) are known only as fossils.

ORDER EDENTATA

Placental mammals with reduced dentition, often lacking enamel; heavily clawed forelimbs for burrowing. Coracoid process of the scapula more strongly developed than in other eutherian mammals. About 31 recent species.

†Suborder Palaeanodonta

Fossil only. Bony carapace and the xenarthrous articulations on the vertebrae lacking; no fusion between the ischium and caudal vertebrae; simple teeth, usually single-rooted and lacking enamel; slight inward twist to the forelimbs.

†*Family Metacheiromyidae.* Upper Paleocene to lower Eocene. Small, lightly built edentates with sharply pointed somewhat laterally compressed canines; postcanine teeth reduced or lacking; armadillo-like feet with 4 to 5 digits; the lateral digits reduced, the inner greatly so. North America.

†*Family Epoicotheriidae.* Lower Eocene to lower Oligocene of North America. Very small burrowing edentates, with a domed skull having a depressed snout and possessing a zygomatic arch; having 6 slightly differentiated cheek teeth, nearly cylindrical, lacking enamel, and single-rooted.

Suborder Xenarthra

Some extinct and all living edentates; all with xenarthrous posterior trunk vertebrae. Lack teeth or have only poorly differentiated cheek teeth, lacking enamel and having single roots. Fusion (symphysis) between the ischium and the anterior caudal vertebrae in all except the 2-toed anteater (*Cyclopes*).

Infraorder Loricata

Bony plates covered with horny plates, forming a protective carapace over the trunk.

Family Dasypodidae (armadillos). Lower Eocene to Recent in South America; Pleistocene to Recent in North America; forests, savannas, and plains; primarily in the tropics and sub-

tropics. Teeth more than $\frac{6}{6}$ (6 above and below on each side); top and sides of the body covered with horny scutes over a bony, flexible carapace.

†*Family Peltephilidae.* Upper Oligocene to lower Pliocene (questionably) in South America. Teeth $\frac{7}{7}$; short, broad skull; hornlike structures on the head armour.

†*Family Glyptodontidae* (glyptodonts). Mid-Eocene to Pleistocene in South America; upper Pliocene to Pleistocene in North America. Fusion of bony plates on the back into a solid turtle-like carapace; teeth $\frac{8}{8}$ with 3-lobed pattern.

Infraorder Pilosa

Incomplete zygomatic arch (bony ridge below eye); 4 or 5 simple cheek teeth; expanded acromion, which unites with the coracoid region.

†*Family Orophodontidae.* Oligocene in South America. Bilophodont teeth, $\frac{5}{4}$, made of a mass of compacted dentine; highly specialized astragalus (ankle bone). Small to medium body size.

†*Family Megalonychidae* (ground sloths). Upper Pliocene to Pleistocene in South America; middle Pliocene to Pleistocene in North America; Pleistocene in West Indies. Ground sloths. Large canine-like teeth in the upper and lower jaws and no spoutlike terminus on the lower jaw.

†*Family Megatheriidae* (ground sloths). Upper Oligocene to Pleistocene in South America; Pleistocene in North America. Partial development of upper "canines"; spoutlike terminus on the lower jaw.

†*Family Mylodontidae* (ground sloths). Upper Oligocene to Pleistocene in South America; Pleistocene in North America. Distinguished by a partial development of upper "canines," triangular-shaped cheek teeth, and a hindfoot with the 1st toe reduced.

Family Bradypodidae (tree sloths). South and Central America; tropical forests; no fossil record. Dental formula of $\frac{5}{4\text{-}5}$; 6 to 9 cervical vertebrae.

Infraorder Vermilingua

Toothless and hairy edentates.

Family Myrmecophagidae (anteaters). Lower Miocene to Recent in South America; Recent in Central America; tropical forests and savannas. Teeth lacking; tubular mouth with a small terminal opening.

Critical appraisal. The edentates are a natural assemblage of several seemingly rather diverse groups. Given that the oldest known representatives of the palaeanodonts and the xenarthrans are too advanced for either to be ancestral to the other, it would be of great interest to find the common ancestor of both groups, as well as the insectivore precursor of this ancestor. To date, no good fossil candidates have been unearthed in either category.

Authorities disagree on the relationships and degree of distinctness of some groups of fossil edentates. The palaeanodonts are believed by some to be closer to the Old World pangolins (order Pholidota) than to the New World edentates. The armadillo-like *Pseudophorodon* from the upper Oligocene of South America is sometimes placed in its own family but is considered here to belong in the Peltiphilidae. The palaeopeltids, known especially from polygonal plaques of bone, are considered by some to have been aberrant glyptodonts but are considered by Romer to have been the distinct superfamily Palaeopeltoidea. The unusual Miocene ground sloth *Entelops*, due to the presence of teeth on the premaxillary bone, is sometimes separated from the Mylodontidae as sole member of the family Entelopsidae.

The degree of relatedness of sloths and anteaters has been the subject of disagreement. It has been shown that despite certain adaptations of the front limbs and digestive tract for insect eating, anteaters are close to the most primitive xenarthran edentates in many characteristics of the forefeet and hindfeet. Hofstetter has suggested that the evolutionary lines that produced the modern Myrmecophagidae and Bradypodidae separated early and that, therefore, the two families are not closely related but should be placed in separate suborders.

The origin of the tree sloths and the relationship of the currently recognized two genera is a question of some interest. As mentioned above, the two genera may be independently derived from two extinct families of ground sloths. *Choloepus* shows affinities to the megalonychids and *Bradypus* is more like the megatheriids.

At one time, the Old World scaly anteaters or pangolins (order Pholidota) and the South African aardvarks (order Tubulidentata) were included in the Edentata. The adaptations for an insectivorous diet that these two groups seem to have in common with the New World anteaters and the armadillos are most assuredly a result of convergent evolution, and not a reflection of any close phylogenetic relationship.

BIBLIOGRAPHY. The following works, although broad in scope, are particularly rich in information on this group: J.C. BARLOW, "Edentates and Pholidotes," in S. ANDERSON and J.K. JONES, JR. (eds.), *Recent Mammals of the World* (1967), a discussion of the taxonomy of this order and its families; J. DORST, *South America and Central America: A Natural History* (1967), many fine photographs of edentates; E.R. HALL and K.R. KELSON, *The Mammals of North America*, 2 vol. (1959), descriptions and distributional information on North and Central American members of this order; R. HOFFSTETER, "Xenarthra," and R. SABAN, "Palaeanodonta," in J. PIVETEAU (ed.), *Mammifères evolution*, pt. 6, vol. 2 of *Traité de Paleontologie* (1958), detailed accounts of early edentates; and E.P. WALKER, *Mammals of the World*, 2nd ed., vol. 1 (1968), an account of the natural history and distribution of every genus in the order.

(J.C.B.)

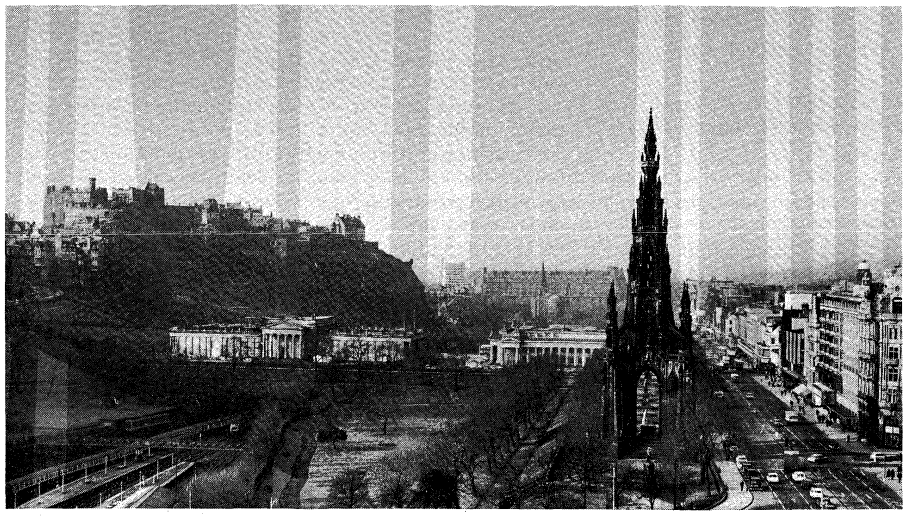
Edinburgh

Edinburgh, the capital of Scotland, like many of its northern European counterparts, is a dramatic and spacious city. While its character may seem lacking in warmth, it exhibits a striking, if somewhat aloof, nobility. The dramatic qualities of the city lie in its site: a rough valley, punctuated by upthrusting crags (there are five extinct volcanoes within the city limits) and scoured by Ice Age glaciers, slopes down to the port of Leith, two miles away on the southern shore of the Firth of Forth, an irregularly diminishing arm of the North Sea that thrusts westward into the heart of the Scottish lowlands. The wind, frequently whistling in from the sea and the abrupt heights, sounds a wild note among the urbane expanses of the modern city.

The city, seen from on high, still gives the impression of being rather sparsely strewn amid open country, although its population had reached almost 500,000—nearly 10 percent of the inhabitants of Scotland—by the 1970s. Only a little more than half of the area of the district, 64,000 acres (26,000 hectares), is occupied by buildings; there are more than 550 miles of city streets, and some of the free space remains relatively untamed hill and moorland. Within the confines of Edinburgh there exist entirely rural villages such as Duddingston, sited beside a loch behind the 824-foot (251-metre) hill known as Arthur's Seat.

Edinburgh Castle, 443 feet (135 metres) above sea level, dominates the central city from atop basalt cliffs, which themselves drop a sheer 270 feet to a grassy valley planted with trees and flowers, the drained North Loch. On the opposite shore of this former lake, northward toward the sea, there rises an elegant gray Georgian town of the 18th century. On either side of the valley, east–west thoroughfares, one called the Royal Mile, the other named Princes Street, form the main axes of Edinburgh. The Royal Mile descends Castle Hill, a spur of rock linking the castle with the Palace of Holyroodhouse; the towering, 16th-century-style tenements of the original burgh, the Old Town, are pressed together on this crest.

Changes in light and weather make for a constant and subtle shift in the vistas stretching away across the valley separating the Old Town from the New. The ravine itself, drained in 1816, now houses the Princes Street Gardens, one of the most remarkable of urban parks, with main railway lines running deep along its centre. To the south of the castle, beyond the Meadows, which occupy the site of the drained South Loch, there are additional splendid views of the Pentland Hills (1,898 feet [579 metres] high), formed of igneous rocks heaved up by



Princes Street, showing Castle Rock overlooking the city, the Scott Monument (foreground), the National Gallery of Scotland, and the Royal Scottish Academy, Edinburgh.
A.G. Ingram Ltd., Edinburgh

earth movements and leading to the Southern Uplands beyond. Although Edinburgh absorbed surrounding villages and the Firth of Forth ports between 1856 and 1920, the mood of the whole is still generated within the small compass of the Old Town and New Town, the former huddled high as a result of medieval fears of attack, the latter serenely spread out in a magnificent succession of streets, crescents, and terraces, somehow reflecting the optimism of the 18th-century Enlightenment.

"This profusion of eccentricities, this dream in masonry and living rock is not a drop-scene in a theatre," wrote Robert Louis Stevenson, the 19th-century Scottish novelist, essayist, and poet, who was born in the New Town, "but a city in the world of reality." The contrasts that make Edinburgh unique also make it typically Scottish, for, despite its reserved exterior, it is also a city capable of great warmth and, upon occasion, gaiety, while its history demonstrates that its citizens have also been capable of great passion, especially in matters royal or religious. In 1736 the burgh almost lost its royal charter following the lynching of one John Porteous, captain of the city guard. This occurrence can be viewed as a violent gesture common to the history of most old cities, yet, even in this moment of conventional madness, the city manifested its complex character: needing a hanging rope, the mob descended on a shop and bought one.

A wealthy city by European standards, Edinburgh shuns ostentation. As far back as 1820, Stendhal, one of the greatest of French novelists, noted that the city had "emancipated itself from the mean omnipotence of gold." In a society relatively unmarked by modern class conflict, one in which public politeness and private consideration still prevail, it is not necessarily held against a man that he has had a head start in life. Having for centuries been a bulwark of free education for all, and endowed with a traditionally superior state school system, the Edinburgh of the 1970s still had a greater proportion of private school students than any other city of the British Isles—20 percent of the school population, including children of families with very modest means.

Long history of free education

A city long renowned for a somewhat inflexible respectability—when west Princes Street Gardens were turned over to the general public in 1876, smoking was forbidden—Edinburgh concurrently maintained a fascinating netherworld of ribaldry and drunkenness. A poet or jurist of sufficient distinction might succeed in inhabiting both worlds. Such "Edinburgh characters" abounded during the flourishing neoclassical period of the 18th and 19th centuries known as the Augustan age, when the city's authors, critics, publishers, teachers, physicians, and scientists formed an intellectual elite of world influence. With the subsequent relapse of the city into a more provincial role, such noted eccentrics have become virtually extinct. Since 1947, Edinburgh has shone anew as an international focal point for the arts during the three weeks of its annual summer festival. Almost 100,000 visitors come for the theatre, ballet, music, films, art expositions, and the general excitement, which closes with a skirl of the Scottish bagpipes before the castle gate.

Regular attendance at church is still common to the citizens of Edinburgh, although Sabbath observances are less austere than those molding the social life in previous centuries. Even in the later years of the 19th century, a noted foreign visitor leaving church was advised not to proceed too fast, "lest people think we are just out for a walk." Men, whose costume during the week may well be a sombre business suit, sometimes dress for church by donning the kilt, once prohibited by the English. The kilt is seen on other days of the week on soldiers of the Scottish regiments (the army has its national Scottish headquarters in the city) and is also worn by visitors from the Highlands and islands. As late as 1970, services at Tolbooth St. John's in the Old Town were still being conducted in Gaelic, the fast-diminishing native tongue of Scotland. A measure of the relaxation of sabbatarian severity is evidenced by the number of ladies who now attend church hatless, while Sunday cricket and soccer matches are allowed on the municipal playing fields after 2:00 PM, weather permitting. Costumes other than the workaday or the tartan are rarely seen on the streets of Edinburgh except during the important high tourist season: of only marginal industrial importance, the city has not attracted the great numbers of Commonwealth immigrants who have been drawn to other leading British cities. A cosmopolitan group around the university has nevertheless leavened Edinburgh society to some extent. The medical school, in particular, has attracted overseas students, and one of the first Africans to take a medical degree in Britain, Dr. James Africanus Beale Horton, was graduated from there in 1859. This tradition has found its latest expression in the Centre for African Studies, established in 1963.

Social observances and the kilt

GROWTH AND TOPOGRAPHY

The castle and the earliest period. The Roman invaders of Britain established a camp on the shores of the Firth of Forth, where the little Almond River debouches, and the village of Cramond on this site, together with the port of Leith and other, smaller, settlements, became part of expanding Edinburgh in 1920. There is no evidence that the legions installed themselves at Castle Rock, four miles further inland, where the towering crag was to become the nucleus of the city. The Gaelic name Dun-eideann means "hill fort of Eiden," and, as far as historical records indicate, fortifications have existed on the rock since at least the 6th century AD. The oldest surviving construction is 11th century; the impressive cliffhanging ramparts, with attendant sentinel towers, are large-

ly of 18th-century origin. The history of the rock and of its stony spine trailing off to the east is, for the first thousand years, not only the history of a city but also that of the Scottish nation, a nation successively seeking to be born, to defend its independence, and to preserve its identity.

The king of the castle in the 11th century was Malcolm III Canmore (1058–93), whose father, Duncan, had been murdered by Macbeth, the dark event providing the inspiration for the Shakespearean tragedy. Malcolm's Saxon queen, St. Margaret, erected a small, square chapel in the castle grounds at the highest point of the rock. Rebuilt a century later, the simple building is the oldest to survive in the city. It was not harmed when Edward I of England (1239–1307), after an assault with stone-slinging engines, took the place by storm in 1296 at the start of the bloody Wars of Independence (the Scots took it back 16 years later). When the Scottish king Robert I, the Bruce, (1274–1329)—who granted Edinburgh's first extant charter in the year of his death—ordered the castle razed for military reasons, he specifically spared the chapel, in which the saint had heard her dying mass in 1093. Services are still occasionally conducted there.

Nearby are the massive ruins of David's Tower, erected by David II in 1367–79. The tower's ruins are now entirely enclosed by the Half-Moon Battery (1574), from which a modern 25-pounder fires the daily one o'clock signal. Simultaneously, a white ball drops from the top of the 102-foot-high Nelson monument, three-quarters of a mile away on Calton hill. Since this firing began, in 1858, it has been an Edinburgh pleasantry to pretend to admire the gunnery of the castle cannoneers.

Palace Yard, or Crown Square, occupies much of the summit of Castle Rock and is surrounded on three sides by the 16th-century royal palace, with the Scottish National War Memorial (1927) completing the closure on the north. On the south side, the Great Hall, restored in 1888, preserves the features of an imposing medieval banqueting hall, and ceremonial gatherings are still held under the original richly carved hammer-beam roof. On the east, in the King's Lodging, the "Honours of Scotland" (royal crown, sceptre, sword of state, and other 16th-century regalia), hidden from 1707 to 1817 after the Union with England, are on public view, as are the royal apartments. These include the quarters of Mary of Guise (Lorraine), who died, it is said, when she scanned the Forth for sight of a French fleet and saw instead an English one. Her daughter Mary, Queen of Scots (1542–87), gave birth in 1566 to James VI (James I of England) in a room unregally tiny but extraordinarily secure, with a single window giving onto the perpendicular cliff. The Scottish United Services keep a museum of Jacobite relics and Highland weapons and uniforms in the palace.

The extensive precincts of the rock are not all given over to sight-seeing: in addition to the regional command offices of the army and the small guard, the Piobaireached Society continues the School of Piping. Outside the gates is the bare, broad Esplanade, from which tourist buses and parked cars are banned during traditional military displays. A corner near the drawbridge contains Nova Scotian soil, as mandated by Charles I (1600–49). To attract Scottish settlers to the North American colony, baronetcies were offered for sale. Before such title had legal force, each new baronet was obliged to take seisin —*i.e.*, to accept a morsel of his lands from the crown—hence the alien soil on Scottish ground. In the northeast corner of this open space a metal wellhead marks the spot where 300 or more alleged witches were burned at the stake between 1479 and 1722.

Importance of sight-seeing

The medieval Old Town. The original burgh of Edinburgh, now known as the Old Town, rose under the castle walls on the slopes of the ridge to the east, bounded by streams on north and south. At the same time (during the reign of David I [1124–53]), the Augustinian Abbey of Holyrood was founded at what is now the eastern end of the Royal Mile. The King granted canons the right to erect their own burgh, which, under the name of Canongate, remained a self-governing entity until its incorporation into the city of Edinburgh in 1856.

The embryonic town had suffered with the castle under the ravages of the Wars of Independence, and it was razed again during the invasion of Richard II in 1385. Only during the reign of James II (1437–60), who was born and crowned at Holyrood, did Edinburgh begin to emerge as the Scottish capital. It became the usual meeting place of the national Parliament, and developed a municipal administration, together with the beginnings of a power struggle, between craft guilds and merchants, that was typical of evolving European cities of that period. The first city wall was constructed, and a second street, the Cowgate, was built on the southern slope of the ridge, parallel to the main street. In the next century, after the terrible defeat of Scottish arms at Flodden in 1513, another wall, thicker, and of wider compass—it reached halfway down the Royal Mile—was hastily thrown about the town as protection against an imminent threat of assault, which did not materialize.

The Flodden wall

In spite of religious and political turbulence—there was further destruction by an English army in 1544, and the tragic reign of Mary—the 16th century was a period of consolidation. In 1505 the Royal College of Surgeons was chartered and the practice of dissection legalized. Little use was made of this privilege for two centuries, after which Edinburgh became one of the great medical teaching centres of the Western world. The College of Justice was founded in 1535, and, in 1560, the first General Assembly of the Church of Scotland convened in the Magdalen Chapel in Cowgate, which still stands. Further, in 1583, James VI settled the squabbling between the craft guilds and the merchants' party over the composition of municipal government, giving it a form it was to keep for 250 years. A year earlier the council had given orders for the construction of the "Toun's College" at the site of the ancient collegiate church of St. Mary in the Fields ("Kirk o' Field"), scene of the unsolved gunpowder murder of Lord Darnley, Queen Mary's husband. The college of 80 students, under a single teacher, was to grow into Edinburgh University.

The church, the law, national and municipal legislatures, and advanced learning were thus all established in Edinburgh when James VI also became James I of England. He moved his court to London in 1603, two years after the union of crowns. There followed a precipitate fall in the prestige and prosperity of Edinburgh, but in 1633 Charles I declared the burgh an episcopal see and made the ancient church of St. Giles a short-lived Episcopal cathedral. Five years later there came the National Covenant, signed—sometimes in the blood of the signatory—by nobles and commoners alike in Greyfriars churchyard. While loyal to the King as their monarch, the signatories rejected his policy of enforcing the Anglican Episcopacy on Presbyterian Scotland. Edinburgh escaped the battles that subsequently flared under both Stuarts but did not escape the dire consequences. From June to September 1679, more than 1,200 Covenanters, captured in the battle of Bothwell Bridge, were imprisoned in the same churchyard where the Covenant had been signed. Many died from starvation and exposure; others were sold into West Indian slavery, and 100 were executed in the Grassmarket.

After the Union with England in 1707, resulting in the abolition of the Scottish Parliament, the town again declined, and riots were common. The town was nevertheless to emerge into its "Golden Age" by the mid-18th century. Then a city of 50,000 people, Edinburgh had grown from fewer than 8,000 inhabitants 100 years earlier and only 25,000 some 50 years before. Almost all these 50,000 were squeezed behind the Flodden walls, in buildings that climbed ever higher. Then, in the 1760s, the old city vaulted its valleys and spread to both the north and the south (see below *The Meadows and the university; Princes Street Gardens;* and *The New Town.*

The Royal Mile bears several names on its way to Holyrood: Castle Hill, Lawnmarket, High Street, and Canongate. The cliff-face of houses is broken by wynds—narrow, winding, stone lanes leading down either side of the ridge, sometimes not much more than long staircases, some with a left-handed spiral that favoured defending

swordsmen—and closes or vennels—entranceways into courtyards, around and behind which are yet more buildings, towering and beetling. Everything was jammed together to make space where there was none, ". . . piled deep and massy, close and high," as Sir Walter Scott, the nation's leading writer, described the area.

Only the wealthiest could afford to occupy one of the heavy-timbered and gabled houses fronting directly on the street, and the closes and courtyards harboured not only families but also workshops, taverns, and clubs. At Anchor Close the parents of Sir Walter Scott lived until 1777, the editor and writer William Smellie brought out the first serials of *Encyclopædia Britannica* in 1768, and the Anchor Tavern provided food, drink, and fellowship. In James' Court, off the Lawnmarket, lived one of the leading 18th-century philosophers, David Hume. In the same building the diarist and biographer James Boswell entertained Samuel Johnson, in 1773, regretting, after the latter's remark on the smells of High Street, that although the magistrates "enforce city laws against throwing foul water from the windows . . . there being no covered sewers the odour still continues." In the same year, Robert Fergusson, one of the leading figures of a renascent Scots poetry, cited "the morning smells that hail our street" where "clarty [filthy] Odours fragrant flow." "Auld Reekie," the affectionate name for Edinburgh, may thus not have referred exclusively to the smoke and sea mist mantling the town.

Water supply was a great problem until it was piped in, in 1681, to serve street wells, which, when they became contaminated under later, more crowded, conditions, spread epidemics of typhoid and cholera. The linkboys, who ran errands and carried jugs of water up to householders, were called "caddies," a Scotch version of the French *cadet*. The name survives on golf courses.

Once the city had spread off the hill, although the town council moved effectively to end overcrowding and to institute modern municipal services, the Royal Mile decayed badly. Courts, the church, and the city government nevertheless remained in place on the ancient thoroughfare, and in the mid-20th century the council and private persons resolved to restore it. Five of the old buildings are now museums, while others are used as offices or for educational purposes. Tenements have been razed, ugly courtyards scoured, and the splendid old courts restored to their original state and made into desirable dwellings.

The High Kirk

In the case of St. Giles, the National Church of Scotland, an earlier, but less thoughtful, generation of rescuers had already (1829) paved the exterior with good intentions. Happily, the magnificent 15th-century crown tower, an open spire with eight flying buttresses supporting a sculptured turret, was left untouched. Many of the church's chapels are authentically late medieval. The official city guide says, "The High Kirk of Edinburgh is the very heart of Scotland." Still in use, it has been a parish church, a collegiate church, and for two short periods (1633–38, 1661–89) an Episcopal cathedral.

Behind St. Giles is Parliament House, built by the Town Council (1632–39), with the aid of private contributions, when Charles I threatened to move the law courts elsewhere. The Scots Parliament met there from 1639 to 1707. The supreme civil and criminal courts of Scotland now sit in the building, administering Scots law, which has been characterized as somewhere midway between the Anglo-American school of legal thought, deriving from English common law and equity, and the continental school, based on Roman law.

Opposite the cathedral are the City Chambers, which is the venue of the City of Edinburgh District Council (from May 1975), consisting of 64 popularly elected members, one from each of the city's wards. Intended to be the Royal Exchange, the building was erected in 1753, using the designs of John Adam, the Neoclassicist architect.

At the foot of the castle crags, on the south slope, the Cowgate emerges at its western end into a broad, tree-planted square, the Grassmarket, where until 1911, cattle and grain were sold weekly. On the north side is the White Hart Inn, patronized by such figures as the poets Robert Burns and William Wordsworth. A flight of steps on the southwest corner is enclosed on the left side by a stretch of the Flodden Wall; on the other side is George Heriot's Hospital, built in 1628, following designs of Inigo Jones, the founder of the classical school of British architecture. This school is not only one of Edinburgh's noblest buildings but also its oldest privately endowed educational charity. Tanner's Close once stood on the same corner of Grassmarket. Here, in 1828, the "Resurrectionists" Burke and Hare purveyed cadavers for dissection to Dr. Robert Knox of the Medical School. At first they robbed graves, then took to obtaining cadavers wholesale, by murder, and their trial was one of the most celebrated in Scots criminal annals. The nearby graveyard of Greyfriars Church still has iron trellises on some tombs to foil body snatchers, as St. Cuthbert's burial ground has a watchman's tower in heavy stone.

The Meadows and the university. Southeast of this neighbourhood, close to the edge of the Meadows (the drained South Loch), the architect James Brown developed a Neoclassical square (1766), which he named George for his brother. This was the first move in developing the town southward beyond the old walls. Today the university's new (1966–75) complex occupies three sides of the square, with a row of Georgian houses preserved along the other. The university's old quadrangle, some 300 yards closer to the Royal Mile, was started in 1789 and completed in 1834. The 1854 Royal Scottish Museum is adjacent, and the nearby Cowgate houses the city's central library, across the street from which is the 1854 National Library of Scotland (3,000,000 volumes, founded 1682), abutting Parliament House.

Holyrood. Away from the crowded buildings, at the lower end of the Royal Mile, is Holyrood. The now-roofless abbey was founded in 1128 and rebuilt about 1220. The substantial remains are impressive examples of the bold, imaginative work of the period. The Palace of Holyroodhouse, which shouldered the abbey aside, is intact in its early 16th-century wing and in the 1671 extensions. It was to this palace that the widowed Mary, Queen of Scots came from France, to find herself the symbol of Catholicism and continental refinement opposed by (among others) the fiery John Knox of St. Giles Kirk, who represented the new Calvinism and the old, rugged Scottish ways. The rooms in which Mary's Italian secretary was dragged from her side and murdered are now on public view, as are the State apartments used during royal visits to Edinburgh. Bonnie Prince Charlie also stayed here during the attempted Stuart restoration of 1745. Its 110 royal portraits, from "Fergus I, 330 AD" to "James VI (1566–1625)," were all painted by the Dutch artist Jacobus de Wet.

Royal traditions

The abbey's monks began making beer in the 12th century, and today the tradition is maintained by a nearby commercial brewery, sharing Holyrood's soaring view of the Salisbury Crags and Arthur's Seat. These two eminences are among the five in the 648-acre (262-hectare) Holyrood Park, which also embraces three lochs.

Princes Street Gardens. The 37 acres (15 hectares) of the Princes Street Gardens, lodged between the Old and New Towns, have a distinct character. The old North Loch, which had been created out of a swamp by James II (1437–60) as the castle's northern defense, was the sump for sewage draining from the Old Town. Doctors clamoured for its destruction long before it was achieved in 1816. Flowers are set out in beds changed 10 or 12 times a year, and a floral clock (1903, first in the world), which embowers a quarter-hour cuckoo, has some 24,000 plants in its 36-foot circumference. Among the lawns, flower beds, and groves are recreational areas, a bandstand, an outdoor dance floor, and numerous memorials, the most conspicuous of which is an 1844 Gothic spire, 200 feet high, that rises above a statue of Sir Walter Scott and his hound, Maida.

The floral clock

For the first 100 years of its existence, west Princes Street Gardens was the private amenity of Princes Street proprietors. In 1876 this tract was opened to the public, which had always had access to the east gardens. The

Mound, a causeway of rubble and earth from New Town construction, which crossed the ooze of the recently exposed lake bottom, forms the division between the two gardens. On the Mound stand two neo-Grecian temples to the arts: the Royal Scottish Academy (1832) and the National Gallery of Scotland (1859). These buildings are now seen to be lightly tinged with ochre, but for more than a century they were soot black: locomotives passed under the Mound, railroad tracks—now almost concealed by landscaping—having been laid through the middle of the gardens in 1847. The rails terminated in the east end of the Park at Waverly Station, which covered some 25 acres, most of it under glass.

There were other attempts to bring progress (as envisaged during the 19th century) to the gardens, notably those to erect buildings on the park side of Princes Street. The fights were hard and bitter, but public-spirited citizens eventually prevailed over private interest.

The New Town. The town council approved plans for the New Town in 1767. Its members broke with tradition in planning a town at all, but exceeded even that act of courage and imagination by planning the converse of the Edinburgh that had existed for 600 years. The architect, James Craig, envisaged a grid five streets deep and seven streets wide, its broad central axis terminating in grand squares at each end. St. George's Church would anchor the western end of the scheme, St. Andrew's the eastern. St. Andrew's Square boasts a column in Roman style, topped by a statue of Henry Dundas, 1st Viscount Melville. A member of his family, Sir Laurence Dundas, refused to let St. Andrew's Church be built on the spot he had selected for his mansion. There was no argument, for Sir Henry was the uncrowned king of Scotland, serving as the English prime minister's proconsul in the Scottish capital. Under him, the corrupt Town Council concealed the fact that the city was bankrupt at the end of the 18th century, a state of affairs not divulged until years later. The Dundas house, a fine design of 1772, later became the headquarters of the Royal Bank of Scotland. Extension schemes by other architects were coupled onto Craig's New Town: in 1804 the original plan was reproduced as well as the terrain to the north allowed, while in 1822 there evolved a sumptuous ensemble around a crescent, an oval, and a polygon, connected by short streets. Later expansion developments lacked the Georgian élan, and elegant style was replaced by unimaginative convention. By the 1970s, a scattering of modern glass-fronted structures had replaced grand residences ill-adapted to the contemporary requirements of, for example, the Stock Exchange or the Scottish Widows' Fund and Life Assurance Society. Robert Adam's superb Charlotte Square (1791), the showpiece of New Town, remains intact. Princes Street (originally it trailed an apostrophe, being named after two sons of King George III of England) became the shopping street and the principal thoroughfare of the city, and there is scarcely one original building behind the shop fronts.

Architectural styles of Calton Hill

At the east end of Princes Street is another of the city's extinguished volcanoes, Calton Hill (355 feet [108 metres] high). Its summit proclaims the passion for the classical that gripped Edinburgh architecture in the first half of the 19th century, before Gothic styles became dominant. With the exception of the Nelson monument (1806–16), a structure exhibiting neither Greek nor any other known architectural style, the hilltop monuments are all of Attic extraction. A small temple gracefully honours the memory of the philosopher Dugald Stewart (1753–1828). A porticoed fortress with cupolas is an observatory designed by Robert Adam for the university but completed by the municipality when the original funds were exhausted in 1792. Within the Adam central enclosure, a cruciform structure, with four pedimented facades supporting a dome, is an observatory that W.H. Playfair designed in 1812 for a local astronomy group (the Royal Observatory, 1859, is across town on Blackford Hill). The 12 stark columns, standing like a ruin from some vanished Aegean civilization, are the beginnings—and remains—of Playfair's Parthenon, a project launched with fervor in 1822 as a memorial to Scots who served in the Napoleonic Wars but tagged "Scotland's Pride and Poverty" when the money ran out in 1830.

In a niche in the hill's south flank is an adaptation (1829) of the Temple of Theseus in Athens. This housed the Royal High School (founded at Holyrood in the 11th century) before it moved to suburban Barnton in 1968.

Across the road from the high school is St. Andrew's House (1937), architecturally undistinguished but politically powerful, containing the administrative centre of government in Scotland.

CONTEMPORARY DEVELOPMENTS

Edinburgh entered the 1970s with 159,099 inhabited dwellings, 48,170 of which had been built by municipal housing authorities, almost 20,000 of them less than a dozen years earlier. Ranging from small cottages, through reconstructed historic precincts, to multi-story complexes, they have been conceived with unusual sympathy for the needs of the inhabitants, the character of the neighbourhood, and the requirements of the site. By far the greatest number have been erected on the periphery, between three and four miles from the city centre. Planning is complete for several renewal projects in older central districts. Some of the recent installations house employees at Sighthill Industrial Estate, where about 45 light industries have been established and more are planned.

The urban economy. In the 1970s, manufacturing employed only one-quarter of the people who worked in Edinburgh. The three main manufacturing groups were food, drink, and tobacco; engineering (including ship repairing and marine engineering), electrical goods, and electronics; and paper, printing, and publishing. The capital is Scotland's centre of medicine, law, banking, insurance, and investment and is second only to London as a United Kingdom tourist city.

The port of Leith, about 30 miles from the open sea, has been part of the six-port Forth Ports Authority since 1968 and handles nearly 2,000,000 tons of cargo annually. Grain, foodstuffs, and wood products were the principal incoming products; outbound shipments being coal, whiskey, iron, and steel. Extensive modernization was completed in the early 1970s. Granton, Edinburgh's other port town, is also under the authority, with special facilities for rapid loading and unloading; a sizable industrial estate to service North Sea petroleum is also under development there, and it is the home port for the Firth of Forth fishing fleet.

Population. In 1801 the city had 67,288 inhabitants: the urbanization associated with the Industrial Revolution stimulated growth to 228,357 by 1881. Significant extensions of the city boundary, to facilitate incorporation of surrounding settlements, occurred in 1896, 1900, and 1920, and the population had grown to 439,010 by 1931. By 1974 it had reached 449,910, and the Edinburgh City District, created in 1975, annexed another 25,000 persons.

Administration and services. Until the reorganization of local government of Scotland in May 1975, the Town Council performed most local governmental functions; its successor, the City of Edinburgh District Council, has somewhat more restricted responsibilities—housing, primarily, as well as local planning, libraries, museums, and sports grounds. Under the new scheme, the Lothian Regional Council has primary responsibility for education, transportation, social services, environment, and police, fire, and water services. The famous Edinburgh teaching hospitals, Western General and Royal Infirmary, are undergoing extensive rebuilding as part of a regional National Health Service scheme, costing over £60,000,000 ($144,000,000) and destined for completion in the 1980s. In the early 1970s, there were eight general hospitals, two maternity hospitals, one children's, and one orthopedic hospital—handling some 2,675,000 in- and out-patients. In addition, there were 29 child health clinics, 14 day nurseries, 33 play centres, and extensive facilities for home visits. The birth rate was 12.1 per 1,000, and the death rate 13.3 per 1,000 (with a low infant mortality rate of 16.6 per 1,000 live births).

The medical tradition

The city is served by Britrail, the national system. Princes Street railroad station was closed in 1965, but Waverly (second largest in Britain) and Haymarket remain. Local, regional, and national bus services are also extensive, and there is a civil airport located at Turnhouse, six miles to the west of the city. By the 1970s, the annual number of licensed motor vehicles in Edinburgh reached 100,000, 80 percent of them private vehicles, with all the attendant problems of parking and congestion.

Edinburgh has a remarkable proportion of recreational facilities per head of population. In addition to the expected spectator sport facilities, it has within its boundaries nine miles of coastline for boating, with a fair number of beaches; six public and 16 private courses for golf, which has been played there since 1457; 50 public bowling greens and as many private ones; 104 public football and hockey fields; 20 public cricket pitches; 113 public tennis courts; 28 public putting greens and short hole golf courses; plus a vast multisport centre constructed for the 1970 Commonwealth Games.

Education. By 1970, Edinburgh University, whose origins have already been described, had nearly 11,000 students from 90 countries, enrolled in 150 departments of study. The nearby Heriot-Watt college, dating from the earliest days of the Industrial Revolution, achieved university status in 1966, becoming one of the first of Britain's new technological universities. This institution, installed in the university area and at Grassmarket, has a 200-acre satellite campus just over the city line at Riccarton, Currie. Other establishments of higher education in the city include the Moray House College of Education; Craiglockhart College of Education, for Roman Catholic women; Dunfermline College of Physical Education, for women; Leith Nautical College, training ship's engineers, cooks, radio operators, and pursers. State nursery, primary, and secondary schools offered, by the early 1970s, some 2,700, 45,000, and 21,000 places, respectively.

The Edinburgh College of Art, in addition to its fine arts faculties, also has, under a faculty of Environmental Studies formed jointly with Heriot-Watt, schools of town and country planning and of architecture.

New College, Church of Scotland, is the largest postgraduate theological college in Scotland. William Henry Playfair, whose Greek galleries are just below on the Mound, designed the twin towers (1846) that glare down on the gardens from Castle Hill, solid, upright, and uncompromising, characteristics that have been of strong—but not overwhelming—significance in molding the life of the entire city.

BIBLIOGRAPHY

General: GEORGE SCOTT-MONCRIEFF, *Edinburgh*, 3rd ed. (1965), a detailed account of every aspect of the city, illustrated from photographs, prints, and paintings; IAN NIMMO, *Portrait of Edinburgh* (1969), the city and its inhabitants, their way of life, and culture; WILLIAM GORDON SMITH, *Edinburgh* (1967), text and photographs by author, the principal districts of the city today and yesterday; ERIC LINKLATER, *Edinburgh* (1961), a personal, essayist treatment of the city in its variety.

The past: A.J. YOUNGSON, *The Making of Classical Edinburgh: 1750–1840* (1966), investigates the growth of Edinburgh from a small, crowded town to a substantial and beautiful city; DOUGLAS YOUNG, *Edinburgh in the Age of Sir Walter Scott* (1965), the city when its vigorous intellectual life gave it the name "Athens of the North."

Chiefly photographic: ANTHONY KERSTING, *Portrait of Edinburgh* (1961), mainly architectural features, with text by GEORGE SCOTT-MONCRIEFF.

Special topics: ROYAL COMMISSION ON THE ANCIENT AND HISTORICAL MONUMENTS AND CONSTRUCTIONS OF SCOTLAND, *An Inventory of the Ancient and Historical Monuments of the City of Edinburgh* (1951), a detailed, illustrated catalog for each district of the city, including a list of monuments the commissioners deemed most worthy of preservation; DAVID KEIR (ed.), *The City of Edinburgh*, vol. 15 of *The Third Statistical Account of Scotland* (1966), a close examination of the life of the city in the second half of the 20th century; ROBERT CHAMBERS, *Traditions of Edinburgh*, new ed. reprinted (1967), written to preserve the stories of some of the characters who lived in Edinburgh at the beginning of the 19th century, first published in 1824, uses the author's final revision of 1868.

(B.E.)

Edison, Thomas Alva

One of the outstanding geniuses in the history of technology, Thomas Alva Edison included among his 1,093 patents the incandescent electric lamp, the phonograph, the carbon telephone transmitter, and the motion-picture projector. In addition, he created the world's first industrial-research laboratory.

Edison was born in Milan, Ohio, on February 11, 1847, the youngest of four children of Samuel Ogden and Nancy Elliott Edison. When Thomas was seven years old his family moved to Port Huron, Michigan. There the boy received his only formal schooling, which ended abruptly after three months, when the schoolmaster expelled him as "retarded." For the next three years his mother, a former schoolteacher, tutored him. "She instilled in me the love and the purpose of learning," Edison later recalled.

By courtesy of the Edison National Historical Site, West Orange, N.J.

Edison demonstrating his tin-foil phonograph, photograph by Mathew Brady, 1878.

First experiments and inventions. Curiosity was his dominant trait. When he was ten, his mother aroused his interest in an elementary physical science book, and in the cellar of his home he set up a small chemical laboratory. He found the study of chemistry and the production of electrical current from voltaic jars especially absorbing; soon he operated a home-made telegraph set.

A mettlesome child, he was given to pranks calculated to frustrate his father, who regularly beat him and on one occasion whipped him before the townspeople in the village square. The boy's middle-ear deafness, which had resulted from scarlatina, also contributed considerably to his troubles.

Early interest in science

As his father was a poor provider, Edison went into "business" for himself at the early age of 12, selling newspapers and candy on the new railroad between Port Huron and Detroit. He installed his little laboratory in a baggage car, where he occupied himself during train layovers. He later described this as "the happiest time" of his life. Because people had to raise their voices above the train's roar, he could hear them speaking, and at the stations he could follow the high click of the telegraph. His impaired hearing disposed him to enjoy solitude that gave him the opportunity to read and work in his laboratory. He also managed to train himself in telegraphy, and from 1863 to 1868 he roamed as a "tramp telegrapher," filling casual jobs in Midwestern and Southern cities, sometimes behind the battlefronts of the Civil War. Though living in the bohemian style of the roving telegraphers, he read all the scientific journals he could find, and in cheap furnished rooms of slum quarters ex-

perimented with modifications of the telegraph that would increase its speed and capacity. His earnings went mostly for electrical equipment and chemicals.

In 1868 he obtained a position in Boston as an expert night operator for Western Union Telegraph Company; by day he slept little, however, for he was gripped by a passion for manipulating electrical currents in new ways.

Influence of Michael Faraday

Soon after his 21st birthday he bought a used copy of Michael Faraday's journals, *Experimental Researches in Electricity*, which he promptly read without stopping. It marked a turning point in his life. The experiments of the self-taught English scientist were set forth so clearly that Edison instructed himself by performing all of them. He had already won the notice of his companions by his unusual powers of observation and his ability to "cure" balky instruments. Now, under the inspiration of Faraday, his experimental work became more painstaking, and he began to keep laboratory notebooks of his findings. By this time he had serious projects in view. Borrowing a small sum from an acquaintance, he gave up his job in the autumn of 1868, and became a free-lance inventor, taking out his first patent for an electrical vote recorder. Although the invention worked well when exhibited before a committee of Congress, no one bought it—a lesson Edison never forgot.

In the summer of 1869 he was in New York, sleeping in a basement below Wall Street, a typical, impecunious inventor. At a moment of crisis on the Gold Exchange caused by the breakdown of the office's new telegraphic gold-price indicator, Edison was called in to try to repair the instrument; this he did so expertly that he was given a job as its supervisor. Soon he had remodelled the erratic machine to such effect that its owners, the Western Union Telegraph Company, commissioned him to improve the crude stock ticker just coming into use. He performed this task by creating the Edison Universal Stock Printer, which, together with several other derivatives of the Morse telegraph, brought him a sudden fortune of $40,000. With this capital he set himself up as a manufacturer in Newark, New Jersey, producing stock tickers and high-speed printing telegraphs. Edison was married in 1871 to Mary Stilwell of Newark, a gentle but undereducated woman. Two sons and a daughter were born of this union, but the often distracted father tended to neglect them.

Western Union, then owned by the Vanderbilts and J.P. Morgan, ranked as America's leading electrical enterprise at the time. A group of talented inventors, including Elisha Gray and J.B. Stearns as well as Edison, was retained to develop its several electrical devices. Their knowledge of the field, like Edison's, flowed from an evolving technological environment. The young inventor's skill in eliminating "bugs" became legendary; he could also describe his electrical researches with considerable imagination. In order to gain permission to work all night at Western Union headquarters, he explained that he "wanted to feel the pulse of my patients." The company's operating chief called him a "genius" and "fiend for work" and remarked upon his unfailing self-confidence. When the financier Jay Gould set up a rival telegraph company to harass Western Union, Edison sometimes served under him also, acting as a scientific soldier of fortune in the camps of the warring "robber barons." His quadruplex telegraph in 1874 was his most ingenious invention of those earlier years: it doubled the capacity of the existing duplex of Stearns, simultaneously sending two messages in each direction over one wire.

In his Newark factory, Edison set aside a room equipped with a few scientific instruments, where he could plan and perform experiments according to his own fancy, unrelated to his duties at Western Union. It only served to sharpen his appetite; and in 1876, asserting that he feared he would be turned into a mere "bloated Eastern manufacturer," he gave up the Newark factory altogether and moved to the village of Menlo Park, New Jersey, to set up a laboratory where he could devote his full time to inventive work.

Years at Menlo Park. In a rectangular wooden-frame structure that looked like a rural meeting house he installed a quantity of fine measuring instruments, such as a galvanometer and an electrometer; supplies of storage batteries, induction coils, copper wire, chemicals, and metals; a Brown steam engine; photographic equipment; and a forge. He also established there a library of scientific publications. The equipment at Menlo Park, costing him $40,000, had few rivals even at the leading universities. About 20 of the skilled clockmakers and machinists whom he had employed at Newark were persuaded to settle, with their families, in his secluded "scientific village," to which in due time he also added a mathematical physicist.

The Menlo Park laboratory

Before moving to Menlo Park, Edison had told one of his scientific friends that his would be an "all-purpose research laboratory" and an "invention factory," and he promised that he would turn out a minor invention every ten days and a "big trick" every six months. He also proposed to "make inventions to order." Before long he had 40 different projects going at the same time and was applying for as many as 400 patents a year. Unlike his contemporaries, who worked alone in an attic, Edison depended in great measure on exact instruments and operated with an organized team of specialists whose tasks were allotted to them by their master—then still in his 20s. The German historian Werner Sombart later wrote that Edison was an example of a man "who made a business of invention." Menlo Park was, in fact, the first industrial-research laboratory, to be imitated 20 years later by giant corporations in both Europe and America.

The rough-hewn Edison used to declare that he was no scientist but a "commercial inventor" who worked for the "silver dollar." What he meant was that he consciously directed his studies to devices that could satisfy real needs and come into popular use. Indeed, it may be said that in applying himself to technology, he was fulfilling the ideals of democracy, for he centred his attention upon projects that would increase the convenience and pleasure of the multitude.

At Menlo Park, Edison, now in his 30s, experienced his finest hours. An astonishing number of new products bearing the mark of his originality were developed systematically from the inventor's growing reservoir of knowledge of various arts and industries. When Alexander Graham Bell won the race for the "speaking telegraph" in 1876, Edison was commissioned to improve the device and quickly added the carbon transmitter, markedly increasing the instrument's audibility. But he also had the gift of serendipity: when some unexpected phenomenon was observed, he did not hesitate to halt work in progress and turn off course in a promising new direction. This was how, in 1877, he achieved his most original invention, the phonograph, based on a fresh observation of the dynamics of sound. The phonograph was also an example of what has been called the wonderful simplicity of many Edison devices. Up to this time he had enjoyed a high repute among American technicians, but the phonograph made Edison internationally famous.

The electric light

Europe was well in advance of the United States in developing dynamos and also the electric arc light, which was the sensation of the Centennial Exposition at Philadelphia in 1876. In September 1878, after having viewed William Wallace's exhibition of a series of eight glaring 500-candlepower arc lights, Edison boldly announced that he would invent a safe, mild, and inexpensive electric light that would replace the gaslight in millions of homes; moreover, he would accomplish this by an entirely different method of current distribution from that used for arc lights.

Edison's past achievements commanded respect for his boastful prophecy. A syndicate of leading financiers, including Morgan and the Vanderbilts, advanced him $50,000 for research and development work on the incandescent electric light, which had been the despair of inventors for 50 years. The contracting parties went so far as to establish the Edison Electric Light Company, one of very few occasions when capital has backed an unachieved invention. There followed the prodigious dragnet search, lasting 14 months, for a suitable filament for his proposed lamp. The gifted physicist Nikola Tesla, who once

worked for Edison, called this his "needle-in-the-haystack" method of experimentation. The newspapers also reported that Edison proposed to connect his lights in a parallel circuit by subdividing the current, so that failure of one light bulb would not cause a whole circuit to fail. Some eminent scientists, including Sir William H. Preece, in a paper read before the Royal Society in London, predicted that such a circuit could never be feasible. On October 21, 1879, Edison demonstrated the carbon-filament lamp, supplied with current by his special high-voltage dynamos. The pilot light-and-power station at Menlo Park glowed with a circuit of 30 lamps, each of which could be turned on or off without affecting the rest. Three years later, the Pearl Street central power station in downtown New York City was completed, initiating the electrical illumination of the cities of the world.

One of the accidental discoveries Edison made in his laboratory in 1883 was beyond his comprehension, as it was beyond that of contemporary scientists. In one of his vacuum bulbs he observed a mysterious force that manifested itself by a flow of current between a hot and a cold electrode; later, this was shown to be due to thermionic emission of electrons from the hot electrode. He patented the bulb, describing what became known as the "Edison effect"; studied by other scientists a few years later, it became the basis of the electron tube and laid the foundation for the electronics industry.

After the death of his first wife in 1884, Edison married Mina Miller of Akron, Ohio, a woman half his age, but of genteel background and firm character. She took the heretofore incorrigible, tobacco-chewing eccentric in hand and made him presentable in polite society, which he had tended to avoid. Two sons and a daughter were born of the second marriage, one of the sons, Charles Edison, later becoming secretary of the navy and governor of New Jersey.

Later life. The second half of Edison's life was active and eventful but brought fewer big surprises. Most of the time he occupied himself with promoting and marketing his many products through companies he formed, such as the Edison General Electric Company, which manufactured his lamps and a variety of central station equipment. Eventually, Edison lost control of this major company which, in 1892, was merged into the General Electric Company.

The Edison laboratory

In 1887 he moved his workshop from Menlo Park to West Orange, New Jersey, where he built the Edison Laboratory (now a national monument), a facility ten times larger than the earlier one. In time it was surrounded with factories employing some 5,000 persons and producing a variety of new products, among them his improved phonograph using wax records, the mimeograph, fluoroscope, alkaline storage battery, dictating machine, and motion-picture cameras and projectors. The motion picture, with which he is so closely identified, was not Edison's invention, but with his rare power of going straight to a practical end he brought it to practical, operational success when many others had failed. The mass entertainment industry that he fathered yielded him rich profits, which, by the turn of the century, he badly needed, for he had overextended himself in the development of a magnetic ore-separating process for low-grade iron deposits. When, after five years, he decided to write off this venture, he found that he had lost the entire fortune of $4,000,000 earned by his electric-light system. His reaction was typical: "Well, it's all gone, but we had a hell of a good time spending it!"

In one important episode his limited knowledge of theoretical science told heavily against him. Toward 1890 the more advanced electrical scientists were introducing alternating current at high voltages for long-distance transmission. Edison, having chosen direct current, launched a press campaign to alarm the public about the alleged hazards of the high-voltage system. As a consequence, he lost the contract for the great Niagara Falls hydroelectric site to the rival entrepreneur George Westinghouse.

Other ventures, such as in engineering a more efficient form of cement kiln and in the poured-concrete structure designed for low-cost housing, were more fruitful. During World War I, the aged inventor headed the Naval Consulting Board and directed research in torpedo mechanisms and antisubmarine devices. It was largely owing to his agitation that Congress, in 1920, established the Naval Research Laboratory, the first institution for military research.

Achievements. Edison's career, the fulfillment of the American dream of rags-to-riches, through hard work and native intelligence, made him a folk hero to his fellow countrymen. He was attractive in appearance, although negligent in dress; in temperament he was an uninhibited egotist, at once a tyrant to his employees and their most entertaining companion, so that there was never a dull moment with him. His shafts at the expense of the "long-haired" fraternity of theorists sometimes led formally trained scientists to deprecate him as anti-intellectual. Yet he employed as his aides, at various times, a number of eminent mathematical physicists, such as Tesla and A.E. Kennelly.

The full record of his 3,400 notebooks shows that he had a wide understanding of the principles of science, as known in his time, and a tremendous faith in the method of scientific experimentation. The mathematician Norbert Wiener characterized Edison as a transitional figure in late 19th-century science, who arrived on the scene after the crude mechanical inventors had done their part, when systematic experiment and research were about to be undertaken by organized groups of specialists. The little industrial-research laboratory at Menlo Park may, therefore, be seen as his greatest invention. Edison was in the forefront of those who pressed on, for better or worse, toward the Age of the Machine. Though he grew up in an era of great acquisitors, his own instincts were predominantly creative, as shown by the manner in which he flung away fortunes, only in order "to invent more," as he said. He continued to work even in his 80s, devoting his last years to efforts to develop rubber latex from domestic weeds. He died in West Orange, New Jersey, on October 18, 1931.

BIBLIOGRAPHY. MATTHEW JOSEPHSON, *Edison: A Biography* (1959), is based on the correspondence and laboratory notebooks of Edison in the Edison Laboratory archives, West Orange, New Jersey. F.L. DYER, T.C. MARTIN, and W.H. MEADOWCROFT, *Edison: His Life and Inventions*, 2 vol. (1929), an authorized biography, done in part with Edison's collaboration, records his career to the 1920s. W.K. and L. DICKSON, *The Life and Inventions of Thomas Alva Edison* (1894), contains a detailed account of early motion picture experiments by a former laboratory assistant. THOMAS A. EDISON, *Diary and Sundry Observations*, ed. by D.D. RUNES (1948), is autobiographical, with fragments and random notes. See also G.P. LATHROP, "Talks with Edison," *Harper's Magazine*, 80: 425–435 (1890); FRANCIS JEHL, *Menlo Park Reminiscences* (1936); and THOMAS A. EDISON, "The Phonograph and Its Future," *North American Review*, 126:527–536 (1878).

(M.Jo.)

Education, Economics of

The economics of education has a long ancestry. Distinguished economists, from the time of, and including the founders of the science, have alluded to the importance of the growth of human knowledge through formal educational processes and to the significance that this plays in the development of the economy. Only in comparatively recent years, however, has a body of specialized knowledge worthy of note been developed.

Educational investment and return. *Economic input.* In recent times education has come to take more and more years of an individual's life span. In North America and western Europe it is not uncommon for a child to enter a nursery school or kindergarten at the age of two or three and to emerge from the educational process only after he has completed studies for a higher degree, as much as 20 years later. Frequently, during the years of his working life, he will attend refresher courses and other postdoctoral courses in order to bring his knowledge up to date. In the developing countries, where once only a small proportion of the community entered into formal education, universal primary education is a high-

priority matter. The latter half of the 20th century has truly been called the period of universal education.

Education as a percentage of GNP

In developed economies, normally between 5 and 7 percent of the Gross National Product may be spent on formal education. As will be seen later, this figure is exclusive of significant additional costs, notably the earnings forgone by students when they are undergoing education, and education that is not organized—*i.e.*, not part of the formal apparatus of schools and colleges. Some authorities, indeed, have estimated that the proportion of a developed country's Gross National Product devoted to education in the broadest sense might be as much as 10 percent. This figure compares with figures of 1 or 2 percent for the latter part of the 19th century in a significant number of advanced industrial countries and with figures of between 2 and 3 percent in many of the currently developing nations. What is significant about the percentage is that it is increasing more rapidly than many other parts of the national income. This large proportion of the national income is, in most countries, channelled through the public budget at central and local levels; in many countries the proportion of the public budget devoted to education may be as high as one-fifth. In real terms education is an industry relying more on labour than on capital, and it thus uses a high proportion of the qualified manpower available.

One major part of the study of the economics of education is the careful calculation of the amount of resources devoted to education.

Economic output. Corresponding to these inputs are the outputs, or consequences, of the educational system. Here, a major problem is posed by the evaluation of the outcomes of education. In part, these outcomes are intangible; in other respects—in terms of examination results or of the differentials in lifetime earnings that may be attributed to education—they are fully measurable. Thus, a student is supposed to forgo the earnings that he might have made between the ages of 16 and 23 in order to earn a higher salary than he otherwise would have done in his later years. Part of these additional earnings may, of course, be due to innate ability, race status, social class, parental background, or other factors besides education. Techniques have now been developed for differentiating between these elements so as to permit the allocation of an appropriate part of the outcome to education.

Individual and social benefits of education

For the individual it is possible that the major part of the economic outcome of education can be measured by extra income, but it does not follow that these are the only benefits of an economic nature that can be said to accrue because of education. If the model society were one in which all incomes were equal, it would still not follow that education had no economic benefit, because the extra skills and abilities resulting from education would still yield outcomes that were beneficial to the community. Thus there is a distinction between the benefits to the individual and those to the community as a whole. The difference may be positive or negative.

Investment and consumption. It is customary to study the consequences of education and the motivation that leads people to undergo education in terms of either investment or consumption. That is to say, people may undertake education purely for its own sake, as when a middle-aged person undertakes the study of a foreign language for recreational purposes. If he decides to devote a portion of his income to this study, the decision to do so is, in principle, no different from a decision to spend more on food or on clothing or on any other item of personal consumption. Part of the demand for education represents collective family decisions to allocate resources in this way. For some people, learning in itself is something that gives pleasure and a sense of achievement.

But in another respect, education can be viewed as an investment in the future. A society that deliberately devoted a high proportion of its resources to improving the skills and abilities of its young people would expect to see a change in its economic and social performance in the years ahead. The same, of course, is true of the family and of the individual.

In this respect, therefore, expenditure on education may be understood as a forgoing of present enjoyment in order to increase the future flow of benefits. The analogy is with the process of investment in physical capital, when current consumption is forgone in order to increase the stock of capital goods that will, in the future, yield a flow of consumption benefits. It is possible, therefore, to examine the process of educational expenditure over a period of time as a process of investment in the future of an individual or of a community. Part of this investment will be merely the replacement of an older generation and its skills by a younger generation and its skills, but another part of it will be genuinely new investment—that is, a net addition to the stock of "human capital." There are, of course, grave dangers in pressing the analogy too closely but within limits it has some merit.

Factors of production. In classical economic theory, human labour added to land produced the final product. In Neoclassical economic theory, land and labour had to be differentiated from physical capital, that is the outfit of plant, equipment, and buildings used in the productive process. It was held that the contribution of each factor of production—land, labour, or capital—could be distinguished from the contribution of others. In recent years the contribution of labour has been further divided into the contribution of labour pure and simple and the contribution of acquired human ingenuity and skill. The latter factor, termed human capital, is widely regarded as one of the most important factors in the process of production.

Human capital

This formulation has significance for a number of aspects of economic theory, notably for the theory of economic growth. It used to be thought that the accumulation of physical capital was the key to economic development. More recent studies have suggested that the contribution of human capital is at least as significant and that a considerable amount of the growth rate can be explained as the yield on human capital. It follows, therefore, that if the accumulation of human capital can be accelerated, the rate of growth will be increased.

Work on the economics of human capital has been extended to cover not only the effects of formal education but of on-the-job training and other methods of skill acquisition, as well as of scientific research and technological development. The development of the new concept of human capital has helped to explain the residual gap that was left after the contributions of land, labour, and capital had been accounted for, a gap which had hitherto remained a mystery. The policy consequences of this change in economic theory are important because they point to additional expenditures on education and on-the-job training and on research and development, rather than on physical capital. In many countries this offers a relatively more simple solution to the problem of growth than the more traditional solutions.

DEMAND AND SUPPLY OF EDUCATIONAL SERVICES

The increasing demand for educational services. Since about 1950 there has been a rapid acceleration in the rate of growth of demand for educational services at all levels, the most notable at the level of higher education in college and university but affecting primary and secondary education, adult education, and other out-of-school educational activities of all kinds. Projections of future demands suggest that while there may be, in some countries, a slowing up in the rate of increase, there is no reason to suppose that the demand will, in any reasonable period of time, reach a plateau.

Population growth. A major cause of the rise in the demand for education is the increase in the size of the child population. In the world as a whole the rise in the number of children and young people has been due first to a marked decline in infantile mortality rates and second to a tendency in a number of countries for the average size of the family to increase. Even in those parts of the world in which the average size of family has

fallen in recent years, the improvement in child health has been such that the size of the child population has nevertheless tended to increase. Most recently the size of the child population in northwestern Europe and North America has tended to stabilize or, even, to decline slightly; but rapid increases in child population continue to be observed throughout Asia, Africa, and Latin America. It is probable that the rate of increase will tend to slow down if the standard of living rises. One aspect of the rise in the standard of living is an increased spread of education. The spread of education tends to raise the cost of having children for each family—in part by diminishing the supply of child labour—and thus gives an additional incentive for family limitation. It has been suggested that this was a major cause of the fall in the size of the family noted in northwestern Europe and North America in the last quarter of the 19th century.

The lengthening of school life

Social demand for education. The rise in the child population has been a major cause of the increased demand for educational services. But added to this in the developed countries is probably the most important single cause of the increasing demand for education: the increase in the length of school life. At a certain level of national income per person, education becomes widespread, and universal, compulsory primary education is introduced. In the developing countries, since about 1950, the process of decolonialization and spread of the concept of economic and social development have led to a demand for universal primary education in all countries, and this demand is in process of realization.

Preceding this process of universal primary education, however, and superimposed upon it has been a demand for more education for children from favoured social groups. The average length of school life has increased, decade by decade, in almost every country. Those parents who have themselves received some education are often those who demand most education for their children. Thus, in countries in which there is wide educational opportunity, the spread of education has been accompanied by a rapid increase in educational provision for those groups—often the most affluent—with the highest degree of experience of education. This feedback or spiralling effect of education means that education has become a self-perpetuating process.

Economic growth and technological development. To some extent the growth in population, and to a great degree the increase in the demand for education arising from growing prosperity, reflects the process of economic growth and technological development. As technology changes and economic growth occurs, the growing complexity of the economic process requires a growing and diverse range of skills. Thus, the production necessary for a highly diverse and complex society based on an industrial and scientific complex requires a broad range of skills, not only in the labour force but among consumers in the ordinary daily process of life. All of this rests on a complex process of instruction and socialization, which implies a highly articulated, complex, and diverse educational structure.

The process of technological and economic change manifests itself first in the growing experience of the working force and of consumers. This is an almost automatic and almost unconscious process in many instances, and over a long period a substantial amount of new knowledge about technical processes is thus acquired by workers and consumers. Then there are formally arranged training procedures in industry and agriculture, of which the most common is the apprenticeship in which young people work alongside master craftsmen. There is also technical education outside industry. Underlying technical education is, of course, the general process of education.

The links between the different kinds of education and training are loose, but the development of a whole series of formal skills, at the more advanced levels, involves a widespread educational background, and the countries with a deep educational background are able to adjust quickly to the demand for new skills and techniques.

The uncertain supply of educational services. *Demographic imbalance.* Educational services respond only comparatively slowly to new demands. It is common for acute shortages of educational facilities to exist in the face of mounting demand. The causes of this imbalance are manifold. Because of different sized age groups, a large child population may be taught by people drawn from smaller age groups within the adult population. In industrialized countries particularly, and in areas in which there is significant urban development, the population may be highly mobile, causing regional imbalances. Further, rapid shifts in the demand for different specialties within education may cause a surplus of teachers in one subject and a shortage of teachers in another subject; and, finally, education is so closely connected with the religious, social, and ethnic background of the community that it is perfectly possible for acute shortages of teachers to arise in one ethnic group, while there are surpluses in another.

Since teachers form so high a proportion of the total stock of educated people, the increase in the demand for educated people for other occupations must tend to reduce the supply of potential teachers.

Competition from other occupations. There tends to be, in the more advanced countries, a persistent trend toward higher salaries in the educational system. Traditionally, because the provision of teacher training facilities has often been accompanied by fairly adequate scholarship programs and free or reduced tuition, the teaching profession has been an avenue of academic and social advancement for clever children from poorer homes. But as the general level of affluence rises in the developed countries, teaching becomes less an avenue of advancement than previously it had been.

Alienation from teaching. As a result of the development of the educational system, something like 20 or 25 percent of the population of advanced countries will be employed full time in the educational system, either as teachers or as taught. In such circumstances teaching may become less a prestige occupation than a mass-production exercise. Because of this the more ambitious may often be deterred from entering it.

Uneven development. Educational systems develop almost inevitably at an uneven rate, so that although in an ideal system there ought to be a balance between the first, second, and third levels and appropriate teachers and buildings should be supplied for each, in fact, because of fluctuations in population structure and social demand and because of varying pressures on the educational system to provide new skills, the levels of education tend to become out of balance. The supply of buildings, teachers, and other facilities may not necessarily be coordinated with each other, so that in some places there may be empty buildings with no teachers and in other places, teachers with no place to teach. The normal development of education, therefore, is not one of smooth equilibrium in demand and supply but one of fairly chronic imbalances.

ECONOMIC PLANNING OF EDUCATION

Various approaches to the determination of education expenditure. There are various approaches for determining the desirable level of expenditure on education. If an individual family buys its own education for its own children, then it might seem that the solution to the desirable level of educational expenditure is a simple one; as in the case of any consumer commodity, the choice is left to the consumer and the sum total of all consumers' expenditures equals the demand for the commodity. In the case of education considered as an investment, the family would make the same choices it would if it were buying a house or a long-term consumer durable, it would appropriately discount the future flow of returns. But for a number of reasons this is not a wholly satisfactory answer. Families might make decisions that, between them, were too high or too low from the point of view of the nation as a whole, determined by its legislature or other appropriate institution. A number of fam-

ilies might be unable to buy either enough or any education. The desirability of the market as a mechanism for allocating resources depends ultimately upon the doctrine of the sovereignty of the individual consumer and in the field of education raises the question of whether the parent should be sovereign on behalf of his children.

The application of market criteria

In any event, the greater part of educational expenditure is undertaken by public authorities or firms on behalf of their citizens or their employees. The question thus arises: What principles, if any, should determine the expenditure? The first that has been suggested is that it should be based on the analogy of the market—that is to say, each family should be asked how much education it wishes to consume, and a general idea should thus be given to the policy-making institutions. Further, it should be arranged that the expenditure should be based on consumer decisions (on behalf of children) resting on consumer preference and, especially for older students, on ordinary investment criteria. These criteria depend on a correlation between educational attainment and subsequent earnings. The sum total of earnings over a lifetime has to be added up and discounted to a present value. Since very few data are available relating any individual's earnings to his own income over a period that may be as long as 70 years, present-day educational experience is taken and correlated with present-day earnings profiles. Thus there has to be some degree of approximation in the statistics.

According to pure theory, differences in expected rates of returns should be reflected in differences in the choices that people make in careers and in courses leading to careers. There is some evidence that this is the case, though whether or not the explanation that is given for these choices is the correct one is still a matter for debate and study.

Optimum educational opportunity. A contrasting approach is based on the concept of optimum educational opportunity. Sociological evidence shows that for children with the same measured ability, differences in educational attainment are dramatic. Urban children tend to perform better than children in rural environments. There are significant ethnic differences such as those revealed between black and white children in the United States, and there are major differences between those from families in the upper professional classes and those from the lower manual working classes. According to the view that seeks the attainment of optimum educational performance for each child, if all social disadvantages were removed from those children who appear to be relatively handicapped socially, the level of attainment of the children of the nation as a whole would approximate the level of attainment of those children from the most favoured social classes. This, of course, would lead to dramatic increases in the numbers seeking full secondary and higher education. There is strong evidence that as generation succeeds generation, other social groups tend to emulate those social groups with the most favourable educational experience. They are, to some degree, chasing a receding target; even the highest and most favoured social groups tend to improve their educational performance through time, so that over the 30-year period between 1938 and 1968, the gap between the educational performance of the most favoured and the least favoured social groups in England and Wales did not diminish, despite a striking improvement in the position of the least favoured social group.

Problems involved in equalizing educational opportunity

The assumption underlying this criterion of public expenditure may be both questionable and expensive: expensive for the obvious reason that the number of children who ultimately might be expected to proceed into higher education would be large, reaching eventually well over half the age group, if present-day educational attainments by the most favoured social groups were to be attained, and questionable because it is not necessarily inevitable that the educational experience of the most favoured social group represents the optimum experience possible for all young people. In the first place, from an economic investment point of view, there may not be occupations available in sufficient numbers for all the people who, under this hypothesis, would be educated to expect to work at those jobs. In the second place, it may well be, too, that educational aims of people from different social groups may legitimately be radically different from each other.

Underlying this concept there is an important controversy in sociology and genetics about the relative importance of nature and nurture. Some sociologists have taken the view that the distribution of innate ability, if such a concept has any meaning, would be found to be the same in all social groups if social handicaps were removed. There is controversy about the proposition, and some sociologists and geneticists are willing to argue that the greater part of ability is genetically controlled. If they are right, there will not be many people in the least advantaged social group able to undertake higher education; for presumably they have inherited fewer abilities than have the upper classes, who are better prepared for higher education.

Manpower planning

Maximizing economic growth. The preceding approaches to the economic planning of education depend ultimately on individual choice. Manpower planning is an approach based on the need to promote economic growth and on the view that economic growth depends, probably in substantial part, on the availability of skilled manpower. Thus it is argued that, since from the point of view of the nation as a whole, the optimum availability of skilled manpower is what matters, the main purpose of state intervention in the educational and training system should be to achieve the optimum supply of skilled manpower. The required total of skilled manpower may be derived from a number of exercises. It is necessary, for instance, (1) to establish the educational attainments of the existing stock of skilled manpower, (2) to determine the relations between the possible growth of the national economy and the requirements for skilled manpower, (3) to estimate changing technological relations in the future, (4) to determine how the future demand for skilled manpower may be affected by future mortality and retirement rates, and so forth and so on. The resulting gross demand for skilled manpower expressed in annual rates represents the desirable output of the education system from the point of view of manpower planning. The utilization of this technique requires many approximations of what are in fact only loose relationships, both between educational attainment and subsequent occupational classification and between the level of output and the level of skill input. In addition, the long-range projection of the national income and its various sectors is both complex and uncertain.

Even if manpower projections were accurate, however, and could be neatly related to desirable educational outputs, there would still remain the problem of relating this hypothetical desirable output to individual preferences for specific education. This could be done by propaganda, by altering income and wage differentials, or by direction; all of these techniques have been used in the U.S.S.R., which has partly based its educational program on manpower requirements. In France, where manpower planning has been carried on with a considerable degree of sophistication, the educational system bears little relation to the manpower plans.

International comparisons. Another criterion used for determining desirable levels of educational expenditure is international comparison. Comparisons are based on data regarding educational expenditures in different countries representing similar stages of economic growth. The hypothesis is that countries at similar levels of economic development will tend to spend a similar amount on education and that their educational patterns will resemble each other in certain broad respects. The central problem of this method is that, although countries with roughly similar levels of national per capita income do tend to have *similar consumption* patterns (as do the countries of western Europe and North America, where mutual cultural influences are strong), these countries may very well have considerably *different economic* structures. It

follows, therefore, that insofar as the educational system is related to the economic structure, educational expenditures will differ from country to country. Furthermore, there is no necessary relation between the educational structure and the level of national income; the relation is far more closely affected by different historical circumstances than by the social and economic structures, which differ widely from country to country.

International comparisons of national income are notoriously complex and difficult to make and rest upon questionable assumptions about the comparability of welfare. In addition, despite attempts to achieve an agreed-upon international definition, educational expenditures are in fact defined in different countries in part according to their own educational practices, and therefore the range of expenditure may well reflect more the differences in accounting procedures than differences in actual outlays.

Other approaches. Some of the other approaches to determining a desirable level of educational spending represent a compromise between the different approaches so far discussed. Some educational expenditure is mainly a matter for individual decision; some is related to a policy of raising the level of those with the lowest educational attainments; some, particularly industrial training and technical education, is related to the requirements of skilled manpower; and some is based on international comparisons.

Cost–benefit analysis

There are new approaches to public finance that suggest that there may be further possibilities of calculating optimum levels of educational expenditure. Cost-benefit analysis is an instance. It is suggested that there should be a quantified relation between expected total benefits (including private and social benefits) arising from a given level of public expenditure and the cost of that public expenditure. It is generally agreed, however, that cost–benefit analysis can fully be used only within a sector, such as the educational sector, and that it is not valid as a method of comparison between sectors, such as between formal education and other kinds of welfare programs. Nevertheless, it may be some guide to relative expenditures on different levels of education.

Decisions about educational spending are usually marginal decisions. The greater part of educational expenditure in any one year represents previous commitments and inescapable changes in costs that the system has already agreed to incur—for example, wage increases or rises in the cost of building construction. The area of choice is represented by only a small margin in any one year, though over a period the margin may be considerable. It follows, therefore, that since the greater part of educational expenditure represents a pattern that is the cumulative result of many small decisions, it is difficult to change the pattern radically in any short period of time.

In addition, the greater part of educational policy making concentrates on changes in particular areas of the educational system (for example, primary education or higher education); and although changes at one level inescapably have effects ultimately at all other levels, many of these relations have not been carefully assessed and are only generally considered by policy makers.

Various approaches to determining fees and revenue collection: "consumption" and "investment" issues. Much work in the economics of education deals specifically with the issues of raising revenue for the educational system and with the manner in which expenditure on education is undertaken. The first principle of revenue collection is concerned with the issue of who should bear the cost. One of the principles of taxation is that the cost of a service should be borne by those who ultimately benefit from it. If students ultimately benefit from education, it seems logical that they and their families should pay for it. On the other hand, part of educational expenditure is undertaken not only because it is in the interests of students but also because it is in the interests of the community, either locally or nationally. Ultimately, enterprises both public and private benefit from the availability of skilled and educated manpower. Nor must it be forgotten that among those concerned with education are charitable and, especially, religious foundations: indeed, the origin of education in modern Western culture has largely been due to religious bodies.

The difficulty of trying to recoup the cost of education from students is that they themselves rarely have incomes sufficient to meet the cost of their education, though there is a sense in which they already do bear a major part of educational costs because, while they are studying, they are unable to earn significant amounts. The cost of studying is, in some degree, represented by the earnings that they have forgone by becoming students. If this item were added to the total cost of education, as it sometimes is, it would increase the cost substantially. The families of the students, particularly the more affluent families, sometimes bear the full cost of tuition. For the majority of the population, however, the costs of tuition are mainly or wholly borne by public bodies. It would, in principle, be possible to have this payment through taxation replaced by a system of family loans that would be later repaid out of the enhanced incomes that successful students expect to earn. Loans raise major administrative problems as well as questions of social value, however, because they may discourage potential students, and many people hold the view that, in principle, education should be free. It is difficult, if not impossible, to make education compulsory without at the same time making it free. Loan schemes have generally been attached to those optional parts of education (such as technical and university education) in which the decision to attend is, to a greater degree, a free choice. It has been proposed that the full cost of higher education should be recouped from fees charged to students and that loans should be paid back by a given percentage of earnings over the lifetime—that is to say, in effect, an addition to the income tax for those people who have borrowed money. Thus those groups, such as women who marry and are no longer employed or people who enter low-income professions, such as priests or social workers, would not be faced by large repayments, whereas successful businessmen and professional workers would pay back substantially more than they had borrowed. Merely to state the case is to indicate the political difficulty inherent in it.

Tuitions, taxations, and loans

In some countries with a unitary system of administration, such as France, the cost of education is a national responsibility. In others, such as the United States, there is a division between local and central government with education frequently a local responsibility, with the costs borne locally. This raises serious problems because the financial resources of state and local government are often fairly narrow and often based upon the property tax, which has an inelastic yield and does not rise as fast as national income and tends to be politically unpopular. As society becomes increasingly mobile, education itself is a major factor in geographical mobility, and people pay taxes to communities that have not borne the expense of their education. Increasingly, therefore, education has become a matter for national finance. It has already been said that business represents a major beneficiary of the educational system, and businesses make considerable contributions both to central and local taxation. In some countries, such as France and the United Kingdom, there is additional taxation for broadly educational purposes, represented in the United Kingdom by the levy imposed upon most employers to finance the cost of industrial training. It may well be that this form of taxation will become more popular, particularly as a means of financing technical education.

National and local financing

Churches in many countries have been a major source of educational finance, both through the unpaid labour of teachers in religious orders and through direct expenditures from their own charitable endowments. But as modern education has become more complex and expensive, this source of finance has diminished. Private endowments from foundations have been significant in stimulating educational innovation in particular areas of

an education system. There are also long-standing charitable endowments to individual schools and colleges that are used to subsidize the fees of students.

What form should revenue collection take? Most revenue is collected from taxes, though there are also sources of income from fees and from the yield on capital endowments; and to complicate the picture there are also grants-in-aid, loans, and subsidies among various parts of the educational system. The detailed arrangements vary from country to country and from institution to institution. Much of education is financed by property taxes, particularly at the local level; and therefore because primary education in many countries is mainly a local responsibility, primary education rests in many countries mainly on property tax. Insofar as the central government is responsible for education, the general fund of taxation (which includes most forms of taxation) is the main source of income.

THE BALANCE BETWEEN DIFFERENT FORMS OF EDUCATION

Developed countries. *The consumer society.* There is an educational pyramid ranging from education at the preprimary level up to education at the postdoctoral level. There is also a horizontal division between different types of education—notably the differences between a professionalized and a general education or between different forms of vocational education.

Formal education in its origin was in one sense vocational, since it was concerned with the production of priests and other people who could perform high clerical functions. Training in medieval society largely took the form of apprenticeship; during the 18th and 19th centuries in western Europe and North America the upper class concerned itself primarily with general preparation for entering the high culture, and the other classes concerned themselves with apprenticeship. The middle class took an increasingly vocational education in the modern sense of the word, preparing for specific occupations. During the 19th century there was a spread of elementary education to include children from the poorer classes, and a growing development of technical education in the modern sense, resting on the emerging concepts of engineering and science.

During the 20th century, as the level of national income has risen and as the consumer society has become widespread, the demand has increasingly been for a more general education for all children, such as was formerly the prerogative of the upper and middle class and superimposed upon this has been growing technical specialization enabling more people to enter high-level occupations. During the development of the modern consumer society, the notion of a cultured person has radically changed. There has thus arisen in the public mind a growing confusion about the ends of education, at the same time that the demand for it has become an almost unquenchable thirst. The result is that a large and growing portion of the population is spending a longer part of its life in education, not only during childhood and youth but also during maturity, and this experience of education is in itself radically altering the nature of modern culture. The development of powerful student groups has been noted by many observers as a prevalent feature of modern society, and the concept of man as the perpetual student is not as farfetched as it might once have seemed.

The technological society. The prevalence of a technological society—one that has led to the superabundance of commodities and services available to many in modern society—has required an increasing number of specialists of all kinds, particularly in the fields of science, engineering, and associated areas. Such age-old professions as medicine have also been radically transformed by the growth of modern scientific medicine and by the technological applications associated with it. At the same time, there have arisen entirely new professions, such as that of computer specialists, which are increasingly important in determining the nature and rate of development of modern society. Thus, there is a conflict between the growing concern about general cultural levels and the growing specialization of people's working lives. The long period of preparation that is necessary for people to enter a specialized working life is also the period in which they assimilate cultural values. The problem facing education is that as technology and knowledge change, the content of education also changes; and thus there is a contrast between the desirability of continuity implicit in a mature culture—the need to hand on traditions from one generation to another—which is one of the primary purposes of education, and the need for constant change to keep people up-to-date with rapid technological change. This striking contrast represents a major source of tension in the curriculum and in techniques of teaching. There are other subsidiary contrasts that are equally important; increasingly, education is concerned with freedom of expression by the student, but the close and intimate relation between teacher and taught often contrasts with the authoritarian and hierarchic structure of the work that is undertaken in life.

In the developed countries, education has become a major source of economic development. It is regarded as one of the primary sources of skills and of new ideas; it takes an increasingly important part of the national income; it becomes increasingly a focus of national and individual and family concern. At the same time it is a source of revolutionary discontent and of growing concern with man's condition. It follows, therefore, that the community's attitude toward education is shot through with ambivalence—on the one hand, recognition of its economic importance to the individual and to society and, on the other hand, resentment of many of its aspects of change.

Developing countries. *The aspiring society.* In developing countries many of the same tensions are seen. In all of them without exception—but especially among those that have moved from a colonial status to political independence—the demand for universal literacy has been widespread. The right to free primary education is not only written into the Universal Declaration of Human Rights but is also often embodied in the basic constitutional documents of all developing countries. The result has been an explosion in the demand for universal literacy and particularly for universal primary education. Many countries are therefore devoting substantial amounts of their public resources to the spread of universal primary education and to adult literacy programs. The adult literacy programs have generally been a failure despite repeated experiments to raise the efficacy of various methods of teaching adults to read and write. Generally speaking, attempts to achieve universal adult literacy are a failure unless there is a direct correspondence between the achievement of literacy and its subsequent use in daily life. In many cases the achievement of literacy to a standard enabling a person to pass formal tests has been followed by a lapse into illiteracy. The concept of "functional literacy," in which a person is defined as literate if he is able to fulfill a task for which literacy is required, has achieved widespread popularity.

Literacy programs

The basic attack on illiteracy must ultimately come from the spread of universal primary education. Many countries are now within striking distance of achieving this, though the primary educational system in most countries, judged by the standards of the developed countries, is often inefficient. It is marked by a high level of truancy, old-fashioned teaching methods, and widespread problems of untrained teachers and inadequate buildings. Although the manpower planners suggest that the developing countries should be putting their first efforts into training highly skilled and semiskilled specialists, there is no doubt that the demand for primary education is overwhelming and cannot be resisted.

Indeed, instead of a desire for skills for the purposes of economic development, the demand for secondary and higher education in the developing nations often takes the form of a demand for prestigious education, preparing children of the elite for elite positions. Many highly educated people in the developing nations consciously reject manual labour and technical and scientific skills; the

Elitist versus technical education

emphasis in their education is upon the humanities and especially upon law. It is indeed true that education in the humanities and law is often required to develop a politically conscious elite—one that is able to take over the administration of a country from the previous rulers. But it is also true that the major public services—such as the armed forces, the medical services, together with the whole complex of skills required to service the modern sector of the economy—depend heavily upon expatriates trained in science and technology. The replacement of these expatriates by domestically trained specialists often leads to a fall in the efficiency of services.

In any event, the process of decolonialization is often accompanied by an urgent need to produce highly skilled technical manpower. But much of this technical manpower has to be trained overseas, and the process of training overseas encourages the resulting highly skilled specialists to remain overseas or re-emigrate. Many programs for overseas training have, in fact, accelerated the degree of migration of skill from the developing to the developed countries.

A developing society may therefore almost be defined as a country that is deficient in the technical skills required not only to service it at its present level of development but also to accelerate its development. It follows, therefore, that from the point of view of the economy, one of the urgent requirements of the developing countries is the development of cadres of technically skilled people. The International Bank for Reconstruction and Development (the World Bank) has attached priority in its development plans to the provision of technical specialists. In the developing nations, despite the need to develop cultural autonomy and the priority therefore given to cultural objectives in education, there is no doubt that the highest priority must be given to the development of technical specialists who can lead the country into self-sufficiency in technical, scientific, and other specialist skills. But the emphasis on this aspect of education is dangerous if it leads to an imbalance in the education service, because the development of a high-level training program presupposes a solid basis of secondary and primary education. It is for this reason that there is a need to develop plans that coordinate the different levels of education and the different parts of education, so that the product of one level becomes the proper input of a next higher level. Further, it is necessary to see that these plans are carefully coordinated with other programs of economic and social development, so that the educational system is not out of kilter with the other programs in the development process.

BIBLIOGRAPHY. M. BLAUG, *Economics of Education: A Selected Annotated Bibliography* (1966, periodically revised), a full listing, reasonably annotated, of books and articles in the field; and (ed.), *Economics of Education*, 2 vol. (1969), a collection of readings with emphasis on solutions to the problems of education by applying the market criteria; *The Economics of Education* (1970), a textbook that complements the volumes of readings edited by M. BLAUG; M. LEITE *et al.*, *The Economics of Educational Costing*, 4 vol. (1969–70), represents a complete survey of work in the economics of education; E.A.G. ROBINSON and J.E. VAIZEY (eds.), *The Economics of Education* (1966), a collection of papers read at a conference of the International Economic Association; T.W. SCHULTZ, *Economic Value of Education* (1963), an essay representing the Neoclassical tradition of human capital theory that has come to be associated with the University of Chicago; J.E. VAIZEY, *The Economics of Education*, 2nd ed. (1962), a survey that attempts to cover most of the major topics in a way designed both for elementary and advanced students.

(J.V.)

Education, History of

Education can be viewed as the transmission of the values and accumulated knowledge of a society. In this sense, it is equivalent to what social scientists term socialization or enculturation. A child—whether conceived among the African Bushmen, the Renaissance Florentines, or the middle classes of Manhattan—is born without culture. Education is designed to guide him in learning a culture, molding his behaviour in the ways of adulthood, and guiding him toward his eventual role in society. In the most primitive cultures, there is often little formal learning, little of what one would ordinarily call school or teachers; for, frequently, the entire environment and all activities are school, and many or all adults are teachers. As societies grow more complex, however, the quantity of knowledge to be passed on from one generation to the next becomes more than any one person can know; and hence there must evolve more selective and efficient means of cultural transmission. The outcome is formal education—the school and the specialist called the teacher.

As society becomes ever more complex and schools become ever more institutionalized, educational experience becomes less directly related to daily life, less a matter of showing and learning in the context of the workaday world, and more abstracted from practice, more a matter of distilling, telling, and learning things out of context. This concentration of learning in a formal atmosphere allows the child to learn far more of his culture than he could by merely observing and imitating. As society gradually attaches more and more importance to education, it also tries to formulate the overall objectives, content, organization, and strategies of education. Literature becomes laden with advice on rearing the younger generation. In a word, there develop philosophies of education.

This article deals with the evolution of the formal teaching of knowledge and skills in all parts of the world and with the philosophies that have inspired it. The article is divided into the following sections:

I. Education in primitive and early civilized cultures
  Prehistoric and primitive cultures
  Education in the earliest civilizations
    The Old World civilizations of Egypt, Mesopotamia, and North China
    The New World civilizations of the Maya, Aztecs, and Incas
II. Education in classical cultures
  Ancient India
    The Hindu tradition
    The introduction of Buddhist influences
    Classical India
    Indian influences on Asia
  Ancient China
    The Chou period
    The Ch'in–Han period
  Ancient Hebrews
  Ancient Greeks
    Origins
    Sparta
    Athens
    The Hellenistic age
  Ancient Romans
    Early Roman education
    Roman adoption of Hellenistic education
    Education in the later Roman Empire
III. Education in Persian, Byzantine, early Russian, and Islāmic civilizations
  Ancient Persia
  The Byzantine Empire
    Influences on Byzantine education and culture
    Primary and secondary education
    Higher education
  Early Russian education: Kiev and Muscovy
    Influences on Russian education and culture
    Education under Kievan federation
    Muscovy educational development to 1725
  The Islāmic Era
    Influences on Muslim education and culture
    Aims and purposes of Muslim education
    Organization of education
    Major periods of Muslim education and learning
    Influence of Islāmic learning on medieval and modern Europe
IV. The European Middle Ages
  The background of early Christian education
    From the beginnings to the 4th century
    5th to 8th centuries
    The Irish and English revivals
  The Carolingian renaissance and its aftermath
    The cultural revival under Charlemagne and his successors
    Influences of the Carolingian renaissance abroad

Education of the laity in the 9th and 10th centuries
The medieval renaissance
Changes in the early schools
The development of the universities
Lay education and the lower schools
V. Education in Asian civilizations: *c.* 700 to the eve of Western influence
India
The foundations of Muslim education
The Mughal period
China
The T'ang dynasty (AD 618–907)
The Sung (960–1279)
The Mongol period (1279–1368)
The Ming period (1368–1644)
The Manchu period (1644–1911)
Japan
The ancient period to the 12th century
The feudal period (1192–1867)
VI. European Renaissance and Reformation
The channels of development in Renaissance education
The Arabic influence
The secular influence
The Humanistic tradition in Italy
Early influences
Humanistic boarding schools
Nonscholastic traditions
The Humanistic tradition of northern and western Europe
Dutch Humanism
Juan Luis Vives
The early English Humanists
Education in the Reformation and Counter-Reformation
Martin Luther and the German Reformation
The English Reformation
The French Reformation
The Calvinist Reformation
The Roman Catholic Counter-Reformation
The legacy of the Reformation
VII. European education in the 17th and 18th centuries
The social and historical setting
The new scientism and rationalism
The Protestant demand for universal elementary education
Education in 17th-century Europe
Central European theories and practices
French theories and practices
English theories and practices
The academies
Education in 18th-century Europe
General condition of education during the Enlightenment
The background and influence of Pietism
The background and influence of naturalism
The influence of nationalism
European offshoots in the New World
Spanish and Portuguese America
French Québec
British America
VIII. Western education in the 19th century
The social and historical setting
The early reform movement: the new educational philosophers
Pestalozzianism
Froebel and the kindergarten movement
Herbartianism
Other German theorists
French theorists
Development of national systems of education
Germany
France
England
Russia
The United States
The British dominions
Spread of Western educational practices to Asian countries
India
Japan
IX. Education in the 20th century
Social and historical background
Summary of major trends and problems
Major intellectual movements
Influence of psychology on education
Traditional movements
Experimental movements
Western patterns of education
United Kingdom
Germany
France
Other European countries
United States
Elder members of the British Commonwealth
Revolutionary patterns of education
Russia: from tsarism to Communism
China: from Confucianism to Communism
Modernizing Afro-Asian patterns of education
Japan
India
South Africa
Education in colonies and newly emerging nations
Patterns of African education
Patterns of education in the Arab world
Patterns of education in Latin America
Patterns of education in Southeast Asia

I. Education in primitive and early civilized cultures

PREHISTORIC AND PRIMITIVE CULTURES

The term education can be applied to primitive cultures only in the sense of enculturation, which is the process of cultural transmission. A primitive man, whose culture is the totality of his universe, has a relatively fixed sense of cultural continuity and timelessness. The model of life is relatively static and absolute, and it is transmitted from one generation to another with little deviation. As for prehistoric education, it can only be inferred from educational practices in surviving primitive cultures.

The purpose of primitive education is thus to mold the child in the manner in which he becomes a good member of his tribe or band. There is a marked emphasis upon training for citizenship, because primitive people are highly concerned with the growth of individuals as tribal members and the thorough comprehension of their way of life during passage from prepuberty to postpuberty.

Participatory learning in primitive societies

Obviously, because of the variety in the countless thousands of primitive cultures, it is difficult to describe any standard and uniform characteristics of prepuberty education. Nevertheless, certain things are practiced commonly within cultures. Children actually participate in the social processes of adult activities, and their participatory learning is based upon what the American anthropologist Margaret Mead has called empathy, identification, and imitation. Primitive children, before reaching puberty, learn by doing and observing basic technical practices. Their teaching personnel are not strangers but, rather, their immediate community.

In contrast to the spontaneous and rather unregulated imitations in prepuberty education, postpuberty education in some cultures is strictly standardized and regulated. The teaching personnel may consist of fully initiated men, often unknown to the initiate though they are his relatives in other clans. The initiation may begin with the initiate being abruptly separated from his familial group and sent to a secluded camp where he joins other initiates. The purpose of this separation is to deflect the initiate's deep attachment away from his family and to establish his emotional and social anchorage in the wider web of his culture.

The initiation "curriculum" does not usually include practical subjects. Instead, it consists of a whole set of cultural values, tribal religion, myths, philosophy, history, rituals, and other knowledge. Primitive people in some cultures regard the body of knowledge constituting the initiation curriculum as most essential to their tribal membership. Within this essential curriculum, religious instruction takes the most prominent place.

EDUCATION IN THE EARLIEST CIVILIZATIONS

The Old World civilizations of Egypt, Mesopotamia, and North China. The history of civilization started in the Near East about 3000 BC, whereas the North China civilization began about a millennium and a half later. The Mesopotamian and Egyptian civilizations flourished almost simultaneously during the first civilizational phase (3000–1500 BC). Although these civilizations differed, they shared monumental literary achievements. The need for the perpetuation of these highly developed civilizations made writing and formal education indispensable.

Priestly control of Egyptian and Babylonian education

Egypt. Egyptian culture and education were preserved and controlled chiefly by the priests, a powerful intellectual elite in the Egyptian theocracy who also served as the political bulwarks by preventing cultural diversity. The humanities as well as such practical subjects as science, medicine, mathematics, and geometry were in the hands of the priests, who taught in formal schools. Vocational skills relating to such fields as architecture, engineering, and sculpture were generally transmitted outside the context of formal schooling.

Egyptians developed two types of formal schools for privileged youth under the supervision of governmental officials and priests: one for scribes and the other for priest trainees. At the age of five, pupils entered the writing school and continued their studies in reading and writing until the age of 16 or 17. At the age of 13 or 14, the schoolboys were also given practical training in offices for which they were being prepared. Priesthood training began at the temple college, which boys entered at the age of 17, the length of training depending upon the requirements for various priestly offices. It is not clear whether or not the practical sciences constituted a part of the systematically organized curriculum of the temple college.

Rigid method and severe discipline were applied to achieve uniformity in cultural transmission, since deviation from the traditional pattern of thought was strictly prohibited. Drill and memorization were the typical methods employed. But, as noted, Egyptians also used a work–study method in the final phase of the training for scribes, when the trainees were required to take part in actual professional work.

Mesopotamia. As a civilization contemporary with Egyptian civilization, Mesopotamia developed education quite similar to that of its counterpart with respect to its purpose and training. Formal education was practical and aimed to train scribes and priests. It was extended from basic reading, writing, and religion to higher learning in law, medicine, and astrology. Generally, youth of the upper classes were prepared to become scribes, who ranged from copyists to librarians and teachers. The schools for priests were said to be as numerous as temples. This indicates not only the thoroughness but also the supremacy of priestly education. Very little is known about higher education, but the advancement of the priestly work sheds light upon the extensive nature of intellectual pursuit.

As in the case of Egypt, the priests in Mesopotamia dominated the intellectual and educational domain as well as the applied. The centre of intellectual activity and training was the library, which was usually housed in a temple under the supervision of influential priests. Methods of teaching and learning were memorization, oral repetition, copying of models, and individual instruction. It is believed that the exact copying of scripts was the hardest and most strenuous and served as the test of excellence in learning.

North China. In North China, the civilization of which began with the emergence of the Shang era, complex educational practices were in effect at a very early date. In fact, every important foundation of the formation of modern Chinese character was already established, to a great extent, more than 3,000 years ago.

Moral emphases of Chinese education

Chinese ancient formal education was distinguished from the education discussed earlier by the markedly secular and moral character of its educational purpose. Its paramount purpose was to develop a sense of moral sensitivity and duty toward people and the state. Even in the early civilizational stage, harmonious human relations, rituals, and music formed the curriculum.

Formal colleges and schools probably antedate the Chou dynasty of the 1st millennium BC, at least in the imperial capitals. Local states probably had less organized institutions, such as halls of study, village schools, and district schools. With regard to actual methods of education, ancient Chinese learned from bamboo books and obtained moral training and practice in rituals by word of mouth and example. Rigid rote learning, which typified later Chinese education, seems to have been rather condemned. Education was regarded as the process of individual development from within.

The New World civilizations of the Maya, Aztecs, and Incas. The outstanding cultural achievements of the pre-Columbian civilizations are often compared with those of Old World civilizations. The ancient Mayan calendar, which surpassed Europe's Julian calendar in accuracy, was, for example, a great accomplishment demonstrating the extraordinary degree of knowledge of astronomy and mathematics possessed by Maya. Equally impressive are the sophistication of the Incas' calendar and their highway construction, the development of the Maya's complex writing system, and the magnificent temples of the Aztecs. It is unfortunate that archaeological findings and written documents hardly shed sufficient light upon education among the Maya, Aztecs, and Incas. But from available documents it is evident that these pre-Columbian civilizations developed formal education for training the nobility and priests. The major purposes of education were cultural conservation, vocational training, moral and character training, and control of cultural deviation.

The Maya. Being a highly religious culture, the Maya regarded the priesthood as one of the most influential factors in the development of their society. The priest enjoyed high prestige by virtue of his extensive knowledge, literate skills, and religious and moral leadership, and high priests served as major advisers of the rulers and the nobility. To obtain a priesthood, which was usually inherited from his father or another close relative, the trainee had to receive rigorous education in the school, where priests taught history, writing, methods of divining, medicine, and the calendar system.

Priestly control of Maya and Aztec education

Character training was one of the salient features of Mayan education. The inculcation of self-restraint, cooperative work, and moderation was highly emphasized in various stages of socialization as well as on various occasions of religious festivals. In order to develop self-discipline, the future priest endured a long period of continence and abstinence, and, to develop a sense of loyalty to community, he engaged in group labour.

The Aztecs. Among the Aztecs, cultural preservation relied heavily upon oral transmission and rote memorization of important events, calendrical information, and religious knowledge. Priests and noble elders, who were called conservators, were in charge of education. Since one of the important responsibilities of the conservator was to censor new poems and songs, he took the greatest care in teaching poetry, particularly divine songs.

At the *calmecac*, the school for native learning where apprenticeship started at the age of ten, the history of Mexico and the content of the historical codices were systematically taught. The *calmecac* played the most vital role in ensuring oral transmission of history through oratory, poetry, and music, which were employed to make accurate memorization of events easier and to galvanize remembrance. Visual aids, such as simple graphic representations, were used to guide recitation phases, to sustain interest, and to increase comprehension of facts and dates.

The Incas. The Incas did not possess a written or recorded language as far as is known. Like the Aztecs, they also depended largely on oral transmission as a means of maintaining the preservation of their culture. Inca education was divided into two distinct categories: vocational education for common Incas and highly formalized training for the nobility. As the Inca Empire was a theocratic, imperial government based upon agrarian collectivism, the rulers were concerned about the vocational training of men and women in collective agriculture. Personal freedom, life, and work were subservient to the community. At birth an individual's place in the society was strictly ordained, and at five years of age every child was taken over by the government, and his socialization and vocational training were supervised by government surrogates.

Education for the nobility was clearly defined in terms of the curricula and rituals. The Inca nobility consisted of three classes: the imperial *ayllus*; the *curacas*, who

were the tribal chiefs; and distinguished common Indians. The sons of the *ayllus* and the *curacas* received education for four years in learning centres called the *cuzco* schools. In the first year the pupils learned the Quechua language, the language of the nobility. The second year was devoted to the study of religion, and the third year to learning about the quipus, a complex system of knotted coloured strings or cords used for sending messages and recording historical events. In the fourth year major attention was given to the study of history, with additional instruction in sciences, geometry, geography, and astronomy. The instructors were highly respected encyclopaedic scholars known as *amautas*, to whom even the supreme Inca listened. After the completion of this education, the pupils were required to pass a series of rigorous examinations in order to attain full status in the life of the Inca nobility. (N.S.)

II. Education in classical cultures

ANCIENT INDIA

The Hindu tradition. India is the site of one of the most ancient civilizations in the world. About the 2nd millennium BC, the Aryans entered the land and came into conflict with the *dāsas*, or the non-Aryan tribes. They defeated them, spread far and wide in the country, established large-scale settlements, and founded powerful kingdoms. In the course of time, a section of the intellectuals, the Brahmins, became priests and men of learning; another group, nobles and soldiers, became Kṣatriyas; the agricultural and trading class was called Vaiśyas; and finally the *dāsas* were absorbed as Śūdras, or domestic servants. Such was the origin of the division of the Hindus into four *varṇas*, or "classes." By about 500 BC, the classes became hardened into castes. The Brahmins occupied the most venerable position in society.

The Vedic tradition in Hindu education

Religion was the mainspring of all activities in ancient India. It was of an all-absorbing interest and embraced not only prayer and worship but everything—philosophy, morality, law, and government. Religion saturated educational ideals, too, and the study of Vedic literature was indispensable to higher castes. The stages of instruction were very well defined—namely, up to the age of seven at home, from eight to 16 at school, and then at a university. During the first period, the child received primary education at home. Formal schooling, however, began with a ritual known as the *upanayana*, or thread ceremony, which was more or less compulsory for the three higher castes. The Brahmin children had this ceremony at the age of eight, the Kṣatriya children at the age of 11, and the Vaiśya children at the age of 12 years. The *upanayana* ceremony marked the beginning of secondary education. The boy would leave his father's house and enter his preceptor's *āśrāma*, or home, situated amid sylvan surroundings. The *ācārya* would treat him as his own son, give him free education, and not charge anything for his boarding and lodging. The pupil had to tend the sacrificial fires, do the household work of his preceptor, and look after his cattle.

The study at this stage consisted of the recitation of the Vedic *mantras*, or "hymns," and the auxiliary sciences—phonetics, the rules for the performance of the sacrifices, grammar, astronomy, prosody, and etymology. The character of education, however, differed according to the needs of the caste. For a boy of the priestly class, there was a definite syllabus of studies. The *trayī-vidyā*, or the knowledge of the three Vedas, the most ancient of Hindu scriptures, was obligatory for him. During the whole course at school, as at college, the student had to observe *brahmacarya*—that is, wearing a simple dress, living on plain food, using a hard bed, and leading a celibate life.

The period of studentship normally extended to 12 years. For those who wanted to continue their studies, there was no age limit. After finishing their education at an *āśrama*, or forest school, they would join a higher centre of learning or a university presided over by a *kulapati* (a founder of a school of thought). Advanced students would also improve their knowledge by taking part in philosophical discussions at a *parisad*, or "academy." Occasionally, a great king would summon a national gathering or congress to which the representative thinkers of the various schools of thought in the country were invited. Education was not denied to women, but normally girls were instructed at home.

The method of instruction differed according to the nature of the subject. The first duty of the student was to learn by heart the particular Veda of his school. Every word and line of the text had to be learned from the lips of the teachers, and so correct pronunciation was stressed. In the study of such literary subjects as law, logic, rituals, and prosody, comprehension played a very important role. A third method was the use of parables, which were employed in the personal spiritual teaching relating to the *Upaniṣads*, or conclusion of the Vedas. In higher learning, such as in the teaching of *śāstras* ("sciences"), the most popular and useful method was catechism—the pupil asking questions and the teacher discoursing at length on the topics referred to him. Memorization, however, played the greatest role. It could not be avoided even when a topic was thoroughly grasped.

The introduction of Buddhist influences. By about the end of the 6th century BC, the Vedic rituals and sacrifices had gradually developed into a highly elaborate cult that profited the priests but antagonized an increasing section of the people. Education became generally confined to the Brahmins, and the *upanayana* was being gradually discarded by the non-Brahmins. The formalism and exclusiveness of the Brahmanic system was largely responsible for the birth of two new religious orders, Buddhism and Jainism. Neither of them recognized the authority of the Vedas, and both challenged the exclusive claims of the Brahmins to priesthood. Their founders, Buddha and Mahāvīra, taught through the common language of the people and gave education to all, irrespective of caste, creed, or sex. Buddhism also introduced the monastic system of education. Monasteries attached to Buddhist temples served the double purpose of imparting education and of training persons for priesthood. A monastery, however, educated only those who were its members, living under a common roof. It did not admit day scholars. Thus, it was not possible for the Buddhist order to cater to the needs of the entire population.

Buddhism and Jainism in Indian education

Meanwhile, significant developments were taking place in the political field that had repercussions on education. The establishment of the imperialistic Nanda dynasty in *c.* 413 BC and then of the even stronger Mauryas some 40 years later shook the very foundations of the Vedic structure of life, culture, and polity. The Brahmins in large numbers gave up their ancient occupation of teaching in their forest retreats and took to all sorts of occupations; the Kṣatriyas also abandoned their ancient calling as warriors; and the Śūdras in their turn rose from their servile occupations. These forces produced revolutionary changes in education. Schools were established in growing towns, and even day scholars, including married and regular householders, were admitted. Studies were chosen freely and not according to caste, and teachers admitted students of all castes. Taxila had already acquired an international reputation in the 6th century BC as a centre of advanced studies and now improved upon it. It did not possess any college or university in the modern sense of the term, but it was a great centre of learning with a number of famous teachers, each having a school of his own; it included special institutions of law, medicine, and military science.

In the 3rd century BC Buddhism received a great impetus under India's most celebrated ruler, Aśoka. After his death, Buddhism evoked resistance, and a counter-reformation in Hinduism began in the country. About the 1st century AD, there was also a widespread lay movement among both Buddhists and Hindus. As a result of these events, Buddhist monasteries began to undertake secular as well as religious education, and there began a large growth of popular elementary education along with secondary and higher learning.

Classical India. The 500 years from the 4th century AD to the close of the 8th, under the Guptas and Harṣa

The university at Nālandā

and their successors, is a remarkable period in Indian history. It was the age of the universities of Nālandā and Valabhī and of the rise of Indian sciences, mathematics, and astronomy. The university at Nālandā owed its foundation to six generations of the Gupta kings. It housed a population of several thousand teachers and students, who were maintained out of the revenues from more than 100 villages. Because of its fame, Nālandā attracted students from abroad, but the admission test was so strict that only two or three out of ten attained admission. More than 1,500 teachers discussed over 100 different dissertations every day. These covered the Vedas, logic, grammar, philosophy (*sāṃkhya, nyāya,* etc.), astronomy, and medicine. Other great centres of Buddhist learning of the post-Gupta era were Vikramaśīla, Odantapurī, and Jagaddala. The achievements in science were no less significant. Āryabhaṭa in the late 5th century was the greatest mathematician of his age. He introduced the concepts of zero and decimals. Varāhamihira of the Gupta age was a profound scholar of all the sciences and arts, from botany to astronomy and from military science to civil engineering. There was also considerable development of the medical sciences. According to contemporaries, more than eight branches of medical science, including surgery and pediatrics, were practiced by the physicians.

These were the main developments in education prior to the Muslim invasions, beginning in the 10th century. Nearly every village had its schoolmaster, who was supported from local contributions. The Hindu schools of learning, known as *pathasala*s in western India and *tol*s in Bengal, were conducted by Brahmin *ācāryas* at their residence. Each imparted instruction in an advanced branch of learning and had a student enrollment of not more than 30. Larger or smaller establishments, specially endowed by *rajas* and other donors for the promotion of learning, also grew in number. The usual centres of learning were either some king's capital, such as Kanauj, Dhār, Mithilā, or Ujjayinī, or a holy place, such as Vārānasi, Ayodhyā, Kāñcī, or Nasik. In addition to Buddhist *vihāras* ("monasteries"), there sprang up Hindu *maṭhas* (monks' residences) and temple colleges in different parts of the country. There were also *agrahāra* villages, which were given in charity to the colonies of learned Brahmins in order to enable them to discharge their scriptural duties including teaching. Girls were usually educated at home, and vocational education was imparted through a system of apprenticeship.

Indian influences on Asia. An account of Indian education during the ancient period would be incomplete without a discussion of the influence of Indian culture on western and Southeast Asia. It was achieved partly through cultural or trade relations and partly through political influence. Khotān in Central Asia had a famous Buddhist *vihāra* as early as in the 1st century AD. A number of Indian scholars lived there, and many Chinese pilgrims, instead of coming to India, stayed there. Indian pandits (scholars) were also invited to China and Tibet, and many Chinese and Tibetan monks studied in Buddhist *vihāra*s in India.

Indianization of Southeast Asia

The process of Indianization was at its highest in Southeast Asia. Beginning in the 2nd century AD, Hindu rulers reigned in Indochina and in the numerous islands of the East Indian archipelago, from Sumatra to New Guinea, for a period of 1,500 years. These regions were peopled by primitive races, who adopted the civilization of their masters. A greater India was thus established by a general fusion of cultures. Some of the inscriptions of these countries, written in flawless Sanskrit, show the influence of Indian culture. There are references to Indian philosophical ideas, legends, and myths and to Indian astronomical systems and measurements. Hinduism continued to wield its influence on these lands so long as the Hindus ruled in India. This influence ceased by the 15th century AD. (S.N.M.)

ANCIENT CHINA

According to the prehistory records, education in China was early recognized as an important function of the state. Good government consisted of teaching the people the proper conduct so that they would know how to live harmoniously and happily. A good ruler was one who set a good example for the people to follow, and officers of the state were specifically charged with the duty of attending to the instruction of the people.

Ancient Chinese education served the needs of a simple agricultural society with the family as the basic social organization. Paper and the writing brush had not been invented, and the "bamboo books" then recorded to be in existence were of limited use at best. Oral instruction and teaching by example were the chief methods of education.

The molding of character was a primary aim of education. This was done by imparting knowledge of the laws of the land and teaching the acceptable modes of behaviour. Even more important than knowledge was actual participation in the rituals, the prescribed customs, and the music and dance that were actually a part of the rituals. Through such participation correct behaviour was taught and learned.

Ethical teachings stressed the importance of human relations and the family as the foundation of society. Filial piety, especially emphasizing respect for the elderly, was considered to be the most important virtue. Special institutions were designated for the care of the elderly and the performance of rituals reflecting the respect for age.

Education and public service

Two characteristics of Chinese education were already evident in the early period: the emphasis on ethical and moral values and the close link between education and public service. It was the responsibility of government to provide instruction so that the talented would be able to enter government service and thus perpetuate the moral and ethical foundation of society. The selection of educated persons for government service foreshadowed the elaborate examination system of later periods. The basis of selection was usually the recommendation of local officials. It was recorded, however, that a model emperor of the early legendary period personally examined his officers every third year and decided upon their promotion or dismissal after three examinations.

The Chou period. *Early Chou (1122–721 BC).* This was the feudal age, when the feudal states were ruled by lords who paid homage to the king of Chou and recognized him as the "Son of Heaven." The relations between the feudal lords and the king were governed by prescribed rules and rituals assuring the stability of government and society.

Schools were established for the sons of the nobility, but the education of the common people was not neglected. The former were located in the capital city of Chou and the capital cities of the feudal states; within the feudal states there were schools in villages and hamlets that were attended, according to written records, by men and women after their work in the fields. There were elementary and advanced schools for both the ruling classes and the common people. Separate studies were prescribed for girls; they were concerned chiefly with homemaking and the feminine virtues that assured the stability of the family system.

The content of education for the ruling classes consisted of the "six arts"—rituals, music, archery, charioteering, writing, and mathematics. They constituted what may be called the "liberal education" of the period. Mere memory work was condemned. As Confucius said of the ancient spirit of education, "learning without thought is labour lost."

Later Chou (771–221 BC). This was a period of social change brought about by the disintegration of the feudal order, the breakdown of traditional loyalties, the rise of cities and urban civilization, the growth of commerce, and the appearance of scholar-philosophers offering a variety of solutions for the problems of the day. The decline of the authority and prestige of the king was proportionate to the increasing power and arrogance of the feudal states, until the latter practically ignored the royal position and engaged in strife and warfare leading to the collapse of the Chou dynasty.

The instability and the perplexing problems of the times challenged scholars to propose various remedies. The absence of central control facilitated independent and creative thinking. Thus appeared one of the most creative periods in China's intellectual history, when a Hundred Schools of thought vied with one another to expound their views and their proposals for attaining a happy social and political order. Some urged a return to the teachings of the sages of old, while others sought better conditions by radical change. Among the major "schools" of this age were Taoism, Confucianism, Moism, and Legalism (see CHINESE RELIGION; CHINESE PHILOSOPHY). No one school was in the ascendancy. Orthodoxy was yet unknown.

The Hundred Schools of thought

Each major school had its followers and disciples, among whom there was a vigorous program of instruction and intellectual discussion. The breakdown of the feudal system meant the neglect of education by the government, and this period saw the decline of state schools and the growth of private ones. Most active in the establishment of private schools were Confucius and his disciples, but the Taoists, the Moists, and the Legalists also maintained their own teaching institutions.

Another form of educational activity was the practice of the contending feudal states of luring to their domain a large number of scholars, partly to serve as a source of ideas for enhancing the prosperity of the state and partly to gain an aura of intellectual respectability in a land where the respect for scholars had already become an established tradition. The princes and dukes not only provided the scholars with material sustenance and comfort but strove to accord them due respect and honour. The age of political instability and social disintegration therefore turned out to be an age of free and creative intellectual activity. Conscious of their importance and responsibility, the scholars developed a tradition of self-respect and fearless criticism. It was this tradition that Confucius had in mind when he said that the educated person was not a utensil to be used, and it was this spirit that the Confucian philosopher Mencius described when he said that the great man was a man of principles whom riches and position could not corrupt, whom poverty and lowliness could not swerve, whom power and force could not bend.

The teachings of the Hundred Schools and the records of the feudal states meant a marked increase in literature and, consequently, in the materials for instruction. The period of the Eastern Chou (so named from the location of its capital) is the classical age of China, which bequeathed to subsequent periods an intellectual and educational legacy of inestimable value. Its scholars propounded theories of government and of social and individual life that were as influential in China and East Asia as the Greek philosophers of almost contemporary age were in the Western world. The classics that were to become the main content of education until modern times were written or compiled during this period.

The Ch'in–Han period. *Ch'in autocracy (221–206 BC).* Of the various schools of thought that arose in China's classical age, Legalism was the first to be accorded official favour. The policies of the Ch'in dynasty were based on Legalist principles stressing a strong state with a centralized administration. Many of its policies were so different from past practices that they incurred the criticism of scholars, especially those who upheld the examples of the ancient sages. To stop the criticism, the ruler, who called himself the first emperor, acting upon the advice of a Legalist minister, decreed a clean break with the past and a banning of books on history and of classics glorifying past rulers. Numerous books were collected and burned, and hundreds of scholars were put to death for failure to surrender the banned books.

Legalism and banning of books

Though condemned for the burning of books and the persecution of scholars, the Ch'in dynasty laid the foundation for a unified empire and made it possible for the next dynasty to consolidate its power and position at home and abroad. In education, the unification efforts included a reform and simplification of the written script and the adoption of a standardized script intelligible throughout the country. First steps were taken toward uniform textbooks for the primary schools. The invention of the writing brush made of hair, as well as the making of ink, led to the replacement of the clumsy stylus and bamboo slips with writing on silk.

Scholarship under the Han (206 BC–AD 220). The Ch'in dynasty did not last long. Capitalizing on the popular discontent with the short-lived regime, the Han dynasty reversed many of the policies of its predecessor. The most important change was a shift from Legalism to Confucianism. The banned books were now highly regarded, and the classics became the core of education. An assiduous effort was made to recover the prohibited books and to discover books and manuscripts that scholars had concealed in secret places. Much painstaking work was done in copying and editing, and the textual and interpretative studies of the Han scholars accorded a new importance to the study of the classics. The making of paper further stimulated this revival of learning. Critical examination of old texts resulted in the practice of higher criticism long before it developed in the West.

Shift to Confucianism

There were historians, philosophers, poets, artists, and other scholars of renown in the Han dynasty. Deserving special mention is Ssu-ma Ch'ien, author of a monumental history of China from the earliest times to the 1st century BC, whose high level of scholarship earned him the title "Chinese Father of History." An illustrious woman of letters, Pan Chao, was named poet laureate. A bibliographer collected and edited ancient texts and designated them as classics. The first dictionary of the Chinese language was written. Books multiplied; historical and literary works became models of critical scholarship. Since the discovery and interpretation of ancient texts had largely been the work of Confucian scholars, Chinese scholarship from now on became increasingly identified with Confucianism. Most of the Han rulers gave official sanction to Confucianism as a basis of conducting government and state affairs. There was, however, no action to exclude other schools of thought.

There were a variety of schools on the national and local levels. The trend of increasing activity in private education continued, and much of the study of the classics and enriched literature of the day was done in private schools. Of considerable influence in the country and abroad was a national university with an enrollment that soared to 30,000. The classics now became the core of the curriculum, but music, rituals, and archery were still included. The tradition of all-round education in the six arts had not vanished.

Recruiting educated persons for state service had always been important. Now competitive examinations were coming into use, and the combination of examinations with proper recommendation was considered a good way to insure the selection of the worthy for government service.

Introduction of Buddhism. The Han dynasty was a period of territorial expansion and growth in trade and cultural relations. Buddhism was introduced at this time.

Early information about Buddhism was probably brought into China by traders, envoys, and monks. By the 1st century AD, an emperor became personally interested and sent a mission to India to seek more knowledge and bring back Buddhist literature. Thereafter, Indian missionaries as well as Chinese scholars translated Buddhist scriptures and other writings into Chinese, and the way was opened for the spread of the new faith among the people.

Buddhist and Indian influences in China

Indian missionaries not only preached a new faith but also brought in new cultural influences. Indian mathematics and astronomical ideas enriched Chinese knowledge in these fields. Chinese medicine also benefitted. Architectural and art forms reflected Buddhist and Indian influence. Hindu chants became a part of Chinese music. Translation of new ideas and new terms also resulted in new vocabulary and writing forms.

For a couple of centuries after its introduction, however, Buddhism showed no signs of popular appeal. Han scholarship was engrossed in the study of ancient classics

and was dominated by Confucian scholars who had scant interest in Buddhist teachings that were unconcerned with the practical issues of moral and political life. Moreover, the Buddhist view of evil and the Buddhist espousal of celibacy and escape from earthly existence were alien to China's traditions. The Taoist scholars, finding in Buddhism much that seemed not too remote from their own spiritual message, were much more inclined to study the new philosophy, and some of them aided in the translation of Buddhist texts. The Taoists, however, were not in the centre of the Han stage.

The fall of the Han dynasty was followed by a few hundred years of division, strife, and foreign invasions. China was not united again until the end of the 6th century. It was during this period of confusion and disunity that Buddhism gained a foothold in China. Political chaos and foreign invasions produced a widespread feeling of uncertainty and insecurity, and Buddhist asceticism seemed to fit the mood of the times.

The literary efforts of Chinese monks produced a Chinese Buddhist literature, and this marked the beginning of a process that transformed an alien importation into a Chinese religion and system of thought. The non-Chinese characteristics of Buddhism were soon to be submerged under the growth of distinctly Chinese sects. There was much study and teaching of Buddhism by scholars and monks and religious converts, but most of this educational activity took place in monasteries and temples rather than in schools which were attended by the people. The content of the examinations, however, generally reflected the growth of Buddhist thought and Buddhist literature. (T.H.C.)

ANCIENT HEBREWS

Like all preindustrial societies, ancient Israel first experienced a type of education that was essentially familial; that is to say, the mother taught the very young and the girls, while the father assumed the responsibility of providing moral, religious, and handcraft instruction for the growing sons. This characteristic remained in Jewish education, for the relation of teacher to pupil was always expressed in terms of parenthood and filiation. Education, furthermore, was rigid and exacting; the Hebrew word *musar* signifies at the same time education and corporal punishment.

The education of scribes

Once they were established in Palestine—at the crossroads of the great literate civilizations of the Near East, in the beginning of the 1st millennium BC—the Jewish people learned to develop a different type of education—one that involved training a specialized, professional class of scribes in a then rather esoteric art called writing, borrowed from the Phoenicians. Writing was at first practical: the scribe was the man who wrote letters and drew up contracts, who kept accounts, maintained records, prepared orders. Because he could receive written orders, he eventually became entrusted with their execution; hence the importance of scribes in the royal administration, well-attested since the times of David and Solomon. The training given these scribes was not merely technical or utilitarian; it included training of character and instilling the high ideal of wisdom, as would befit the servants of the king.

The evolution of Hebrew schools

Writing found another avenue of application in Israel—in religion. And the scribe again was the agent of education. He was the man who copied the sacred Law faithfully and established the canonical text. He was the one who read the Law to himself and to the people, taught it, and translated it when Hebrew ceased to be the vernacular or "living language" (into Greek in Alexandria, into Aramaic in Palestine); he explained it, commented on it, and studied its application in particular cases. After the downfall of Israel in 722–721 BC and Judah in 586 BC and their subjection to foreign rule, Jewish education became characterized more and more by this religious orientation. The synagogue in which the community assembled became not merely a house of prayer but also a school, with a "house of the book" (*bet sefer*) and a "house of instruction" (*bet ha-midrash*) corresponding roughly to elementary and secondary or advanced levels of education.

The role of writing in this Oriental world should not be exaggerated, of course; oral instruction still held first place by far. Although a pupil might learn to read aloud, or rather to intone his text, his main effort was to learn by heart fragment after fragment of the sacred Law. Alongside this written Law, however, there developed interpretations or exegeses of it, which at first were merely oral but which progressively were reduced to writing—first in the form of memoranda or *aides-mémoire* inscribed on tablets or notebooks, then in actual books. The diffusion of this religious literature called for an expansion of programs of instruction, evolving into diverse stages: elementary, intermediate, and advanced, the latter in several centres in Palestine, later in Babylonia. This religiously based education was to become one of the most important factors enabling Judaism to survive the national catastrophes of AD 70 and 135, involving the capture and subsequent destruction of Jerusalem. In their dispersion, the Jews clung to Hebrew, their only language for worship, for the study of the Law, for tradition, and consequently for instruction. From this evolved the respect with which the teacher was and is surrounded in Jewish communities. Teaching is ennobled by its association with the sacred text; as the Talmud records, "One must have as much respect for the teacher as for God."

ANCIENT GREEKS

Origins. The history of the Hellenic language and therewith of the Hellenic people goes back to the Mycenaean civilization of about 1400–1100 BC, which itself was the heir of the pre-Hellenic civilization of Minoan Crete. Because the Mycenaean civilization consisted of little monarchies of an Oriental type, with an administration operated by a bureaucracy, one may assume an educational system designed for the training of scribes, analogous to those known from the ancient civilizations of the Near East. But continuity did not exist between this education and that which was to develop after an interruption of obscure centuries known as the Greek dark ages (11th to 9th centuries BC). When the Greek world reappeared in history, it was an entirely different society, one headed by a military aristocracy as idealized in Homer's *Iliad* and *Odyssey*. During this period, sons of the nobility received their education at the court of the prince in the setting of a guild companionship of warriors: the young nobleman was educated through the counsel and example of an older man to whom he had been entrusted or had entrusted himself, a senior admired and loved. It was in this atmosphere of virile camaraderie that there developed the characteristic ideal of Greek love that was enduringly to mark Hellenic civilization and to deeply influence its conception of education itself—for example, in the relation of master to pupil. Yet these warriors of the archaic period were not coarse barbarians but already in some sense knights. Their education was not confined to an apprenticeship in arms and horsemanship but gave much attention to sports (which survived, little changed, into classical Greece), to the dance, instrumental music, and oral poetry, and, finally, to the taste for words, the germ of eloquence. There was already a certain refinement of manners, a delicacy that showed itself particularly in the respect accorded to women. More characteristic and perhaps more important was the ethical ideal that characterized this knightly education, an ethic based on the notion of valour (*aretē*). The cult of military or sporting prowess by which one could assert oneself, could outclass others, was a source of prestige and glory. It was an ethic of honour, which made virtues of pride and of jealousy as the inspiration of great deeds and which accepted it as natural that one would be the object of jealousy or of hatred. The cult of Homer, which until the end of antiquity (and in Byzantium even later) was to constitute the basis of Greek culture and therewith of Greek education, would maintain from generation to generation this "agonistic" ideal: the cult of the hero, of

Education in the Greek archaic ages

the champion, of high performance, which found an outlet outside the sphere of battles in games or contests (*agōnes*), particularly in the realm of sports, the most celebrated being the Olympic Games, dating traditionally from 776 BC.

Profound changes were introduced into Greek education as a result of the political transformations involved in the maturing of the city-state. There then developed a collective ideal of devotion to the community: the city (*polis*) was everything to its citizens; the city made its citizens what they were—mankind. This subordination of the individual exploit to collective discipline was reinforced by the strategic military revolution that saw the triumph of heavy infantry, the hoplites, foot soldiers heavily armed and in tight formation.

Sparta. It is in Sparta, the most flourishing city of the 8th and 7th centuries BC, that one sees to best advantage the richness and complexity of this archaic culture. Education was carried to a high level of artistic refinement, as evidenced by the events organized within the framework of the city's religious festivals. The young men and women took part in processions, dances, and competitions in instrumental music and song. Physical education had a like part, equally for both sexes, given status by national or international contests (the Spartans regularly took more than half of the first places at the Olympic games); and military and civic education had a large place, as was proper for the citizen-soldier to be ready to fight and if necessary to die for his country.

Spartan education for a warrior caste

This last aspect became not merely dominant but exclusive from the time (about 550 BC) when a conservative reaction triumphed at Sparta, bringing to power a militarist and aristocratic regime. Arts and sports gave way completely to an education appropriate to men of a warrior caste. The education of girls was subordinated to their future function as mothers; a strict eugenic regime pitilessly eliminated sickly and deformed children. Up to the age of seven, children were brought up by the women, already in an atmosphere of severity and harshness. Education, properly speaking, *agōgē*, lasted from ages seven to 20 and was entirely in the hands of the state.

The male youth of Sparta were enrolled into formations corresponding to successive age classes, divided into smaller units as in modern Scouting, under the authority of comrades of their own age or of young officers. It was a collective education, which progressively removed them from family life and subjected them to garrison life. Everything was organized with a view to preparation for military service: ill-clothed, bedded on the bare ground, the child was poorly fed, told to steal to supplement his rations, and subjected to rigorous discipline. His virility and combativeness were developed by hardening him to blows—thus the role of ritual brawls between groups of boys and of the institution of the *krypteia*, a nocturnal expedition designed both to terrify the lower classes of slaves (helots) and to train the future fighter in ambushes and the ruses of warfare. He was also, of course, directly apprenticed to the military craft, using arms and manoeuvring in close formation. This puritanical education, proceeding in a climate of austerity, had as its sole norm the interests of the state, erected into a supreme category; the Spartan was trained under a strict discipline to obey blindly the orders of his superiors. Curiously, the child was at the same time trained to dissimulation, to lying, to theft—all virtues when directed toward the foreigner, toward whom distrust and Machiavellianism were encouraged.

This implacably logical education enabled Sparta to remain for long the most powerful city, militarily and diplomatically, of the entire Greek world and to triumph over its rival Athens after the long struggle of the Peloponnesian War (431–404 BC); but it did not prevent Sparta's decadence. Not that Sparta ever relaxed its tension: on the contrary, in the course of centuries, the rigour and ferocity were accentuated even as such behaviour became more and more anachronistic and without real use. Rites of initiation were transformed into barbarous tests of endurance, the boys undergoing flagellation and competing in enduring it, sometimes to the very death, under the eyes of tourists attracted by the sadistic spectacle. This occurred in times of complete peace when, under the Roman Empire, Sparta was nothing but a little provincial city with neither independence nor army.

Athens. Beginning at a date difficult to fix precisely (at the end of the 7th or during the 6th century), Athens, in contrast to Sparta, became the first to renounce education oriented toward the future duties of the soldier. The Athenian citizen, of course, was always obligated, when necessary and capable, to fight for the fatherland, but the civil aspect of life and culture was predominant: armed combat was only a sport. The evolution of Athenian education reflected that of the city itself, which was oriented toward increasing democratization—though it should be noted that the slave and the resident alien always remained excluded from the body politic. The Athenian democracy, even in its most complete form, attained in the 4th century BC, was to remain always the way of life of a minority—about one-tenth, it is conjectured, of the total population. Athenian culture continued to be oriented toward the noble life, that of the Homeric knight, minus the warrior aspect, and this orientation determined the practice of elegant sports. Some of these, such as horsemanship and hunting, always remained more or less the privilege of an aristocratic and wealthy elite; the various branches of athletics, however, originally reserved for the sons of the great families, became more and more widely practiced.

Athenian education for a democratic minority

Education of youth. Under such conditions, there arose the need of a collective education—and of the school. It was at first a school of gymnastics, or *palaistra*, under the authority of the boys' coach, or *paidotribēs;* it was for the youngsters what the gymnasium was for the adults. (From 632 BC the Olympic Games added contests for children.) But education also became musical; a second category of specialized teacher appeared, that of the teacher of music, including dance, choral singing, instrumental music on the lyre or the *aulos* (a kind of oboe), lyric music, and finally poetry. By way of this latter poetry, a literary aspect was introduced into education. Certainly the poetry, sung or recited, was at first entrusted primarily to memory, but recourse to reading and writing gradually increased, and there appeared a third type of school—third in time, as in the importance credited to it—in which the child learned to read, write, and count.

The moral aspect of education was not neglected. The Athenian ideal was that of the *kalos kathagos*, the "wise and good" man. The teachers were as much preoccupied with overseeing the child's good conduct and the formation of his character as with directing his progress in the various subjects taught him. Poetry served to transmit all the traditional wisdom, which combined two currents: the ethic of the citizen expressed in the moralizing elegies of the 6th-century lawmaker Solon and the old Homeric ideal of the value of competition and the exploit. But this ideal equilibrium between the education of the body and that of the mind was interrupted before long as a result on the one hand of the development of professional sports and the exigencies of its specialization and on the other by the development of the strictly intellectual disciplines, which had made great progress since the time of the first philosophers of the 5th century BC.

Higher education. With the great Sophists, contemporaries and adversaries of Socrates (*c.* 470–399 BC), one sees the emergence of a system of higher education open to all—to all, at any rate, who had the leisure and necessary money. Until then, the higher forms of culture had retained an esoteric character, being transmitted by the master to a few chosen disciples, as in the first schools of medicine at Cnidus and at Cos, or within the framework of a religious confraternity involving initiate status. The Sophists proposed to meet a new need that was generally felt in Greek society, particularly in the most active cities such as Athens, where political life had been intensively developed. Henceforth the management of public affairs became the supreme occupation engaging

Sophistic education

the ambition of Greek man; it was no longer in sports and elegant leisure activities that his valour, his desire to assert himself and to triumph, would find expression but rather in political action.

The Sophists, professional educators, introduced a form of higher education whose success attested to its social utility and practical efficacy—a commercial success, promoted by an advertising setup: they inaugurated the literary genre of the public lecture, which was to experience a long popularity. It was a teaching process oriented in an entirely realistic direction, the education of statesmen. The Sophists pretended neither to transmit nor to seek for the truth concerning man or existence; they offered simply an art of success in political life, which meant, above all, being able on every occasion to make one's point of view prevail. Two principal disciplines constituted the program: the art of persuasion, or dialectic, and the art of speaking, or rhetoric—the two most flourishing humanistic sciences of antiquity. These disciplines the Sophists founded by distilling from experience their general principles and logical structures, thus making possible their transmission on a theoretical basis from master to pupil.

To the pedagogy of the Sophists there was opposed the activity of Socrates, who, as inheritor of the old aristocratic tradition, was alarmed by this radical utilitarianism. He doubted that virtue could be taught, especially for money, a degrading substance. An heir also of the old sages of former times, Socrates held that the supreme ideal of man and hence of education was not the spirit of efficiency and power but the distinterested search for the absolute, for virtue—in a word, for wisdom.

It was only at the beginning of the 4th century BC, however, that the principal types of classical Greek higher education became organized on definitive lines. This was the result of the joint and rival efforts of the two great educators, the philosopher Plato (*c.* 428–348/347), who opened his school, the Academy, probably in 387, and the orator Isocrates (436–338), who founded his school *c.* 390.

Platonic education

A disciple of Socrates and like him an heir of the aristocratic tradition, Plato had early tried to take advantage of the ambition of men of his social class and to dream of the creation of an elite of men of politics, councillors to the sovereign, or leaders of the people. Blocked by the triumph of democratic ideas in Athens, however, he turned to a more personal conception of the ideal of the Greek man—as a philosopher. Plato's Academy did not take the form of a commercial enterprise but of a *thiasos*, a religious association devoted to the cult of the Muses (and, after the death of the master, to the cult of Plato himself, venerated as a hero). A notion of what his curriculum may have been can be gleaned from the idealized programs offered in the utopias of his books *The Republic* and *Laws*. The curriculum was founded on the traditional education of the Athenian child: for education of the body, it included gymnastics (which was broadened, however, to embrace hygiene and dietetics); for education of the spirit or soul, it included *mousikē*, "the Muses' thing"—dance, choral and solo singing, playing the lyre, and poetry. Then—this was the essential feature—there followed a long preparatory education that was characterized essentially by the study of mathematics and that, in principle, was to last as long as 30 years. Only then, when his intellect had been thoroughly tested and accustomed to moving in the rarefied atmosphere of abstract rational thought, would the disciple be able to attack the study of philosophy itself. Overall, this plan of education was doubly aristocratic: its origins were class-conscious; and its aims were to weed out those persons whose mental capacities and philosophical leanings proved insufficient to the tasks of long years of training.

The education of Isocrates

The rival school of Isocrates—a school open to fee-paying customers and not having an esoteric character like that of the Academy—was much more down-to-earth and practical. If it too aimed at a form of wisdom and claimed for itself the name of philosophy in defining its ideal, it was a wisdom of a much more practical order, based on working out commonsense solutions to life's problems. In contrast to Plato, Isocrates sought to develop the quality of grace, cleverness, or finesse rather than the spirit of geometry. The program of study that he enjoined upon his pupils was more literary than scientific. Besides gymnastics and music, its basics included the study of the classics (the poets first of all, with Homer in the foreground), the great prose authors, the historians, and the philosophers. Its higher levels consisted of an extensive study of the queen of disciplines, rhetoric—a study consisting of five or six years of theory, analysis of the great classics, imitation of the classics, and finally practical exercises. Such was the practical training of the Athenian citizen.

The two parallel forms of culture and of higher education must not be considered totally opposed: both opposed the cynical pragmatism of the Sophists; each influenced the other. Isocrates did promote elementary mathematics as a kind of mental training or mental gymnastics and did allow for a smattering of philosophy to illumine broad questions of human life. Plato, for his part, did recognize the usefulness of the literary art and philosophical rhetoric. The two traditions appear as two species of one genus; their debate, continued in each generation, enriched classical culture without jeopardizing its unity.

Aristotelian education

Before leaving the Hellenic age, there is one other great figure to appraise—one who was a bridge to the next age since he was the tutor of the young prince who became Alexander the Great of Macedonia. Aristotle (384–322 BC), who was one of Plato's pupils and shared some of his opinions about education, believed that education should be controlled by the state, and that it should have as a main objective the training of citizens. The last chapter of his *Politics* opens with these words:

> No one will doubt that the legislator should direct his attention above all to the education of youth. . . . The citizen should be moulded to suit the form of government under which he lives.

He shared some of Plato's misgivings about democracy; but, because he was no recluse but a man of the world acquainted with public affairs, he declared his preference for limited democracy over other forms of government. His worldliness also led him to be less concerned with the intuition of universals, in the Platonic mode, and more concerned with the observation of specific things. His urge for logical structure and classification, for systematization, was especially strong.

This systematization extended to a youth's education. In his first phase, from birth to age seven, he was to be physically developed, learning how to endure hardship. From age seven to puberty, his curriculum would include the fundamentals of gymnastics, music, reading, writing, and enumeration. During the next phase, from puberty to age 17, the student would be more concerned with exact knowledge, not only carrying on with music and mathematics but also exploring grammar, literature, and geography. Finally, in young manhood, only a few superior students would continue into higher education, developing encyclopaedic and intensely intellectual interests in the biological and physical sciences, ethics, and rhetoric, as well as philosophy. Aristotle's school, the Lyceum, was thus much more empirical than Plato's Academy.

The Hellenistic age. Alexander the Great's conquest of the Persian Empire between 334 and 323 BC abruptly extended the area of Greek civilization by carrying its eastern frontier from the shores of the Aegean to the banks of the Syrdarya and Indus rivers in Central Asia. Its unity rested henceforward not so much on nationality (it incorporated and assimilated Persians, Semites, and Egyptians) nor on the political unity soon broken after the death of Alexander in 323 but on a common Greek way of life, the fact of sharing the same conception of man. This ideal was no longer social, communal in character, as had been that of the city-state; it now concerned man as an individual—or, better, as a person. This civilization of the Hellenistic age has been defined as a civilization of *paideia*—which eventually denoted the condition of a person achieving enlightened, mature self-ful-

The concept of *paideia*

fillment but which originally signified education per se. The Greeks succeeded in preserving their distinctive national way of life amid this immense empire because, wherever numbers of them settled, they brought with them their own system of education for their youth, and they not only resisted being absorbed by the "barbarian," non-Hellenic peoples but succeeded somewhat in spreading Greek culture to many of the alien elite. It is important to note that, although Hellenism was finally to be swept away in the Near East by the Persian national renaissance and the invasions originating from Central Asia beginning in the 2nd century BC, it continued to flourish and even expand in the Mediterranean world under Roman domination. Hellenistic civilization and its educational pattern were prolonged to the end of antiquity and even beyond; it was to be a slow metamorphosis and not a brutal revolution that would later give birth to the civilization and education strictly called Byzantine.

The institutions. Hellenistic education comprised an ensemble of studies occupying the young from age seven to age 19 or 20. To be sure, this entire program was completed only by a minority, recruited from the rich aristocratic and urban bourgeois classes. The students were mostly boys (girls occupied only a very modest place), and of course they were usually free citizens (masters, though some slaves were given a professional education occasionally reaching a high level).

As in the preceding era, education continued to be dependent upon the city, which remained the primary frame of Greek life. One of the results of the implantation of Hellenism in the Near East was the foundation of a network of cities or communities organized and administered in the Greek manner. In effect, the creation of vast kingdoms did not eliminate the role of the city, even if the latter was not altogether independent; the Hellenistic state was not at all totalitarian and sought to reduce its administrative machinery to a minimum. It relied upon the cities to assume responsibility for public services, that of education in particular. The city in turn looked to the contributions of the richest and most generous private individuals, either by requiring them to fill magistracies and supply costly services or by appealing to their voluntary generosity; the proper functioning of the Hellenistic city presupposed the willing contributions of "benefactors." Thus, certain educational institutions were supported—and in fact sometimes set up—by private foundations that specified exactly the use to be made of the income from their gift of capital. Many schools were private, the role of the city being limited to inspections carried out by special magistrates and to the organization of athletic and musical competitions and festivals.

Physical education. The Hellenistic school *par excellence* was still the school of gymnastics, the practice of athletic sports and the nudity that they required being the most characteristic feature contrasting the Greek way of life with that of the "barbarians." It consisted of an establishment called in some contexts *palaestra* and in others *gymnasion*. There were, at least in sufficiently large cities, several of these, separately for the different age classes and on occasion for the sexes. They were essentially sports grounds, generally square, surrounded by colonnades in which were set up the necessary services: cloakrooms, washstands, training rooms, massage rooms. At the side there was a track for footraces.

The *palaestra* or *gymnasion*

The foundation of the training always consisted of the sports properly called gymnastic and field. Horsemanship remained an aristocratic privilege. Nautical sports had a very modest role—a curious thing for a nation of sailors like the Greeks, but the fact is that they were by origin Indo-Europeans from the interior of the Eurasian continent. The other sports—ball games, hockey—were considered merely diversions or at best preparatory exercises. As the competition of professional sports grew, however, education based on sports progressively, though no doubt very slowly, lost its pre-eminent position. Although the popularity of athletic sports as spectacle endured, educational sports moved into the background, disappearing altogether in the Christian period (in the 4th century AD) in favour of literary studies.

There was a similar progressive decline, a similar final effacement, of artistic, particularly musical, education, the other survivor from the archaic culture. The art of music continued to flourish, but like sports it became the concern of professional practitioners and a feature of public spectacles rather than an art generally practiced in cultivated circles.

The primary school. The child from seven to 14 years of age went to the school of letters, conducted thither by a special slave, the *paidagōgos*, whose role was not limited to accompanying the child: he had also to educate him in conduct, good manners, and morals and finally to act as a lesson coach. Ignorance—or disdain—of child psychology resulted in the use of teaching methods that were scarcely efficient: flogging was the only aid used, and it was used generously. The teaching of reading involved an analytical method that made the process very slow. First the alphabet was taught from alpha to omega and then backward, then from both ends at once—alpha–omega, beta–psi, and so on to mu–nu. (A comparable progression in the Latin alphabet would be *A–Z*, *B–Y*, and so on to *M–N*.) Then were taught simple syllables—*ba, be, bi, bo*—followed by more complex ones, and then by words, successively of one, two, and three syllables. The vocabulary list included rare words (*e.g.*, some of medical origin), chosen for their difficulty of reading and pronunciation. It took several years for the child to be able to read connected texts, which were anthologies of famous passages. With reading was associated recitation and, of course, practice in writing, which followed the same gradual plan.

The program in mathematics was very limited; rather than computation, the subject, strictly speaking, was numeration: learning the whole numbers and fractions, their names, their written notations, their representation in finger counting (in assorted bent positions of the fingers and assorted placements of either hand relative to the body). The general use of tokens and of the abacus made the teaching of methods of computation less necessary than it became in the modern world.

Secondary education. Between the primary school and the various types of higher education, the Hellenistic educational system introduced a program of intermediate, preparatory studies—a preliminary education, a kind of common trunk preparing for the different branches of higher culture, *enkyklios paideia* ("general," or "common, education"). This general education, far from having "encyclopaedic" ambitions in the modern sense of the word, represented a reaction against the inordinate ambitions of philosophy and, more generally, of the Aristotelian ideals of culture, which had demanded the large accumulation of intellectual attainments. The program of the *enkyklios paideia* was limited to the common points on which, as noted earlier, the rival pedagogies of Plato and of Isocrates agreed, namely, the study of literature and mathematics. Specialized teachers taught each of these subjects. The mathematics program had not changed since the ancient Pythagoreans and comprised four disciplines—arithmetic, geometry, astronomy, and music (not the art of music but the theory of the numerical laws regulating intervals and rhythm). The primary function of the *grammatikos*, or professor of letters, was to present and explicate the great classic authors: Homer first of all, of whom every cultivated man was expected to have a deep knowledge, and Euripides and Menander—the other poets being scarcely known except through anthologies. Although poetry remained the basis of literary culture, room was made for prose—for the great historians, for the orators, Demosthenes in particular, even for the philosophers. Along with these explications of texts, the students were introduced to exercises in literary composition of a very elementary character (for example, summarizing a story in a few lines).

The program of this intermediate education did not attain its definitive formulation until the second half of the 1st century BC, after the appearance of the first manual devoted to the theoretical elements of language, the slim grammatical treatise by Dionysius Thrax. The program then consisted of the seven liberal arts: the three literary

The seven liberal arts

arts of grammar, rhetoric, and dialectic and the four mathematical disciplines noted above. (These were, respectively, the trivium and the quadrivium of medieval education, though the latter term did not appear until the 6th century and the former not until the 9th century.) The long career of this program should not conceal the fact that in the course of the centuries it fell into disuse and became rather largely a theory or abstraction; in reality, literary studies gradually took over at the expense of the sciences. Of the four mathematical disciplines, only one remained in favour—astronomy. And this was not merely because of its connections with astrology but primarily because of the popularity of the basic textbook used to teach it—the *Phaenomena*, a poem in 1,154 hexameters by Aratus of Soli—whose predominantly literary quality was suited to textual explications. Not until about the 3rd and 4th centuries AD was the need of a sound preparatory mathematical education again recognized and put into practice.

Higher education. Higher education appeared in several forms, complementary or competitive. First was the *ephebeia* ("youth" culture), a kind of civic and military training that completed the education of the young Greek and prepared him to enter into life; it lasted two years (from 18 to 20) and corresponded quite closely to the obligatory military service of modern states. It was a survival from the regime of the old Greek city-states, but in the Hellenistic age the absence of national independence erased all reason for this military training; between the 3rd and 2nd centuries BC, the Athenian *ephebeia* (eventually reduced to a single year) was transformed into a peaceful civilian college where a minority of rich young men came to be initiated into the refinements of the elegant life. Military training came to play only a modest role and gave way to sport as such. To this were added lectures on scientific and, above all, literary subjects, assuring the ephebe a polish of general culture. The same evolution took place in other cities: the *ephebeia* became everywhere more aristocratic than civic, more sporting than military. What the Greeks, especially those who had emigrated to the "barbarian" lands, demanded of it was above all that it initiate their sons into Greek life and its characteristic customs, beginning with athletic sports. Especially in Egypt, it was intended to legitimize the privileged status of the Hellene relative to the "native" Egyptian. In any event, the *ephebeia* no longer was the setting for the highest forms of education.

Education in science and medicine

Formal education in science also lacked any institutionalization. There were, to be sure, certain establishments having scientific staffs of high competence, such as the Museum established at Alexandria, richly endowed by the Ptolemies; but, at least initially, it was an institute for advanced studies rather than a university properly speaking. If the savants endowed there were also professors, this meant only that they dispensed instruction to a small circle of chosen disciples. The same informal character of personal training was to be seen in all the special disciplines—medicine, for example, which saw such a fine development between the time of Hippocrates (5th century BC) and that of Galen (2nd century AD). If there were in the Hellenistic era certain "schools" of medicine —old (Cnidus, Cos) and new (Pergamum, Alexandria)— these were less the equivalent of today's medical faculties than simply centres to which the presence of numerous qualified masters attracted a large number of aspirants. Whatever theory these "students" were able to learn, they learned largely through self-training and practice, by associating themselves with a practicing physician whom they accompanied to the bedsides of patients, taking part in his consultations, profiting by his experience and advice.

Philosophy and rhetoric were subjects of education most highly institutionalized. Although philosophy was taught privately by individual masters-lecturers, who could be either itinerants or residents of one place, these teachers were well organized and, in groups, possessed a kind of institutional character. On the model of Plato's Academy, the new Athenian schools of philosophy—Aristotle's Lyceum, Epicurus' Garden, the Porch (*stoa*), which gave its name to the Stoics—were brotherhoods in which the posts in both teaching and administration were passed from generation to generation as a kind of heritage. It was in philosophy that the personalistic character of the Hellenistic era most clearly asserted itself, in contrast to the more communal idea of the preceding period; when philosophy turned to the problem of politics, for instance, it dealt less with the citizens of a republic and more with the sovereign king, his duties and character. The central problem was henceforth that of wisdom, of the purpose that man should set for himself in order to attain happiness, the supreme ideal. The teaching of philosophy was not entirely contemplative: it involved the disciple in a conversion analogous to a religious conversion, a decision implying a revision of his life and the adoption of a generally ascetic way of life. Such a vocation, however, could obviously appeal only to a moral, intellectual, and financially secure elite; philosophers were never to be more than a small number within the Hellenistic (and Roman) intelligentsia.

The prestige of rhetoric

The reigning discipline was thus that of rhetoric. The prestige of the oratorical art outlived those social conditions that had inspired it; political eloquence operated only in the context of an embassy coming to plead the cause of a particular city or pressure group at the court of the sovereign. Legal eloquence did retain its function, and the profession of advocate retained its attractiveness; but it was above all the eloquence of showy set speeches, the art of the lecturer, that experienced a curious blossoming. Also, as a result of the customary habit of reading aloud, there was no sharp line between speech and the book; thus, eloquence imposed its rule upon all literary genres—poetry, history, philosophy. Even the astronomer and the doctor became lecturers.

Hence the importance attached to the teaching of rhetoric, which developed from century to century with an ever more rigorous technicalism, precision, and systematization. The study of rhetoric had five parts: invention (the art of finding ideas, according to standard schemes), disposition (the arrangement of words and sentences), elocution, mnemonics (memory training), and action. Action was the art of self-presentation, the regulation of voice and delivery, and above all the art of reinforcing the word with the expressive power of gesture. Each of these parts, equally systematized to the tiniest detail, was taught with a technical vocabulary of extreme precision. Such an education, which in addition to theory comprised a study of the great examples to be imitated and exercises in practical application, required many years of study; in fact, even in maturity, the cultivated Hellene continued to deepen his knowledge of the art, to drill himself, to "declaim."

A rivalry existed between philosophy and rhetoric, each trying to draw into its orbit the best and the most students. Even in the time of Plato and Isocrates, this rivalry did not proceed without mutual concessions and reciprocal influences, but it remained one of the most constant characteristics of the classical tradition and continued until the end of antiquity and beyond. The long summer of Hellenic civilization was extended under the Roman domination; the great centres of learning also experienced a long prosperity. Athens in particular was the unchallenged capital of philosophy; its *ephebeia* welcomed foreigners to come to crown their culture in the "school of Greece." Its masters of eloquence also had a solid reputation, even though they had competition from such schools of Asia Minor as those of Rhodes (in the 1st century BC) and Smyrna (in the 2nd century AD). Under the later Roman Empire, Alexandria, already famous for medicine, competed with Athens for pre-eminence in philosophy. Other great centres developed: Beirut, Antioch, and the new capital Constantinople. The quality of the teachers and the number of students attending permits one to apply to these centres, without too much anachronism, the modern designation of "universities."

ANCIENT ROMANS

Early Roman education. The quality of Latin education before the 6th century BC can only be conjectured.

Rome and Roman civilization were then dominated by a rural aristocracy of landed proprietors directly engaged in exploiting their lands, even after the establishment of the republic. Their spirit was far removed from Greece and Homeric chivalry; ancient Roman education was instead an education suitable for a rural, traditional people —instilling in youth an unquestioned respect for the "customs of the ancestors."

Education had a technological aspect, involving instruction on how to direct the use of one's lands, how to oversee the work of slaves, and how to advise the tenant farmers or one's steward. It had a legal aspect; in contrast to Athenian law, which relied more on equity than on written law, Roman justice was much more formalistic and technical and demanded much more study on the part of the citizen. Education had a moral aspect, aiming at inculcating rustic virtues, a respect for good management of one's patrimony, and a sense of austerity and frugality. Roman education, however, did not remain narrowly utilitarian; it broadened in urban Rome, where there developed the same ideal of communal devotion to the public weal that had existed in Greece—with the difference that in Rome such devotion would never be called into question. The interests of the state constituted the supreme law. The ideal set before youth was not that of the chivalrous hero in the Homeric manner but that of the great men of history who, in difficult situations, had by their courage and their wisdom saved the fatherland when it was in danger. A nation of peasants, Rome was also a nation of soldiers. Physical education was oriented not toward self-realization or competitive sport but toward military preparedness: training in arms, toughening of the body, swimming across cold and rapid streams, and horsemanship, involving such performances as mounted acrobatics and cavalry parades under arms.

The familial character of Roman education

Differing from the Greeks, the Romans considered the family the natural milieu in which the child should grow up and be educated. The role of the mother as educator extended beyond the early years and often had lifelong influence. If, in contrast to the girl, the boy at seven years of age escaped from her exclusive direction, it was to pass under that of his father; another trait of Roman pedagogy was that it was the father who closely supervised the development and the studies of his son, giving him instruction in an atmosphere of severity and moral exigency, through precept but even more through example. The young Roman noble accompanied his father as a kind of young page in all his appearances, even within the Senate.

Familial education ended at 16, when the adolescent male assumed the toga of manhood, the attire of the adult. He devoted one year to an apprenticeship in public life, no longer at his father's side but placed in the care of some old friend of the family, a man of politics laden with years and honours. Then came military service, first as a simple soldier (it was well for the future leader to learn first to obey), encountering his first opportunity to distinguish himself by courage in battle, but soon thereafter as a staff officer under some distinguished commander. Civil and military, the education of the young Roman was thus completed in the entourage of some high personage whom he regarded with respect and veneration, without ceasing, however, to gravitate toward the family orbit. The young Roman was brought up not only to respect the national tradition embodied in the example of the illustrious men of the past but also, very specifically, to respect the particular traditions of his own family, which too had had its great men and which jealously transmitted a stereotype, a specific attitude toward life. If ancient Greek education can be defined as the imitation of the Homeric hero, that of ancient Rome took the form of imitation of one's ancestors.

Roman adoption of Hellenistic education. Something of these original characteristics was to survive always in Roman society, so ready to be conservative; but Latin civilization did not long develop autonomously.

It assimilated, with a remarkable faculty for adaptation, the structures and techniques of the much further evolved Hellenistic civilization. The Romans themselves were quite aware of this, as evidenced by the famous lines of Horace: "Captive Greece captivated her rude conqueror and introduced the arts to rustic Latium" "Graecia capta ferum victorem cepit et artis intulit agresti Latio" (*Epistles*, II, i, 156).

Greek influence was felt very early in Roman education and grew ever stronger after the long series of gains leading to the annexation of Macedonia (168 BC), of Greece proper (146 BC), of the kingdom of Pergamum (133 BC), and finally of the whole of the Hellenized Orient. The Romans quickly understood what advantages they could draw from this more mature civilization, richer than their own national culture. The practical Romans grasped the advantages to be drawn from a knowledge of Greek, an international language known to many of their adversaries, soon to be their Oriental subjects, and grasped the related importance of mastering the art of oratory espoused by the Greeks. Second-century Rome assigned to the spoken word, particularly in political and legal life, as great an importance as had Athens in the 5th century. The Roman aristocrats quickly understood what a weapon rhetoric could be for a statesman.

Rome doubly adopted Hellenistic education: on the one hand, it came to pass that a Roman was considered truly cultivated only if he had the same education, in Greek, as a native Greek acquired; on the other hand, there progressively developed a parallel system of instruction that transposed into Latin the institutions, programs, and methods of Hellenistic education. Naturally, only the children of the ruling class had the privilege of receiving the complete and bilingual education. From the earliest years, the child, boy or girl, was entrusted to a Greek servant or slave and thus learned to speak Greek even before he was able to express himself well in Latin; he also learned to read and write in the two languages, with Greek again coming first. (Alongside this private tutoring there soon developed, from the 3rd century BC, a Greek public education in schools aimed at a socially broader clientele, but the results of this schooling were less satisfactory than the direct method enjoyed by the children of the aristocracy.) In following the normal course of studies, the young Roman was taught successively by an instructor of Greek letters (*grammatikos*) and then by a Greek rhetorician. Those desiring more complete training did not content themselves with the numerous and often highly qualified Greeks to be found in Rome itself but went to Greece to participate in the higher studies of the Greeks themselves. From 119 or 118 BC onward, the Romans secured admission to the Ephebic College at Athens, and in the 1st century BC such young Latins as Cicero were attending the schools of the best philosophers and rhetoricians at Athens and Rhodes.

The tutoring of Roman children

Roman modifications. The adoption of Hellenistic education did not proceed, however, without a certain adaptation to the Latin temperament: the Romans showed a marked reserve toward Greek athleticism, which shocked both their morals and their sense of the deep seriousness of life. Although gymnastic exercises entered into their daily life, it was under the category of health and not that of sport; in Roman architecture, the palaestra or gymnasium was only an appendage of the public baths, which were exaggerations of their Greek models. There was the same reserve, on grounds of moral seriousness, toward music and dance, arts suitable for professional performers but not for freeborn young men and least of all for young aristocrats. The musical arts indeed became integrated into Latin culture as elements of the life of luxury and refinement, but as spectacle rather than as amateur participation; hence their disappearance from programs of education. It must be remembered, however, that athletics and music were in Greece itself survivals of archaic education and had already entered upon a process of decline.

This education in a foreign language was paralled by a course of studies exactly patterned upon those of the Greek schools but transposed into the Latin language. The aristocracy was to remain always attached to the idea of private education conducted within the family, but social pressure brought about the gradual develop-

ment of public education in schools, as in Greece, at three levels—elementary, secondary, and higher; they appeared at different dates and in various historical contexts.

Roman primary and secondary education

Education of youth. The appearance of the first primary schools is difficult to date; but the use of writing from the 7th century BC implies the early existence of some kind of appropriate primary instruction. The Romans took their alphabet from the Etruscans, who had taken theirs from the Greeks; Etruscans and Latins quite naturally copied the pedagogy of the Hellenistic world: the same ignorance of psychology, the same strict and brutal discipline, the same analytical method characterized by slow progress—the alphabet (forward, backward, from both ends toward the middle), the syllabary, isolated words, then short sentences (one-line moral maxims), finally continuous texts—the same method for writing and the same numeration, rather than computation.

It was only between the 3rd and the end of the 1st centuries BC that Latin secondary education developed, staffed by the *grammaticus Latinus*, corresponding to the Greek *grammatikos*. Since the principal object of this education was the explication of poetry, its rise was hindered by the slowness with which Latin literature developed. The first of these teachers, Livius Andronicus, took as his subject matter his own Latin translation of the *Odyssey;* two generations later, Ennius explicated his own poetic works. Only with the great poets of the age of Augustus could Latin literature provide classics able to rival Homer in educational value; they were adopted as basic texts almost immediately after their appearance. Thereafter, and until the end of antiquity, the program was not to undergo further change, the principal authors being first of all Virgil, the comic author Terence, the historian Sallust, and the unchallenged master of prose, Cicero. The methods of the Latin grammarian were copied directly from those of his Greek counterpart; the essential point was the explication of the classic authors, completed by a theoretical study of good language using a grammar textbook and by practical exercises in composition, graduated according to a minutely regulated progression and always remaining rather elementary. Theoretically, the curriculum remained that of the seven liberal arts, but, as in Greece, it practically neglected the study of the sciences in favour of that of letters.

It was only in the 1st century BC that the teaching of rhetoric in Latin was established: the first Latin rhetorician appeared in 93 BC in a political context, namely, as a democratic initiative to counter the aristocratic education given in Greek, and, as such, was soon prohibited by the conservative party in power. It was not until the end of the century and the appearance of the works of Cicero that this education would be revived and become normal practice; first, Cicero's discourses offered the young Latin the equivalent of those of the Greek Demosthenes, and, second, Cicero's theoretical treatises provided a technical vocabulary obviating the need for Greek manuals. But this instruction was to remain always very close to its Hellenistic origins: the terminology used by the great pedagogue Quintilian (*c.* AD 35–*c.* 100) is much more impregnated with Hellenism, much less Latinized, than that which Cicero had proposed. At Rome, too, rhetoric became the form of higher education enjoying the greatest prestige; as in Greece, this popularity outlived the elimination of political eloquence. More than in Greece, legal eloquence continued to flourish (Quintilian had in mind particularly the training of future advocates), but, as in the Hellenic milieu, Latin culture became predominantly aesthetic: from the beginning of the empire, the public lecture was the most fashionable literary genre, and the teaching of rhetoric was very naturally oriented toward the art of the lecturer as the crowning achievement.

Higher education. Precisely because the oratorical art was incontestably the most popular subject of higher education, the Romans did not feel the same urgency to Latinize the other rival branches of knowledge, which interested only a small number of specialists with offbeat vocations. To be sure, the philosophical work of Cicero had the same ambition as his oratorical work and proved by its existence that it was possible to philosophize in Latin, but philosophy found no successors to Cicero as rhetoric did. There was never a Latin school for philosophy. Of course, Rome did not lack philosophers, but many used Greek as their means of expression (even the emperor Marcus Aurelius); those who, like Cicero, wrote in Latin—Seneca, for example—had taken their philosophy studies in Greek. It was the same in the sciences, particularly in the medical sciences; for long, there were no medical books in Latin except encyclopaedias on a popular level.

The innovation of legal education

On the other hand, Rome created in the school of law another type of higher education—the only one that had no equivalent in Hellenistic education. The position of law in Roman life and civilization is, of course, well-known. Perhaps even more than rhetoric, it offered young Romans profitable careers; very naturally, there developed an appropriate education to prepare them. At first elementary in character and entirely practical, it was given within the framework of apprenticeship: the professor of law (*magister juris*) was primarily a practitioner, who initiated into his art the group of young disciples entrusted to him; these audited his consultations and heard him plead or judge. Beginning in Cicero's time and undoubtedly under his influence, this instruction was paralleled by a systematic theoretical exposition. Roman law was thus promoted to the rank of a scientific discipline. True schools were progressively established and took on an official character; their existence is well attested beginning with the 2nd century AD. It was at this same time that legal education acquired its definitive tools, with the composition of systematic elementary treatises such as the *Institutiones* of Gaius, manuals of procedure, commentaries on the law, and systematic collections of jurisprudence. This creative period perhaps reached its peak at the beginning of the 3rd century AD. The works of the great legal authors of this time, which became classics, were offered by the law professor with much interpretation and explication—very similar to the way in which grammarians offered literature.

Rome, the capital, remained the great centre of this advanced study in law. At the beginning of the 3rd century, however, there appeared in the Orient the school of Beirut. The teaching there was in Latin; and, to hear it and profit by the advantages that it offered for a high administrative or judicial career, many young Greeks enrolled there, in spite of the language obstacle to be overcome. Only a legal career could persuade young Greeks to learn this language that was always regarded by them as "barbarous."

The Roman world became covered with a network of schools concurrent with the Romanization of the provinces. The primary school always remained private; on the other hand, many schools of grammar or rhetoric acquired the character of public institutions supported (as in the Hellenic world) either by private foundations or by a municipal budget. In effect, it was always the city that was responsible for education. The liberal central government of the high empire, anxious to reduce its administrative apparatus to a minimum, made no pretense of assuming charge of it. It was content to encourage education and to favour teaching careers by fiscal exemptions; and only very exceptionally did an emperor create certain chairs of higher education and assign them a regular stipend. Vespasian (AD 69–79) created two chairs at Rome, one of Greek rhetoric and the other of Latin rhetoric. Marcus Aurelius (AD 161–180) similarly endowed, in Athens, a chair of rhetoric and four chairs of philosophy, one for each of the four great sects—Platonism, Aristotelianism, Epicureanism, and Stoicism.

The durable character of Greco-Roman education

Education in the later Roman Empire. The dominant fact is the extraordinary continuity of the methods of Roman education throughout such a long succession of centuries. Whatever the profound transformations in the Roman world politically, economically, and socially, the same educational institutions, the same pedagogical methods, the same curricula were perpetuated without great change for 1,000 years in Greek and six or seven

centuries in Roman territory. At most, a few nuances of change need be noted. There was a measure of increasing intervention by the central government, but this was primarily to remind the municipalities of their duties in regard to education, to fix the remuneration of teachers, and to supervise their selection. Only higher education received direct attention: in AD 425, Theodosius II created a true university in the new capital of Constantinople and endowed it with 31 chairs for the teaching of letters, rhetoric (both Greek and Latin), philosophy, and law. Another notable innovation was that the exuberant growth of the bureaucratic apparatus under the later empire favoured the rise of one branch of technical education, that of stenography, necessary for secretaries—a humble career, no doubt, but one that in that absolute autocracy could lead, through ambition and intrigue, to high responsibilities.

The only evolution of any considerable extent involves the use of the Greek and Latin languages. There had always been only a few Greeks who learned Latin, even though the growing machinery of administration and the increasing clientele drawn to the law schools of Beirut and Constantinople tended on the whole to increase the numerical size of this tiny minority. On the other hand, in Latin territory, late antiquity exhibited a general recession in the use of Greek. The ideal remained unchanged, and high culture always proposed to be bilingual: but, except among those aristocratic elites who clung to the old traditions and among a few other exceptional groups, people generally knew Greek less and less well. This retrogression need not be interpreted solely as a phenomenon of decadence: it had also a positive aspect, being an effect of the development of Latin culture itself. The richness and worth of the Latin classics explain why the youth of the West had less time than formerly to devote to the study of the Greek authors. Virgil and Cicero had pushed back Homer and Demosthenes, just as in modern Europe the ancient languages have retreated before the progress of the national languages and literatures. Hence, in the later empire, there appeared specialists in intercultural relations and translations from Greek into Latin. In the 4th and particularly in the 5th century, medical education in Latin became possible, thanks to the appearance of a whole medical (and veterinary) literature consisting essentially of translations of Greek manuals. It was the same with philosophy: resuming Cicero's enterprise at a distance of more than five centuries, Boethius (*c.* 480–524) in his turn sought with his manuals and his translations to make the study of that discipline available in Latin. Although the misfortunes of Italy in the 6th century, including the Lombardian invasion, did not permit this hope to be realized, yet the work of Boethius was destined later to nourish the medieval renaissance of philosophic thought.

Christian use of Greco-Roman education

Nothing better demonstrates the prestige and the persistent allure of classical culture than the attitude taken toward it by the Christians. This new religion could have organized an original system of education analogous to that of the rabbinical school—that is, one in which children learned through study of the Holy Scriptures—but it did not ordinarily do so. Usually, Christians were content to have both their special religious education, provided by the church and the family, and their classical instruction, received in the schools and shared with the pagans. Thus, they maintained the tradition of the empire after it had become Christian. Certainly, in their view, the education dispensed by these schools must have presented many dangers, inasmuch as classical culture was bound up with its pagan past (at the beginning of the 3rd century, the profession of schoolteacher was among the professions that disqualified one from baptism); but the utility of classical culture was so evident that they considered it necessary to send their children to these same schools in which they barred themselves from teaching. From Tertullian to St. Basil the Great of Caesarea, the Christian scholars insisted on the dangers presented by the study of the classics, the idolatry and immorality that they promoted, but they sought to show how the Christian could nevertheless make good use of them. With the passage of time and the general conversion of Roman society and particularly of its ruling class, Christianity, overcoming its reserve, completely assimilated and took over classical education. In the 4th century, Christians were occupying teaching positions at all levels, from schoolmasters and grammarians to the highest chairs of eloquence. In his treatise *De doctrina Christiana* (426), St. Augustine formulated the theory of this new Christian culture: being a religion of the book, Christianity required a certain level of literacy and literary understanding; the explication of the Bible required the methods of the grammarian; preaching a new field of action required rhetoric; theology required the equipment of philosophy. The synthesis of Christianity and classical education had become so intimate that, when the barbarian invasions swept away the traditional school along with many other imperial and Roman institutions, the church, needing a literary culture for the education of its clergy, kept alive the cultural tradition that Rome had received from the Hellenistic world. (H.I.M.)

III. Education in Persian, Byzantine, early Russian, and Islāmic civilizations

ANCIENT PERSIA

The ancient Persian Empire began when Cyrus II the Great initiated his conquests in 559 BC, and it ended when it was overrun by the Muslims in AD 651. Three elements dominated this ancient Persian civilization: (1) a rigorous and challenging physical environment, (2) the activist and positive Zoroastrian religion and ethics, and (3) a militant, expansionist people. These elements developed in the Persians an adventurous personality mingled with intense national feelings.

Zoroastrian influences

In the early period (559–330 BC), known as the Achaemenid period for the dynastic name of Cyrus and his successors, education, sustained by Zoroastrian ethics and the requirements of a military society, aimed at serving the needs of four social classes—priests, warriors, tillers of the soil, and merchants. Three principles sustained Zoroastrian ethics: the development of good thoughts, of good words, and of good actions (see ZOROASTRIANISM AND PARSIISM). Achaemenid–Zoroastrian education stressed strong family ties and community feelings, acceptance of imperial authority, religious indoctrination, and military discipline.

Education was a private enterprise. Formative education was carried on in the home and continued after the age of seven in court schools for children of the upper classes. Secondary and higher education included training in law to prepare for government service, as well as medicine, arithmetic, geography, music, and astronomy. There were also special military schools.

In 330 BC Persia was conquered by Alexander the Great, and native Persian or Zoroastrian education was largely eclipsed by Hellenistic. Greek practices continued during the Parthian Empire (247 BC–AD 224), founded by seminomadic conquerors from the Caspian steppes. And, thus, truly Persian influences were not restored until the appearance of a new, more sophisticated and reform-minded dynasty, the Sāsānids, in the 3rd century AD. In what has been called the neo-Persian empire of the Sāsānids (AD 224–651), the Achaemenian social structure and education were revived and further developed and modified. Zoroastrian ethics, though more advanced than during the Achaemenian period, emphasized similar moral principles, but with new stress upon the necessity for labour (particularly agriculture), upon the sanctity of marriage and family devotion, and upon the cultivation of respect for law and of intellectualism—all giving to education a strong moral, social, and national foundation. The subject matter of basic education included physical and military exercises, reading (Pahlavi alphabet), writing (on wooden tablets), arithmetic, and the fine arts.

Academy of Gondēshāpūr

The greatest achievement of Sāsānid education was in higher education, particularly as it developed in the Academy of Gondēshāpūr. Here, Zoroastrian culture, Indian and Greek sciences, Alexandrian–Syrian thought, medical training, theology, philosophy, and other dis-

ciplines developed to a high degree, making Gondēshāpūr the most advanced academic centre of learning in the later period of Sāsānid civilization. The academy, to which came students from various parts of the world, advanced, among other subjects, Zoroastrian, Greek, and Indian philosophies; Persian, Hellenic, and Indian astronomy; Zoroastrian ethics, theology, and religion; law, government, and finance; and various branches of medicine. In medicine, two kinds of diseases were recognized —internal and external, caused, according to Sāsānids, by stench and dirt, heat and cold, hunger and thirst, anxiety and old age, intemperance and overindulgence. Health was defined as a state of harmony of elements in the body, sustained by good birth and proper nourishment. The necessary qualifications for the practice of medicine included not only mastery of the science of medicine but also indifference to material benefits and continuous professional improvement.

It was partly through the Academy of Gondēshāpūr that important elements of classical Greek and Roman learning reached the Muslims during the 8th and 9th centuries AD and through them to schoolmen of western Europe in Latin translations of Arabic works during the 12th and 13th centuries.

THE BYZANTINE EMPIRE

The Byzantine Empire was a continuation of the Roman Empire in the eastern Mediterranean area, when the empire was finally divided into eastern and western parts in AD 395. If one takes as its starting point the year 330, when Constantine transferred the capital to Byzantium (Constantinople), the Eastern Empire lasted from 330 to 1453, when the capital fell to the Ottoman Turks under Sultan Mehmed II.

Influences on Byzantine education and culture. Byzantine education, though essentially patronized by Greek Christian Orthodoxy, remained basically a continuation of classical education. Even the emperor Justinian's closing of the Neoplatonic schools of Athens in 529 was more a defensive measure against pagan traditions than an attempt to break away from the Greek classical traditions of learning, particularly on levels of higher education. This is evident particularly in the traditions of Byzantine universities, which remained to a large extent aloof from religious involvements.

Nonetheless, religious orthodoxy, or faith, was Byzantium's greatest strength. It held the empire together, now expanded and again reduced, for more than 1,000 years against eastern invaders. Faith was also the Byzantine culture's chief limitation, choking originality in the sciences and the practical arts. But within this limitation it preserved the literature, science, and philosophy of classical Greece in recopied texts, some of which escaped the plunders of the crusaders and were carried to southern Italy, restoring Greek learning there. Combined with the treasures of classical learning that reached Europe through the Muslims, this Byzantine heritage helped to initiate the beginnings of the European Renaissance.

Results of the division between East and West

The highlights of the development of Byzantine culture and education within the restrictions of Eastern Greek Orthodoxy may be shown by its contrast with Latin, or Western, Christianity. In Greek Christianity the emperor appointed the patriarch of Constantinople and had the power to reduce or remove his jurisdiction. In Latin Christianity the two centres of power, the state and the church, competed for supremacy—the church maintaining, by and large, the upper hand throughout the Middle Ages. Intellectually, the consequences of this division were many. (1) Geographically, it divided Europe between Greek and Latin orthodoxies. (2) Linguistically, it identified Byzantine orthodoxy with the Greek language and Western orthodoxy with the Latin. (3) Jurisdictionally, it put Greek orthodoxy under secular laws, with the emperor appointing and disposing the patriarchs, whereas the Latin church dominated the political fortunes of western Europe into the 13th and 14th centuries. (4) Liturgically, in Greek Orthodoxy, ecclesiastical ornaments, vessels, and vestments became more ornate than in the Western church.

Primary and secondary education. More is known about the development of higher education than about the development of education on the primary and secondary levels in the Byzantine world. The best available information provides, nonetheless, certain limited observations. In general, some type of education seems to have been available for all classes within a rigid class structure, which provided better educational opportunities for upper class children. This aristocratic restriction was considerably reduced after the 11th century. Some provisions were also available for the education of girls. The emperor Julian (361–363) introduced state examinations and enforced state appointments of all teachers. Theodosius II (408–450) went so far as to make it a state offense for any teacher to give public instruction without a state license and saw to it that licenses were given only to conformists of orthodoxy. Schools were subsidized by the government.

In addition to some private tutoring, formal education included the following types of schools. Primary schools were available for a wide variety of middle and upper class children, including both urban and rural schools. Certain rudiments of classical learning formed part of the curriculum, in addition to an emphasis on reading and writing. These schools enrolled pupils variously between the ages of six and eight or eight and ten. On the secondary level, schools were open to pupils between the ages of ten and 16. Here there was stronger emphasis on the study of classics, particularly the works of Homer, in addition to instruction in mathematics, grammar, and religion. In patriarchal schools, which were fully developed by the 11th century, religious studies founded upon the Scriptures were taught by teachers chosen by the patriarch. But even in these schools the tradition of classical Humanism was not overlooked. Instruction in religion was complemented in these schools by the arts, including instruction in grammar and mathematics—combining, in a sense, religious and classical studies into a curriculum of Christian Humanism. The monastic schools were the only schools in the Byzantine culture devoted completely to religious disciplines and ascetic practices with total indifference or even hostility to the Humanistic tradition. These schools were established only for those who vowed dedication to a completely religious life. Children of the secular world were not admitted to them. Biblical studies, hymns, prose, and verse were central features of the curriculum.

Patriarchal schools

Higher education. The Byzantines remodelled the literary Humanism and moral philosophy of classical education to accord with Christian educational aims and came up with Christian Humanism, incorporating into this renovated curriculum such studies as literature, rhetoric, grammar, and philosophy. This classical tradition survived in the empire into the 8th century. In higher learning, important elements of classical culture were preserved in the great universities of larger cities, such as the universities of Constantinople, Alexandria, Antioch, and Beirut.

The Byzantine universities

In 425, under Theodosius II, higher education was reorganized in Constantinople, and there was established a university with 31 teachers, of whom 28 were assigned to instruction in Greek and Latin grammar, rhetoric, and literature, two were assigned to law, and one to philosophy. In 526 the grammarian Priscian wrote an extensive work on Greek and Latin grammar, *Institutiones grammaticae*, which became one of the famous texts in this field throughout the Middle Ages. Through such writers many of the Greek "pagan" classics were copied and taught in the schools of Constantinople, over the protest of churchmen, and remained more or less in vogue to the end of the Byzantine Empire.

For more than a thousand years (425–1453), save for temporary discontinuation, the University of Constantinople remained a rich centre of secular learning, embodying some of the best features of classical education, with periods of decline matched by as many periods of revival and reorganization in the 9th, 11th, 13th, and 14th centuries.

In general, the secular curricula of the major Byzantine

universities were quite diversified. Taken together, they included instruction in humanities, Greek classics, rhetoric and grammar, philosophy as a preparation for theology, secular music founded upon mathematical concepts, geometry, astronomy, logic, poetry in Atticized Greek, religious literature as a secular study, and arithmetic, as well as courses of study leading to the professions of law, medicine, architecture, and civil service. The state depended upon the universities to train qualified men for government services.

Thus, even after the educational centres of Athens had been closed, other universities and teaching centres carried much of the classical tradition through secondary and higher education, with the University of Constantinople and the patriarchal schools surviving as centres of Hellenistic Christian education to the end of the empire.

EARLY RUSSIAN EDUCATION: KIEV AND MUSCOVY

Influences on Russian education and culture. The influence of Byzantium, the Greek Orthodox Church, and other external forces upon early Russian education appears vividly in the light of six broad historical periods in Russia, based largely upon the shifting relations between the forest and steppe zones. Of concern here are the first four periods. The first, or prenational, period ended with Prince Svyatoslav's final attempt to unify a broad area encompassing the Dnieper River (Kiev and Novgorod), the Lower Volga, and the lower Danube. With his death in 972 this temporary empire crumbled, giving way to a gradual Christianization of Russia under Byzantine religious, cultural, and educational influences. Slavic mythology and Indo-European and Iranian themes, folklores, and arts were also contributing factors.

The struggle between the forest and the steppes, which continued into the second period, from 972 to 1240, is characterized by the unsuccessful though damaging efforts of Turkish tribes to subdue the Russians, whose ecclesiastical and political capital lay, at this time, in Kiev. The period was also characterized by further Russo-Byzantine cultural, religious, and educational developments, as well as by influences from Hungarians, Czechs, Poles, and Germans. During this period all the important elements of the Byzantine civilization and culture permeated Russia. Russia was officially converted to the Christian faith sustained by Byzantium but not as yet separate from the West. The gradual separation of Russian culture from western Europe began with the breach between the Byzantine and the Roman churches in the middle of the 11th century and the adherence of Russia to Greek Orthodoxy. This separation further widened and deepened until the end of the 17th century, or the reign of Peter the Great.

During the third period, from 1240 to 1452, the Mongol invasion terminated the struggle between the forest and the steppes, uniting the several Russian states under the domination of the Mongols, or Golden Horde, for about a century.

The two and a half centuries that followed the fall of Constantinople in 1453 were characterized by Russian expansion to the southeast. This expansionism continued with colonization of Siberia, the conquest of Kazan and Astrakhan, the reoccupation of the mouth of the Don, and the capture of Azov by Peter the Great in 1696 and other conquests later. During this period the church fell under the administrative control of the state and was divided into two metropolitan districts, Moscow and western Russia, with the patriarchate established in Moscow. These developments made a significant impact upon education.

Education under Kievan federation. As far back as 988, when the Russian church was officially converted to Greek Orthodoxy, Vladimir, metropolitan of Kiev, had advocated some form of formal elementary education. In 1028 Prince Yaroslav the Wise, the greatest ruler in the Kievan period, sent to Novgorod a request for 300 children's books; and in 1096 Prince Vladimir Monomakh wrote a long letter to his children on the advantages of education. This letter is one of the earliest documents on education in Russia. Schools as formal educational institutions, however, were not introduced into southern Russia until the end of the 16th century; the cultural nerve centre remained in Kiev, which was by then exposed to Western culture. In 1624, for instance, a Kievan religious order, the Lutsk Brotherhood, composed an educational document defining the qualifications of a teacher. They included devotion, judiciousness, humility, gentleness, continence, and sobriety. The good teacher should not be a usurer, a fornicator, a liar, or an envious, ridiculous, or irascible person; he should above all conduct himself with religious piety and never fall into heresy.

Clerical monopoly of learning

By the mid-17th century several types of schools flourished in southwestern Russia. These included elementary church schools, preparing students for religious service; schools founded by various brotherhoods advancing religious instruction; special home instruction in wealthy homes by private "masters"; Latin schools, similar to those established in Poland and Germany; and special schools giving literacy and religious instruction to children of the masses and occasionally to children of the nobility. Instruction was in the native language, reaching a larger number of pupils than some of the older schools. All these types of schools were established and supported by the people without state interference, thus anticipating the foundations of a system of public education. The Lutsk Brotherhood School laid down the strictest and most extensive rules for teaching and teacher conduct, school discipline, and parent–teacher cooperation.

The Lavr School, established in 1631, had an eight-year curriculum similar to some Jesuit schools, although it aimed at combatting the increasing influence of the Jesuits by teaching Greek and Latin, Slavonic, and the doctrines of Eastern Orthodoxy. The school attracted some fine teachers and became known as the Kiev Academy.

Muscovy educational development to 1725. The few elementary schools that existed in Muscovy up to the 17th century taught reading, writing, computation, and religion; and deacons and sextons served as teachers. The schools prepared students to become deacons and priests or government clerks. (Similar educational conditions prevailed in the Ukraine but with improved elementary schools.) Moscow established a divinity college in 1687, with emphasis on Greek language.

By the mid-16th century, the illiteracy of the clergy had become so notorious that it was decreed that schools be established in the homes of the better educated priests to teach church doctrines, singing, reading, and writing to any child, but particularly to students who prepared for the priesthood; the decree, however, was not widely carried out. Throughout the 16th and 17th centuries, home instruction was still largely the prevailing form of instruction. Oddly enough, secondary and higher education developed in Russia much earlier than elementary schools.

Romanov reforms

The Romanov family, which came into power in the early 17th century, ruled continuously until 1917, with absolute power for much of this period. The Romanovs developed cultural and educational relations with Germany, France, and England very early, with Peter the Great (tsar from 1682 to 1725) making systematic efforts to "westernize" Russia, inviting Western specialists to help found several schools of engineering and a naval academy. In a new type of school, called the cypher schools, arithmetic dominated the curriculum. It became compulsory for children of nobility to attend these schools. When the church was ordered by Peter in 1721 to facilitate education, diocesan schools were developed that by midcentury came into vogue, replacing the cypher schools and creating competition between the church and the crown for supremacy in education. Peter the Great also initiated efforts to establish an academy of science, which was finally founded by his widow as the Academy of Sciences in St. Petersburg in 1725.

Another school, the Slavonic–Greek–Latin Academy, had been licensed in 1682 and opened in Moscow in 1687. In 1701 Peter the Great transferred the academy to the state and turned it into a Latin school. Between 1630 and

1680 most Russian schools gradually changed from a Greek–Slavonic to a Polish–Latin orientation.

THE ISLAMIC ERA

Influences on Muslim education and culture. The Greco-Byzantine heritage of learning that was preserved through the medium of Near Eastern scholarship was combined with elements of Persian and Indian thought and taken over and enriched by the Muslims. It was initiated as early as the Umayyad caliphate (661–750), which allowed the sciences of the Hellenistic world to flourish in Syria and patronized Semitic and Persian schools in Alexandria, Beirut, Gondēshāpūr, Nisibis, Haran, and Antioch. But the largest share of Islām's preservation of classical culture was assumed by the ʿAbbāsid caliphate (750–*c.* 1100), which followed the Umayyad and encouraged and supported the translation of Greek works into Arabic, often by Nestorian, Hebrew, and Persian scholars. These translations included works by Plato and Aristotle, Hippocrates, Galen, Dioscorides, Alexander of Aphrodisias, Ptolemy, and others. The great mathematician al-Khwarizmi (flourished 9th century) compiled astronomical tables, introduced Hindu numerals (which became Arabic numerals), formulated the oldest known trigonometrical tables, and prepared a geographic encyclopaedia in cooperation with 69 other scholars.

The Muslim blend of scholastic heritages

The transmission of classical culture through Muslim channels can be divided into seven basic types: works translated directly from Greek into Arabic; works translated into Pahlavi, including Indian, Greek, Syriac, Hellenistic, Hebrew, and Zoroastrian materials, with the Academy of Gondēshāpūr as the centre of such scholarship (the works then being translated from Pahlavi into Arabic); works translated from Hindu into Pahlavi, then into Syriac, Hebrew, and Arabic; works written by Muslim scholars from the 9th through the 11th centuries but borrowed, in effect, from non-Muslim sources, with the line of transmission obscure; works that amounted to summaries and commentaries of Greco-Persian materials; works by Muslim scholars that were advances over pre-Islāmic learning but that might not have developed in Islām had there not been the stimulation from Hellenistic, Byzantine, Zoroastrian, and Hindu learning; and, finally, works that appear to have arisen from purely individual genius and national cultures and would likely have developed independent of Islām's classical heritage of learning.

Aims and purposes of Muslim education. Muslim education endeavoured to achieve two basic aims, religious and secular. The religious aims included the study of the Qur'ān as a source or basis of knowledge and the study of sectarian moral and spiritual disciplines, including the basic articles of faith. Secular aims included the synthesis of secular knowledge and religious belief, the conception and extensive practice of equal educational opportunity, and the establishment of teachers as guides and authorities in forms of knowledge based upon revelation, or direct intuition. These teachers were known as the *mutakallimun*, or *loquentes*.

It is important to emphasize the practical aims of Muslim secular education, which was responsible for the application of technological expertise to the development of irrigation systems, architectural innovations, textiles, iron and steel products, earthenware, and leather products; the manufacture of paper and gunpowder; the advancement of commerce; and the maintenance of a merchant marine. After the 11th century, when denominational interests dominated higher learning, academic freedom and practical studies gave way to religious and literary pursuits, Arabic language and grammar, and intolerance toward scientific innovations, secular subjects, and creative scholarship. These denominational colleges spread throughout eastern Islām from Transoxania (roughly modern Uzbek S.S.R.) to Egypt. Some 75 of them were in existence between about 1050 and 1250.

Varieties of Islāmic schools

Organization of education. There were many types of schools in the Muslim world, among them the *ḥalqah*, or circle schools; the *maktab* (*kuttab*), or writing schools; the palace schools; the *masjid*, or mosque schools; the bookshop schools; the *madrasah*, or schools of public instruction; and the universities.

The simplest type of early Muslim education was the circle, or gathering, school (*ḥalqah*), so named because the teacher was, as a rule, seated on a dais or cushion with the pupils gathering in a semicircle before him. The circle was determined by rank. The more advanced a student, the closer he was seated to the teacher, with the most advanced students or visiting scholars seated next to him. The mosque circles varied in approach, course content, and quality of teaching. Teachers were as a rule looked upon as masters of scholarship, and their lectures were meticulously recorded in notebooks, which were periodically examined, corrected, and approved for use in teaching others. The teacher encouraged students to discuss highlights of lecture notes with each other as preparation for advanced instruction in given areas. Advanced students were also encouraged to debate controversial points with their teachers and to feel free to challenge and even correct the teacher's statements. Every subject was dealt with methodically, beginning with a general survey of a subject, then connecting each day's lesson with previous lectures and a hint of the forthcoming lessons, giving in this manner coherence and continuity to a subject. Students often made long journeys to join the circle of a great teacher.

Writing schools (*maktab*, or *kuttab*), in which pupils learned to read and write, date back to the pre-Islāmic period in the Arab world. After the advent of Islām and the introduction of instruction in religion and the Qur'ān, these schools developed into centres for instruction in elementary Islāmic subjects. Some included in their curriculum the study of poetry, elementary arithmetic, penmanship, ethics (manners), elementary grammar, horsemanship, and swimming. *Maktab*s were quite common in almost every town or village in the Middle East, Africa, Sicily, and Spain.

Schools conducted in royal palaces taught not only the curriculum of the *maktab*s but also social and cultural studies designed to prepare the pupil for higher education, for service in the government of the caliphs, or for polite society. The instructors were called *muʾaddib*s, or instructors in good manners. Oratory, history, tradition, formal ethics, and the art of good conversation were often included in the curriculum.

The longest lasting of the elementary schools in Islām were the *masjid*s, which, as the name in Arabic implies, were attached to a mosque. Mosques increased in number under the caliphs, particularly the ʿAbbāsids: 3,000 of them were reported in Baghdad alone in the first decades of the 10th century; as many as 12,000 were reported in Alexandria in the 14th century, most of them with schools attached. Some mosques, such as that of al-Manṣūr, built during the reign of Hārūn-ar-Rashīd in Baghdad, or those in Isfahan, Mashhad, Ghom, Damascus, Cairo, and the Alhambra (Granada), became centres of learning for students from all over the Muslim world.

The high degree of learning and scholarship in Islām, particularly during the ʿAbbāsid period in eastern Islām and the later Umayyads in western Islām, encouraged the development of bookshops, copyists, and book dealers in large, important Islāmic cities such as Damascus, Baghdad, and Córdoba. Scholars and students spent many hours in these bookshop schools browsing, examining and studying available books, or purchasing favourite selections for their private libraries. Book dealers travelled to famous bookstores in search of rare manuscripts for purchase and resale to collectors and scholars and thus contributed the spread of learning. Many such manuscripts found their way to private libraries of famous Muslim scholars such as Avicenna, al-Ghazālī, and al-Fārābī, who in turn made their homes centres of scholarly pursuits for their favorite students.

Fundamental to Muslim education as were the circle schools, the *maktab*s, and the palace schools, they embodied definite educational limitations. Their curricula were limited; they could not always attract well-trained teachers; physical facilities were not always conducive to a congenial educational environment; and conflicts

between religious and secular aims in these schools were almost irreconcilable. These limitations led to the necessity of initiating a new type of school, the *madrasah*, to create a curriculum of studies more representative of the progressive diffusion of knowledge and to organize a body of better paid and better trained teachers to handle this enlarged curriculum more efficiently. These schools were widely established in Islāmic countries until the end of the 12th century.

The great Muslim universities

The crown and glory of medieval Muslim education were their famous research centres, or universities. Important among them were the universities of Niẓāmīyah and Mustanṣirīyah in Baghdad, which developed during the ʿAbbāsid caliphate, and those that developed in the Spanish cities of Córdoba, Seville, Toledo, Granada, Murcia, Almeria, Valencia, and Cádiz, in western Islām, under the Umayyads. The curricula of these schools offered both comprehensive programs of vocational studies and advanced studies in the professions, including courses in algebra, trigonometry, geometry, chemistry, physics, astronomy, medicine (including surgery, anatomy, pharmacy, and specialized medical branches), logic, ethics, metaphysics, geography, political disciplines, philology, poetry, grammar, prosody, law, jurisprudence, and all branches of theology.

The contribution of Muslim universities to the advancement of knowledge was vast. Muslim scholars calculated the angle of the ecliptic; measured the size of the Earth; calculated the procession of the equinoxes; invented the pendulum clock; explained, in the field of optics and physics, such phenomena as refraction of light, gravity, capillary attraction, and twilight; used the globe in teaching the geography of a round Earth; and developed observatories for the empirical study of heavenly bodies. They made advances in the uses of drugs, herbs, and foods for medication; established hospitals with a system of interns and externs; discovered causes of certain diseases and developed correct diagnoses of them; proposed new concepts of hygiene; made use of anesthetics in surgery with newly innovated surgical tools; and introduced the science of dissection in anatomy. They furthered the scientific breeding of horses and cattle; found new ways of grafting to produce new types of flowers and fruits; introduced new concepts of irrigation, fertilization, and soil cultivation; and improved upon the science of navigation. In the area of chemistry, Muslim scholarship led to the discovery of such substances as potash, alcohol, nitrate of silver, nitric acid, sulfuric acid, and mercury chloride. It also developed to a high degree of perfection the arts of textiles, ceramics, and metallurgy.

The Islāmic renaissance and medieval periods

Major periods of Muslim education and learning. The renaissance of Islāmic culture and scholarship developed largely under the ʿAbbāsid administration in eastern Islām and under the later Umayyads in western Islām, mainly in Spain, between 800 and 1000. This latter period, the golden age of Islāmic scholarship, was largely a period of translation and interpretation of classical thoughts and their adaptation to Islāmic theology and philosophy. The period also witnessed the introduction and assimilation of Hellenistic, Persian, and Hindu mathematics, astronomy, algebra, trigonometry, and medicine into Muslim culture and their modification and adaptation to Islāmic needs.

Whereas the 8th and 9th centuries, mainly between 750 and 900, were characterized by the introduction of classical learning and its refinement and adaptation to Islāmic culture, the 10th and 11th were centuries of interpretation, criticism, and further adaptation. There followed a period of modification and significant additions to classical culture through Muslim scholarship. Then, during the 12th and 13th centuries, most of the works of classical learning and the creative Muslim additions were translated from Arabic into Hebrew and Latin. The decline of Muslim scholarship coincided with the early phases of the European intellectual awakening that these translations were partly instrumental in bringing about.

Creative scholarship in Islām from the 10th to the 12th centuries included works by such scholars as Omar Khayyam, al-Bīrūnī, ar-Rāzī (Rhazes), Ibn Sīna (Avicenna), aṭ-Ṭabarī, Ibn Bājja (Avempace), and Ibn Rushd (Averroës).

Influence of Islāmic learning on medieval and modern Europe. The translation into Latin of most Islāmic works during the 12th and 13th centuries had a great impact upon the European Renaissance. As Islām was declining in scholarship and Europe was absorbing the fruits of Islām's three and a half centuries of creative productivity, signs of Latin Christian awakening were evident throughout the European continent.

The 12th century was one of intensified traffic of Muslim learning into the Western world through many hundreds of translations of Muslim works, which helped Europe seize the initiative from Islām when political conditions in Islām brought about a decline in Muslim scholarship. By 1300, when all that was worthwhile in Muslim scientific, philosophic, and social learning had been transmitted to European schoolmen through Latin translations, European scholars stood once again on the solid ground of Hellenistic thought, enriched or modified through Muslim and Byzantine efforts. (M.K.N.)

IV. The European Middle Ages

THE BACKGROUND OF EARLY CHRISTIAN EDUCATION

From the beginnings to the 4th century. At first Christianity found most of its adherents among the poor and illiterate, making little headway, as St. Paul observed (I Cor. 1:26), among the worldly-wise, the mighty, and those of high rank. But during the 2nd century AD and afterward it appealed more and more to the educated class and to leading citizens. These naturally wanted their children to have at least as good an education as they themselves had had, but the only schools available were the grammar and rhetoric schools, with their Greco-Roman, non-Christian culture. There were different opinions among Christian leaders about the right attitude to this dilemma that confronted all Christians who sought a good education for their children. The Greek Fathers, especially the Christian Platonists Clement of Alexandria and Origen, sought to prove that the Christian view of the universe was compatible with Greek thought and even regarded Christianity as the culmination of philosophy, to which the way must be sought through liberal studies. Without a liberal education the Christian could live a life of faith and obedience but could not expect to attain an intellectual understanding of the mysteries of the faith or expect to appreciate the significance of the Gospel as the meeting ground of Hellenism and Judaism. St. Augustine and St. Jerome also encouraged the use of the secular schools by Christians, maintaining that literary and rhetorical culture is valuable so long as it is kept subservient to the Christian life. The Roman theologian Tertullian, on the other hand, was suspicious of pagan culture, but he admitted the necessity, though deploring it, of making use of the educational facilities available.

Views of the early Christian Fathers

In any event, most Christians who wanted their children to have a good education appear to have sent their children to the secular schools; this practice continued even after 313, when the emperor Constantine, who had been converted to Christianity, stopped the persecution of Christians and gave them the same rights as other citizens. Christians also set up catechetical schools for the religious instruction of adults who wished to be baptized. Of these schools the most famous was the one at Alexandria in Egypt, which had a succession of outstanding heads, including Clement and Origen. Under their scholarly guidance, it developed a much wider curriculum than was usual in catechetical schools, including the best in Greek science and philosophy in addition to Christian studies. Other schools modelled on that at Alexandria developed in some parts of the Near East, notably in Syria, and continued for some time after the collapse of the empire in the west.

5th to 8th centuries. The gradual subjugation of the Western Empire by the barbarian invaders during the 5th century eventually entailed the breakup of the educational system that the Romans had developed over the

centuries. But, despite the political and social upheavals, the methods and program of ancient education survived into the 6th century in the new barbarian Mediterranean kingdoms; for the barbarians were frequently impressed and attracted by things Roman. In Ostrogothic Italy (Milan, Ravenna, Rome) and in Vandal Africa (Carthage), the schools of the grammarians and rhetoricians survived for a time, and, even in those places where such schools soon disappeared, as in Gaul and in Spain, private teachers or parents maintained the tradition of classical culture until the 7th century. As in previous centuries, the culture bestowed was essentially literary and oratorical: grammar and rhetoric constituted the basis of the studies. The pupils read, reread, and commented on the classical authors and imitated them by composing certain kinds of exercises (*dictiones*) with the aim of achieving a perfect mastery of their style. In fact, however, the practice was desultory, and the results were mechanical and poor. Furthermore, Greek was ignored more and more, and attempts to revive Hellenic studies were limited to a dwindling number of scholars.

Amid the general decline, the Christian Church persevered. From the 5th century to the time of Charlemagne in the 8th century, the church was able to provide some teaching of a religious character, especially for its clergy; and on those modest foundations a new educational system took shape gradually and fitfully. Learning that had no religious significance was rarely encouraged. But, even during the darkest days from the 5th to the 8th centuries, there were exponents of wider conceptions of education, and among them were such Christian scholars as Boethius and Cassiodorus in Rome, Isidore in Spain, and Bede in England. There were also monasteries in which the lamps of learning burned brightly. Particularly in Ireland, monasticism made its greatest contribution to the preservation and enrichment of culture during the 6th, 7th, and 8th centuries. In pre-Christian times there had been an educational tradition in Ireland, as in other Celtic regions, which the Irish monks took advantage of by developing schools open to the laity as well as to the clergy. They were keen students of the classics as well as of their native language and literature; and, when they crossed to the Continent as missionaries, they carried with them their knowledge of the literature of Greece and Rome, establishing monasteries, such as St. Gall (in Switzerland) and Reichenau (in Germany), that won renown for their scholarship.

Monastic and episcopal schools

The monastic schools, however, may not be as significant in the history of education as the schools founded by bishops, usually in connection with a cathedral. These bishops' schools are sometimes looked upon as successors of the grammar schools of the Roman Empire, and they are also often regarded as the parents of the grammar school of modern times. First specializing in the development of the clergy, these episcopal schools later admitted young lay people when the small Roman schools had disappeared. At the same time there were bishops who organized a kind of boarding school where the aspiring clergyman, living in a community, participated in duties of a monastic character, learned his clerical trade, and received doctrinal instruction.

The influence of monasticism affected the content of instruction and the method of presenting it. Children were to be dutiful, as the Celtic and English monks Columban and Bede were to remark: "A child does not remain angry, he is not spiteful, does not contradict the professors, but receives with confidence what is taught him." In the case of the adolescent destined for a religious profession, the monastic lawgiver was severe. The teacher must know and teach the doctrine, reprimand the undisciplined, and adapt his method to the different temperaments of the young monks. The education of young girls destined for monastic life was not much different from that of the monks. The mistress of the novices recommended prayer, manual work, and study.

Education of the laity

Between the 5th and 8th centuries, the principles of education of the laity likewise evolved. The treatises on education, later called the "mirrors," pointed to the importance of the four moral virtues—prudence, courage, justice, and temperance. The *Institutionum disciplinae* of an anonymous Visigoth pedagogue expressed the desire that all young men "quench their thirst at the quadruple fountain of the virtues." The author of this treatise also insisted on physical education, corresponding to Germanic standards of education. In the 7th and 8th centuries, the moral concepts of antiquity completely surrendered to religious principles. The Christian Bible was more and more considered as the only source of moral life, as the mirror in which men must learn to see themselves. A bishop addressing himself to a son of the Frankish king Dagobert (died 639) drew his examples from the books of the Old Testament. The mother of Didier of Cahors addressed to her son letters of edification on the fear of God, on the horror of vice, and on penitence.

The Christian education of children who were not aristocrats or future clergymen or monks was irregular. Whereas in antiquity catechetical instruction was organized especially for the adult laity, after the 5th century more and more children and then infants received baptism, and, once baptized, a child was not required to receive any particular religious education. His parents and godparents assisted him in learning the minimum, if anything at all. Only by attending church services and listening to sermons did the child acquire his religious culture.

The Irish and English revivals. During the 5th and 6th centuries there was a renaissance of learning in the remote land of Ireland, introduced there initially by the patron saints of Ireland—Patrick, Bridget, and Columcille—who established schools at Armagh, Kildare, and Iona. They were followed by a number of other native scholars, who also founded colleges—the most famous and greatest university being the one at Clonmacnois, on the Shannon River near Athlone. To these and lesser schools flocked Anglo-Saxons, Gauls, Scots, and Teutons from Britain and the Continent. From about AD 600 to 850, Ireland itself sent scholars to the Continent to teach, found monasteries, and establish schools.

Although the very earliest Irish scholars may have aimed primarily at propagating the Christian faith, their successors soon began studying and teaching the Greek and Roman classics, along with Christian theology. Eventually, there was a rich curriculum of mathematics, nature study, rhetoric, poetry, grammar, and astronomy, all studied, it seems, very largely through the medium of the Irish language.

England was next to experience the reawakening, and, though there were notable schools at such places as Canterbury and Winchester, it was in Northumbria that the schools flourished most. At the monasteries of Jarrow and Wearmouth and at the Cathedral School of York, some of the greatest of early medieval writers and schoolmasters appeared, including the Venerable Bede and Alcuin. The latter went to France in 780 to become master of Charlemagne's palace school.

THE CAROLINGIAN RENAISSANCE AND ITS AFTERMATH

The cultural revival under Charlemagne and his successors. Charlemagne (742/743–814) has been represented as the sponsor or even creator of medieval education, and the Carolingian renaissance has been represented as the renewal of Western culture. This renaissance, however, built on earlier episcopal and monastic developments; and, although Charlemagne did help to ensure the survival of scholarly traditions in a relatively bleak and rude age, there was nothing like the general advance in education that occurred later with the cultural awakening of the 11th and 12th centuries.

Educational reforms of Charlemagne

Learning, nonetheless, had no more ardent friend than Charlemagne, who came to the Frankish throne in 768 distressed to find such poor standards of Latin in his correspondence, the offenders including even bishops and abbots from all parts of his empire; he complained that their letters, though often very correct in sentiment, were very incorrect in grammar. He thus ordered that the clergy be educated severely, whether by persuasion or under compulsion. He recalled that, in order to inter-

pret the Holy Scriptures, one must have a command of correct language and a fluent knowledge of the Latin tongue; he later commanded, "in each bishopric and in each monastery let the psalms, the notes, the chant, calculation and grammar be taught and carefully corrected books be available" (capitulary of AD 789). His promotion of ecclesiastical and educational reform bore fruit in a generation of churchmen whose morals and whose education were of a higher standard than before.

The possibility then arose of providing, for the brighter young clerics and perhaps also for a few laymen, a more advanced religious and academic training. It was perhaps to meet this modest need that a school grew up within the precincts of the Emperor's palace at Aachen. In order to develop and staff other centres of culture and learning, Charlemagne imported considerable foreign talent. During the 8th century, England had been the scene of some intellectual activity; thus, Alcuin, who had been the master of the school at York, and other English scholars were brought over to transplant to the Continent the studies and disciplines of the Anglo-Saxon schools. From Moorish Spain came Christian refugees who also contributed to this intellectual revival; disputations with the Muslims had forced them to develop a dialectic skill in which they now instructed Charlemagne's subjects. From Italy came grammarians and chroniclers, men such as Paul the Deacon; the more formalistic classical traditions in which they had been bred supplied the framework to discipline the effervescent brilliance of the Anglo-Saxons. Irish scholars also arrived. Thanks to these foreigners, who represented the areas where classical and Christian culture had been maintained in the 6th and 8th centuries, the court became a kind of "academy," to use Alcuin's term. There the emperor, his heirs, and his friends discussed various subjects—the existence or nonexistence of the underworld and of nothingness, the eclipse of the Sun, the relationship of Father, Son, and Holy Ghost, and so on. Recognizing the importance of manuscripts in the cultural revival, Charlemagne formed a library (the catalog of which is still extant), had texts and books copied and recopied, and bade every school to maintain a scriptorium. Alcuin developed a school of calligraphy at Tours, and its new script spread rapidly throughout the empire; this Carolingian minuscule was more legible and less wasteful of space than the scripts hitherto employed.

Outside the court at Aachen were to be found here and there a few seats of culture, but not many. The Archbishop of Lyon reorganized the schools of readers and choir leaders; Alcuin in Saint-Martin-de-Tours and Angilbert in Saint-Riquier organized monastic schools with relatively well-stocked libraries. It was necessary to wait for the second generation or even the third to witness the greatest brilliance of the Carolingian renewal. Under Charlemagne's son Louis the Pious and especially under his grandsons, the monastic schools reached their apogee in France north of the Loire, in Germany, and in Italy. The most famous were at Saint-Gall, Reichenau, Fulda, Bobbio, Saint-Denis, Saint-Martin-de-Tours, and Ferrières. Unfortunately, the breakup of the Carolingian Empire, following local rebellions and the Viking invasions, ended the progress of the Carolingian renaissance.

Educational accomplishments of Alfred the Great

Influences of the Carolingian renaissance abroad. In England, at least in the kingdom of Wessex, King Alfred the Great stands out as another royal patron of learning, one who wanted to imitate the creativity of Charlemagne. When he came to the throne in 871, cultural standards had fallen to a low level, partly because of the turmoil of the Danish invasions. He was grieved to find so few who could understand their church services in English or translate a letter from Latin into English. To accomplish an improvement, he called upon monks from the Continent, particularly those of Saint-Bertin. Moreover, he attracted to his court certain English clergy and young sons of nobles. Since the latter did not know Latin, he had translated into Wessex English some works of Pope Gregory the Great, Boethius, the theologian and historian Paulus Orosius, Venerable Bede, St. Augustine, and others. He himself translated Boethius' *De consolatione philosophiae*, Gregory the Great's *Cura pastoralis*, and Bede's *Ecclesiastical History of the English People*. This promotion of learning was continued by Alfred's successors and spread elsewhere in England; and in the reformed monasteries at Canterbury, York, and Winchester, the young monks renewed the study of religious and secular sciences. Among the master scholars of the late 10th century was the Benedictine monk Aelfric, perhaps the greatest prose writer of Anglo-Saxon times. In order to facilitate the learning of Latin for young monks, Aelfric composed a grammar, glossary and colloquy, containing a Latin grammar described in Anglo-Saxon, a glossary in which master and pupil could find a methodically classified Latin vocabulary (names of birds, fish, plants, and so forth), and a manual of conversation, inspired by the bilingual manuals of antiquity.

Among the other Saxons, those of the Continent who presided over the destinies of Germany, there were also significant gatherings of masters and students at selected monasteries, such as Corvey and Gandersheim. In any case, wherever teaching became important in the 10th century, it concentrated largely on grammar and the works of the classical authors. Thus when Gerbert of Aurillac, after a course of instruction in Catalonia, came to teach dialectic and the arts of the quadrivium (geometry, arithmetic, music, and astronomy) at Reims, he aroused astonishment and admiration. His renown helped in his later election as Pope Sylvester II. The first half of the 11th century contained the first glimmerings of a rediscovered dialectic. A new stage in the history of teaching was beginning.

Religious and military emphases of general education

Education of the laity in the 9th and 10th centuries. The clergy who dominated society thought it necessary to give laymen some directives about life comparable to those offered in monastic rules and thus issued what were called *miroirs* ("mirrors"), setting forth the duties of a good sovereign and exalting the Christian struggle. Already the image of the courtly and Christian knight was beginning to take shape. It was not a question of governing a state well but, rather, of governing oneself. The layman must struggle against vice and practice virtue; he must emphasize his religious heritage. Alcuin became indignant when he heard it said that the reading of the Gospel was the duty of the clergy and not that of the layman. Huoda, wife of Bernard, duke of Septimania, addressed a manual to her 16-year-old son, stressing the reading and praying that a young man should do. In the libraries of the laity, the volumes of the Old and New Testaments took first place, along with prayer books, a kind of breviary designed for day-to-day use.

If a minority of aristocrats could receive a suitable moral and religious education, the masses remained illiterate and preferred a military apprenticeship to study. "He who has remained in school up to twelve years without mounting a horse is no longer good for anything but the priesthood," wrote a German poet. Writers of hagiographic texts were fond of contrasting the mother of the future saint, anxious to give education to her son, and the father, who wanted to harden his son at an early age to the chase or to war. The Carolingian tradition, however, was not totally forgotten by princes and others in high places. In Germany, Otto I and his successors, who wished to recreate the Carolingian Empire, encouraged studies at the court; Wipo, the preceptor of Henry III, set out a program of education for the laity in his *Proverbia*. Rediscovering the ancient moralists, he praised moderation as opposed to warlike brutality or even the ascetic strength of the monks. The same tendency is found in other writings. The rediscovery of the ancient moralists Cicero and Seneca inevitably made itself felt in the program of education that the clergy of the 11th century began to offer the laity.

THE MEDIEVAL RENAISSANCE

The era that has been called the "renaissance of the 12th century" corresponds to a rediscovery of studies originating in the 11th century in a West in the process of transformation. The church cast off the tutelage of lay power, and there was general acceptance of the authority

of the church in matters of belief, conduct, and education; the papacy took over the direction of Christianity and organized the Crusades to the East; the monarchies regrouped the political and economic forces of feudal society; the cities were reanimated and were organized into communes; the merchants traced out the great European trade routes and, before long, the Mediterranean ones. Soon contact with the East, by trade and in the Crusades, and with the highly cultivated Moors in Spain further stimulated intellectual life. Arabic renderings of some of the works of Aristotle, together with commentaries, were translated into Latin, exercising a profound influence on the trend of culture. It was inevitable that the world of education would take on a new appearance.

Changes in the early schools. In the first place, the monastic reformers made the decision to close their schools to those who did not intend to enter upon a cloistered life. According to their idea of solitude and sanctity, recalling the words of St. Jerome, "the monk was not made to teach but to mortify himself." Divine works were to be the only object of study and meditation, and Pierre de Celle asserted that "divine science ought to mould rather than question, to nourish conscience rather than knowledge."

The scholarly monks completed their studies before being admitted to the monastery—the age of entrance, for instance, being fixed at 15 years at Cîteaux and 20 years at Cluny. If there were admitted a few oblates (laymen living in monasteries under modified rules), they were given an ascetic and moral education and were taught to read the Holy Writ and, what was still more desirable, to "relish" it. In the Carthusian monastery the four steps of spiritual exercise were reading, meditation, prayer, and contemplation. Thus, there existed a monastic culture, but there were no truly monastic studies such as existed in the 9th and 10th centuries. The rich libraries of the monasteries served only a few scholarly abbots, while the monks searched for God through prayer and asceticism.

In the cities, on the contrary, the schools offered to all the clergy who so desired the means of satisfying their intellectual appetite. More and more of them attended these schools, for the studies were a good means of social advancement or material profit. The development of royal and municipal administrations offered the clergy new occupations. Hence the success of the schools for notaries and the schools of law and rhetoric. These schools were organized under the protection of the collegiate churches and the cathedrals. The schools for secular subjects were directed by an archdeacon, chancellor, cantor, or cleric who had received the title of *scholasticus*, *caput scholae*, or *magister scholarum* and who was assisted by one or more auxiliary masters. The success of the urban schools was such that it was necessary, in the middle of the 12th century, to define the teaching function. Only those could teach who were provided with the *licencia docendi* conferred by the bishop or, more often, by the *scholasticus*. Those who were licensed taught within the limits of the city or the diocese, whose clerical leaders supervised this monopoly and intervened if a cleric set himself up as master without having the right. The popes were sufficiently concerned about licensing that the Lateran Council of 1179 gave this institution universal application.

Study of the liberal arts and philosophy

The pupils who attended these urban schools learned in them their future occupation as clerics; they learned Latin, learned to sing the service, and were initiated into Holy Writ. The more gifted ones extended their studies further and applied for admission to the liberal arts (the trivium, made up of grammar, rhetoric, and logic; and the quadrivium, including geometry, arithmetic, music, and astronomy) and, upon completion of the liberal arts, to philosophy. Philosophy had four branches: theoretical, practical, logical, and mechanical. The theoretical was divided into theology, physics, and mathematics; the practical consisted of morals or ethics (personal, economic, political). The logical, which concerned discourse, consisted of the three arts of the trivium. Finally, the mechanical included the work of processing wool, of navigation, of agriculture, of medicine, and so on. This was an ambitious humanistic program. In fact, the students became specialized in the study of one art or another according to their tastes or the presence of a renowned master, such as Guillaume de Champeaux at Paris and St. Victor for rhetoric and theology; Peter Abelard at Paris for dialectic and theology; Bernard de Chartres for grammar; William of Conches at Chartres for grammar, ethics, and medicine; Thierry de Chartres for rhetoric. In particular teachers of the "literary" arts, grammar and rhetoric, always had great success in a period of enthusiasm for the ancient authors. It may be noted that Bernard de Chartres organized his literary teaching in this fashion: grammatical explanations (*declinatio*), studies of authors, and each morning the correction of the exercises given the day before.

The third art of the trivium, logic (or dialectics), was nevertheless a strong competitor of the other two, grammar and rhetoric. Since the 11th century, Aristotle's *Posterior Analytics*, translated centuries earlier by Boethius, had developed the taste for reasoning, and, by the time that Abelard arrived in Paris around 1100, interest in dialectics was flourishing. The written words of the Scriptures and of the Fathers of the Church were to be subjected to the scrutiny of human reason; a healthy skepticism was to be the stepping-stone to knowledge, aided by an understanding of critical logic. While dialectic—or logic—reigned in Paris, the masters at Chartres offered a study of the whole of the quadrivium—arithmetic, music, geometry, and astronomy. This interest in the sciences, which had been manifest at Chartres since the early 11th century, had been favoured by the stimulus of Greco-Arabic translations. The works of Euclid, Ptolemy, Hippocrates, Galen, and other Hellenic and Hellenistic scholars, as preserved in the Arabic manuscripts, were translated in south Italy, Sicily, and Spain and were gradually transmitted north. The scientific revival allowed the Chartrians to Christianize Greek cosmology, to explain Genesis according to physics, and to rediscover nature. Another discovery was that of law. The conflicts in the second half of the 12th century between the church and the lay powers encouraged on both sides a new activity in the juridical field. The princes found in the Corpus Juris Civilis, the 6th-century Roman code of the emperor Justinian, the means of legitimizing their politics, and the papacy likewise used Roman sources to promote its claims.

The development of the universities. At the end of the 12th century, the multiplicity of students and masters, their rivalries, and the conflicts in which they opposed the religious and civil authorities obliged the world of education to reorganize. To understand the reorganization, one must review the various stages of development in the coming together of students and masters. The first stage, already alluded to, occurred when the chancellor of a cathedral or some other authority began to accord to other masters permission to open other schools than the episcopal school in the neighbourhood of his church. A further stage was reached when a license to teach—granted only after a formal examination—empowered a master to carry on his vocation at any similar centre.

The *studia generalia* and the *universitas*

A further development came when it began to be recognized that, without a license from pope, emperor, or king, no *studium generale* could be formed possessing the right of conferring degrees, which originally meant nothing more than licenses to teach.

Students and teachers, as *clerici* ("clerks," or members of the clergy), enjoyed certain privileges and immunities, but, as the numbers travelling to the *studia generalia* increased, they needed additional protection. In 1158 Emperor Frederick I Barbarossa of the Holy Roman Empire granted them privileges such as protection against unjust arrest, trial before their peers, and permission to "dwell in security." These privileges were subsequently extended and included protection against extortion in financial dealings and the *cessatio*, or the right to strike, discontinue lectures, and even to secede to protest against grievances or interference with established rights.

In the north of Europe licenses to teach were granted

by the chancellor, *scholasticus*, or some other officer of a cathedral church; in the south it is probable that the guilds of masters (when these came to be formed) were at first free to grant their own licenses, without any ecclesiastical or other supervision. Gradually, however, toward the end of the 12th century, a few great schools claimed, from the excellence of their teaching, to be of more than merely local importance. In practice, a doctor of Paris or Bologna would be allowed to teach anywhere; and those great schools began to be known as *studia generalia;* that is, places resorted to by scholars from all parts. Eventually the term came to have a more definite and technical significance. The emperor Frederick II in 1225 set the example of attempting to confer upon his new school at Naples, by an authoritative bull, the prestige that the earlier *studia* had acquired by reputation and general consent. Pope Gregory IX did the same for Toulouse in 1229, and he added to its original privileges in 1233 a bull by which anyone who had been admitted to the doctorate or mastership in that university should have the right to teach anywhere without further examination. Other *studia generalia* were subsequently founded by papal or imperial bulls, and in 1292 even the oldest universities, Paris and Bologna, found it desirable to obtain similar bulls from Pope Nicholas IV. From this time the notion began to prevail that the essence of the *studium generale* was the privilege of conferring a universally valid teaching license and that no new *studium* could acquire that position without a papal or imperial bull. There were, however, a few *studia generalia* (such as Oxford) the position of which was too well established to be seriously questioned, although they had never obtained such a bull; these were held to be *studia generalia* by repute. A few Spanish universities founded by royal charter were held to be *studia generalia* for the kingdom.

The word *universitas* originally applied only to the scholastic guild (or guilds)—that is, the corporation of students and masters—within the *studium*, and it was always modified, as *universitas magistrorum*, or *universitas scholarium*, or *universitas magistrorum et scholarium*. In the course of time, however, probably toward the latter part of the 14th century, the term began to be used by itself, with the exclusive meaning of a community of teachers and scholars whose corporate existence had been recognized and sanctioned by civil or ecclesiastical authority.

The Italian universities. The earliest *studia* arose out of efforts to provide instruction beyond the range of the cathedral and monastic schools for the education of priests and monks. Salerno, the first great *studium*, became known as a school of medicine as early as the 9th century, and, under the teaching of Constantine the African (died 1087), its fame spread throughout Europe. In 1231 it was licensed by Frederick II as the only school of medicine in the kingdom of Naples. It remained a medical school only.

The Bologna prototype of universities

The great revival of legal studies that took place at Bologna about the year 1000 had been preceded by a corresponding activity at Pavia and Ravenna. In Bologna a certain Pepo was lecturing on parts of the Corpus Juris Civilis about the year 1076. The secular character of this new study and its close connection with the claims and prerogatives of the Western emperor aroused papal suspicion, and for a time Bologna and its students were regarded by the church with distrust. The students found their first real protector in the emperor Frederick I Barbarossa. The immunities and privileges he conferred eventually extended to all the other universities of Italy.

The first university of Bologna was not constituted until the close of the 11th century—the "universities" there being student guilds, formed to obtain by combination that protection and those rights that they could not claim as citizens. As the number of students increased, the number of *universitates*, or societies of scholars, increased, each representing the national origin of its members (France, England, Provence, Spain, Italy). The practice at Bologna was adopted as other *studia generalia* arose.

These confederations were presided over by a common head, the *rector scholarium*, and the different nations were represented by their *consiliarii*, a deliberative assembly with which the rector habitually took counsel. The students at Bologna were mostly of mature years. As the civil law and the canon law were, at first, the only branches of study, the class they attracted was often composed of men already filling office in some department of the church or state—archdeacons, heads of schools, canons of cathedrals, and like functionaries.

About 1200 the two faculties of medicine and philosophy were formed. The former was developed by a succession of able teachers, among whom Thaddeus Alderottus was especially eminent. The faculty of arts, down to the 14th century, scarcely attained equal eminence.

The term college at Bologna long had a different meaning from the ordinary modern one. The masters formed themselves into *collegia* (that is, organizations), chiefly for the conferment of degrees. Places of residence for students existed at Bologna at a very early date, but it was not until the 14th century that they possessed any organization; the humble *domus*, as it was termed, was at first designed solely for necessitous students who were not natives of Bologna; a separate house, with a fund for the maintenance of a specified number of scholars, was all that was originally contemplated.

From the 13th to the 15th century a number of universities in Italy originated from migrations of students; others were established by papal or other charters. Almost all the schools taught civil or canon law or both. Of these institutions the most important were Padua, Piacenza, Pavia, Rome, Perugia, Pisa, Florence, Siena, and Turin.

The Paris prototype of universities

The French universities. The history of the University of Paris well illustrates the fact that the universities arose in response to new needs. The schools out of which the university arose were those attached to the cathedral on the Île de la Cité and presided over by its chancellor. Although, in the second decade of the 13th century, some masters placed themselves under the jurisdiction of the abbot of the monastery of Sainte-Geneviève on the Left Bank of the Seine, it was around the bestowal of the license by the chancellor of Notre Dame that the university grew up. It is in this license that the whole significance of the master of arts degree was contained; for admission to that degree was the receiving of the chancellor's permission to "incept"; and by "inception" was implied the master's formal entrance upon the functions of a duly licensed teacher and his recognition as such by his brethren in the profession. The stage of bachelordom had been one of apprenticeship for the mastership; and his emancipation from this state was symbolized by the placing of the magisterial cap (biretta) upon his head. The new master gave a formal inaugural lecture, and he was then welcomed into the society of his professional brethren with set speeches and took his seat in his master's chair.

Some time between 1150 and 1170 the University of Paris came formally into being. Its first written statutes were not, however, compiled until about 1208, and it was not until long after that date that it possessed a "rector." Its earliest recognition as a legal corporation belongs to about the year 1211, when a brief of Innocent III empowered it to elect a proctor to be its representative at the papal court. With papal support Paris became the great trans-Alpine centre of orthodox theological teaching. Successive pontiffs, down to the Great Schism of 1378, cultivated friendly relations with the university and systematically discouraged the formation of theological faculties at other centres. In 1231 Gregory IX, in the bull *Parens Scientiarum* ("Mother of Learning"), gave full recognition to the right of the several faculties to regulate and modify the constitution of the university. The fully developed university was divided into four faculties: three superior, those of theology, canon law, and medicine; and one inferior, that of arts, which was divided into four nations (French, Picard, Norman, and English), which included both professors and scholars from the respective countries. The head of each faculty

was the dean; of each nation, the proctor. The rector, in the first instance head of the faculty of arts, eventually became the head of the collective university.

After the close of the Middle Ages, Paris came to be virtually reduced to a federation of colleges, though at Paris the colleges were less independent of university authority than was often the case elsewhere. Other major French universities of the Middle Ages were Montpellier, Toulouse, Orléans, Angers, Avignon, Cahors, Grenoble, Orange, and Perpignan.

The English offshoots of Paris: Oxford and Cambridge

The English universities. The University of Paris became the model for French universities north of the Loire and for those of central Europe and England; Oxford would appear to have been the earliest. Certain schools, opened early in the 12th century within the precincts of the dissolved nunnery of St. Frideswide and of Oseney Abbey, are supposed to have been the nucleus around which it grew. But the beginning may have been a migration of English students from Paris about 1167 or 1168. Immediately after 1168, allusions to Oxford as a *studium* and a *studium generale* begin to multiply. In the 13th century, mention first occurs of university "chests," which were benefactions designed for the assistance of poor students. Halls, or places of licensed residence for students, also began to be established. Against periodic vicissitudes such as student dispersions and plagues, the foundation of colleges proved the most effectual remedy. The earliest colleges were University College, founded in 1249, Balliol College, founded about 1263, and Merton College, founded in 1264.

The University of Cambridge, although it came into existence somewhat later than Oxford, may reasonably be held to have had its origin in the same century. In 1112 the canons of St. Giles crossed the River Cam and took up their residence in the new priory in Barnwell, and their work of instruction acquired additional importance. In 1209 a body of students migrated there from Oxford. Then about 1224 the Franciscans established themselves in the town and, somewhat less than half a century later, were followed by the Dominicans. At both the English universities, as at Paris, the mendicants and other religious orders were admitted to degrees, a privilege that, until the year 1337, was extended to them at no other university. Their interest in and influence at these three centres were consequently proportionately great.

In 1231 and 1233 royal and papal letters afford satisfactory proof that the University of Cambridge was already an organized body, with a chancellor at its head.

Although both Oxford and Cambridge were modelled on Paris, their higher faculties never developed the same distinct organization; and, while the two proctors at Cambridge originally represented north and south, the nations are scarcely to be discerned. An important step was made, however, in 1276, when an ordinance was passed requiring that everyone who claimed to be recognized as a scholar should have a fixed master within 15 days after his entry into the university. The traditional constitution of the English universities was, in its origin, an imitation of the Parisian, modified by the absence of the cathedral chancellor. But the feature that most served to give permanence and cohesion to the entire community at Cambridge was, as at Oxford, the institution of colleges. The earliest of these was Peterhouse, in 1284. All the early colleges were expressly designed for the benefit of the secular clergy.

Universities elsewhere in Europe. From the 13th to the 15th centuries, *studia generalia* or universities proliferated in central and northern Europe and were usually modelled on the University of Paris. Although the earliest was Prague, which existed as a *studium* in the 13th century and was chartered by Pope Clement VI in 1348, perhaps no medieval university achieved a more rapid and permanent success than Heidelberg. The University of Heidelberg, the oldest in the German realm, received its charter in 1386 from Pope Urban VI as a *studium generale* and contained all the recognized faculties—theology, canon law, medicine, and the arts, as well as civil law. In the subsequent 100 years, universities were founded at Cologne, Erfurt, Leipzig, Rostock, Freiburg, Tübingen, Ofen (Budapest), Basel, Uppsala, and Copenhagen.

Higher education in Spain

Spain was also an important scene of developments in higher education. Valladolid received its charter in 1346 and attained great celebrity after it obtained the rank of *studium generale* and a *universitas theologiae* by a decree of Pope Martin V in 1418. Salamanca was founded in 1243 by Ferdinand III of Castile with faculties of arts, medicine, and jurisprudence, to which theology was added through the efforts of Martin V. The College of St. Bartholomew, the earliest founded at Salamanca, was noted for its ancient library and valuable collection of manuscripts. Other important early Spanish and Portuguese schools were Seville, Alcalá, and Lisbon.

General characteristics of medieval universities. Generally speaking, the medieval universities were conservative. Alexander Hegius and Rudolf Agricola carried on their work as reformers at places such as Deventer, in the Netherlands, remote from university influences. A considerable amount of mental activity went on in the universities; but it was mostly of the kind that, while giving rise to endless controversy, turned upon questions in connection with which the implied postulates and the terminology employed rendered all scientific investigation hopeless. At almost every university the Realists and Nominalists represented two great parties occupied with an internecine struggle (see EPISTEMOLOGY).

In Italian universities such controversies were considered endless and their effects pernicious. It was resolved, accordingly, to expel logic and allow its place to be filled by rhetoric, thereby effecting that important revolution in academic studies that constituted a new era in university learning and largely helped to pave the way for the Renaissance. The professorial body in the great Italian universities attained an almost unrivalled reputation throughout Europe. For each subject of importance there were always two, and sometimes three, rival chairs. While other universities became sectarian and local, those of Italy continued to be universal, and foreigners of all nations could be found among the professors.

Student life and studies

The material life of the students was difficult. In order to aid the poorest, some colleges founded by clerical or lay benefactors offered board and lodging to a number of foundationers. Courses, too, could occasionally be difficult. The courses in theology were particularly long —eight years at the minimum (one could not be a teacher of theology in Paris before the age of 35). Many students preferred the more rapid and more lucrative paths of law and medicine. Others led the life of perpetual students, of vagabond clerics, disputatious goliards, the objects of repeated but ineffectual condemnation.

The methods of teaching are particularly well-known in the case of Paris. The university year was divided into two terms: from St. Remi (October 1) to Lent and from Easter to St. Pierre (June 29). The courses consisted of lectures (*collatio*), but more often of explications of texts (*lectio*). There were also discussions and question periods. Examinations were given at the end of each term. The student could receive three degrees: the *determinatio*, or baccalaureate, gave him the right to teach under the supervision of a master; the *licencia docendi* was literally the "license to teach" and could be obtained at 21 years of age; then there was the doctorate, which marked his entrance into mastership and which involved a public examination.

Lay education and the lower schools. The founding of universities was naturally accompanied by a corresponding increase in schools of various kinds. In most parts of western Europe, there were soon grammar schools of some type available for boys. Not only were there grammar schools at cathedrals and collegiate churches, but many others were founded in connection with chantries and craft and merchant guilds and a few in connection with hospitals. It has been estimated, for example, that, toward the close of the Middle Ages, there were in England and Wales, for a population of about 2,500,000, approximately 400 grammar schools, rather better provision than in Victorian days of the 19th century. Knowl-

Early grammar schools

edge of the teaching provided in the grammar schools at this period is too slight to justify an attempt at a description. No doubt the curriculum varied, but religion was all-important, with Latin as a written and spoken language the other major element in the timetable. There might have been instruction in reading and writing in the vernacular, but, in addition to the grammar schools, there were writing and song schools and other schools of an elementary type. Elementary teaching was given in many churches and priests' houses, and children who did not receive formal scholastic instruction were given oral teaching by parish priests in the doctrines and duties of the faith. The evidence of accounts, bills, inventories, and the like suggests that there was some careful teaching of writing and of an arithmetic that covered the practical calculations required in ordinary life.

Girls were sometimes taught in nunneries, but usually in the home, and those of the upper classes, at least, generally learned to read, write, and keep accounts, to perform household duties, and to obtain such elementary medical knowledge as might be helpful when called on to treat minor accidents or ailments. For young and grown-up alike, there was much that was educative in medieval society, and the influence of the community life of church, guild, and village was an important factor. Medieval education had many shortcomings, and its benefits were too often confined to a few; but on the whole the mass of the people were by no means sunk in brutish ignorance. The number of people who could read the vernacular, as evidenced by the demand for books in the vulgar tongue as soon as printing made them available, is proof that the latter part of the Middle Ages was certainly not a time of general illiteracy.

In fulfillment of its responsibility for education, the church from the 11th century onward made the establishment of an effective education system a central feature of ecclesiastical policy. During the papacy of Gregory VII (1073–85), all bishops had been asked to see that the art of grammar was taught in their churches, and a Lateran Council in 1215 decreed that grammar school masters should be appointed not only in the cathedral church but also in others that could afford it. Solicitude at the centre for the advancement of education did not, however, result in centralized administration. It was the duty of bishops to carry out approved policy, but it was left to them to administer it, and they in turn allowed schools a large measure of autonomy. Such freedom as medieval schools enjoyed was, however, always subject to the absolute authority of the church, and the right to teach, as earlier noted, was restricted to those who held a bishop's license. This device was used to ensure that all teachers were loyal to the doctrines of the church.

Training for feudalism

In medieval society there was an entirely different training for boys of high rank, and this created a cultural cleavage. Instead of attending the grammar school and proceeding to a university, these boys served as pages and then as squires in the halls and castles of the nobility, there receiving prolonged instruction in chivalry. The training was designed to fit the noble youth to become a worthy knight, a just and prudent master, and a sensible manager of an estate. Much of this knowledge was gained from daily experience in the household, but, in addition, the page received direct instruction in reading and writing, courtly pastimes such as chess and playing the lute, singing and making verses, the rules and usages of courtesy, and the knightly conception of duty. As a squire he practiced more assiduously the knightly exercises of war and peace and acquired useful experience in leadership by managing large and small bodies of men. But this was a type of education that could flourish only in a feudal society; and, though some of its ideals survived, it was outmoded when feudalism was undermined by the growth of national feeling. (P.R./Ed.)

V. Education in Asian civilizations: c. 700 to the eve of Western influence

INDIA

During its medieval period, India was ruled by dynasties of Muslim culture and religion. Muslims from Arabia first appeared in the country in the 8th century, but the foundation of their rule was laid much later by Muḥammad Ghūri, who established his power at Delhi in 1192. The original Muslim rule was replaced successively by that of the Muslim Pashtuns and Mughals.

The foundations of Muslim education. Muslim educational institutions were of two types—a *maktab*, or elementary school, and a *madrasah*, or institution of higher learning. The content of education imparted in these schools was not the same throughout the country. It was, however, necessary for every Muslim boy at least to attend a *maktab* and to learn the necessary portions of the Qur'ān required for daily prayers. The curriculum in the *madrasah* comprised Ḥadīth (the study of Muslim traditions), jurisprudence, literature, logic and philosophy, and prosody. Later on, the scope of the curriculum was widened, and such subjects as history, economics, mathematics, astronomy, and even medicine and agriculture were added. Generally, all the subjects were not taught in every institution. Selected *madrasas* imparted postgraduate instruction, and a number of towns—Āgra, Badaun, Bīdar, Delhi, Jaunpur, and a few others—developed into university centres to which students flocked from far and wide for study under renowned scholars. The sultans and *amīrs* of Delhi and the Muslim rulers and nobles in the provinces also extended patronage to Persian scholars who came to their courts from other parts of Asia under the pressure of Mongol inroads. Delhi vied with Baghdad and Cordova as an important centre of Islāmic culture. Indian languages also received some attention. The Muslim rulers of Bengal, for example, engaged scholars to translate the Hindu classics, the *Rāmāyaṇa* and the *Mahābhārata*, into Bengali.

Developments under the Pathans

Under the Pathan Lodis, a dynasty of Afghan foreigners (1451–1526), the education of the Hindus was not only neglected but was often adversely affected in newly conquered territories. The rulers generally tolerated Sanskrit and vernacular schools already in existence but did not help the existing ones with money or build new ones. At early stages, the *maktabs* and *madrasas* were attended by Muslims only. Later, when Hindus were allowed into high administrative positions, Hindu children began to receive Persian education in Muslim schools.

The Mughal period. The credit for organizing education on a systematic basis goes to Akbar the Great (lived 1542–1605), a contemporary of Queen Elizabeth I of England and undoubtedly the greatest of Mughal emperors. He treated all his subjects alike and opened a large number of schools and colleges for Muslims as well as for Hindus throughout his empire. He also introduced a few curricular changes, based on students' individual needs and the practical necessities of life. The advanced curriculum included ethics, arithmetic, astronomy, geometry, agriculture, economics, public administration, natural philosophy, abstract mathematics, divinity, and history. The scope of the curriculum was so widened as to enable every student to receive education according to his religion and views of life. The adoption of Persian as the court language gave further encouragement to the Hindus and the Muslims to study Persian.

Akbar's policy was continued by his successors Jahāngīr and Shāh Jahān. But his great-grandson Aurangzeb (1618–1707) changed his policy with regard to the education of the Hindus. In April 1669, for instance, he ordered the provincial governors to destroy Hindu schools and temples within their jurisdiction; and, at the same time, he supported Muslim education with a certain religious fanaticism. After his death, the glory of the Mughal Empire began gradually to vanish, and the whole country was overrun by warlords.

During the Mughal period, girls received their education at home or in the house of some teacher living in close proximity. There were special arrangements for the education of the ladies of the royal household, and some of the princesses were distinguished scholars. Vocational education was imparted through a system of apprenticeship either in the house of *ustāds* (teachers) or in *kārkhānahs* (manufacturing centres).

Muslim rulers of India were also great patrons of litera-

ture and gave considerable impetus to its development. Akbar ordered various Hindu classics and histories translated into Persian. In addition, a number of Greek and Arabic works were translated into Persian. Literary activities did not entirely cease even in the troubled days of later rulers. Men of letters were patronized by such emperors as Bahādur Shāh and Muḥammad Shāh and by various regional officials and landlords.

Such is the history of Muslim education in India. It resembles ancient Indian education to a great extent: instruction was free; the relation between the teachers and the taught was cordial; there were great centres of learning; the monitorial system was used; and people were preoccupied with theology and the conduct of life.

Educational accomplishments of the Muslim period

There were, however, several distinctive features of Muslim education. First, education was democratized. As in mosques, so in a *maktab* or *madrasah*, all were equal, and the principle was established that the poor should also be educated. Second, Muslim rule influenced the system of elementary education of the Hindus, which had to accommodate itself to changed circumstances by adopting a new method of teaching and by using textbooks full of Persian terms and references to Muslim usages. Third, the Muslim period brought in many cultural influences from abroad. The courses of studies were both widened and brought under a humanistic influence. Finally, Muslim rule produced a cross-cultural influence in the country through the establishment of an educational system in which Hindus and Muslims could study side by side and in which there would be compulsory education in Persian, cultivation of Sanskrit and Hindi, and translation of great classics of literature into different languages. Ultimately, it led to the development of a common medium of expression, Urdu.

Education in the Muslim era was not a concerted and planned activity but a voluntary and spontaneous growth. There was no separate administration of education, and state aid was sporadic and unsteady. Education was supported by charitable endowments and by lavish provision for the students in a *madrasah* or in a monastery.

The Muslim system, however, proved ultimately harmful. In the early stages genuine love of learning attracted students to the cultural centres, but later on "the bees that flocked there were pre-eminently drones." The whole system became stagnant and stereotyped as soon as cultural communication was cut off from the outside world because of political disturbances and internecine wars. The Indian teachers were reduced to dependence on their own resources, and a hardening tradition that became increasingly unreceptive to new ideas reduced the whole process to mere routine. (S.N.M.)

CHINA

The T'ang dynasty (AD 618–907). The T'ang was one of China's greatest dynasties, marked by military power, political stability, economic prosperity, and advance in art, literature, and education. It was an age in which Buddhist scholarship won recognition and respect for its originality and high intellectual quality, in which the rise of new sects marked the final integration of Buddhism into Chinese life, and in which China superseded India as the land from which Buddhism was to spread to other countries in East Asia.

The T'ang was known for its literature and art and has been called the Golden Age of Chinese poetry. There were thousands of poets of note who left a cultural legacy unsurpassed in subsequent periods and even in other lands. There also flourished prose writers, whose style was studied by young students all the way down to the 20th century, and artists whose paintings reflected the influences of Buddhism and Taoism.

One of the greatest gifts of China to the world was the invention of printing. Block printing was invented in the 8th century and movable type in the 11th century. The first book printed from blocks was a Buddhist *sūtra*, or set of precepts, in 868. Printing met the demand created by the increase in the output of literature and by the regularized civil-service-examination system. It also met the popular demand for Buddhist and Taoist prayers and charms. One historian (Kenneth Scott Latourette) noted that "as late as the close of the eighteenth century the [Chinese] Empire possibly contained more printed books than all the rest of the world put together."

Education in the T'ang dynasty was under the dominant influence of Confucianism, notwithstanding the fact that Buddhism and Taoism both received imperial favours. An academic-examination system was firmly established, and officials were selected on the basis of civil-service examinations. But Confucianism did not dominate to the extent of excluding other schools of thought and scholarship. Renowned scholars were known to spurn public office because they were not satisfied with the identification of public service with a narrow interpretation of Confucianism. Artists and poets were, in general, rebellious against traditional Confucianism.

An emperor in the 5th century ordered the establishment of a "School of Occult Studies" along with the more commonly accepted schools of Confucian learning. It was devoted to the study of Buddhism and Taoism and occult subjects that transcended the practical affairs of government and society. Such schools were often carried on by the private effort of scholars who served as tutors for interested followers.

School structure in the T'ang era

The schools of T'ang were well organized and systematized. There were schools under the central government, others under local management, and private schools of different kinds. Public schools were maintained in each prefecture, district, town, and village. In the capital were "colleges" of mathematics, law, and calligraphy, as well as those for classical study. There was also a medical school.

Semiprivate schools formed by famous scholars gave lectures and tutelage to students numbering in the hundreds and continued to supplement the public schools. Students from Korea and Japan came to study in China and took back the lunar calendar and the Buddhist sects, as well as the examination system and the Confucian theories of government and social life. Chinese culture also penetrated Indochina.

The examination system was at this time given the form that remained essentially unchanged until the 20th century. Examinations were held on different levels, and for each a corresponding academic degree was specified. Interestingly enough, there was provision for three degrees, not unlike the bachelor's, master's, and doctor's degrees of modern times. The first degree was the *hsiu ts'ai* ("cultivated talent"), the second the *ming ching* ("understanding the classics"), and the third the *chin-shih* ("advanced scholar"). The name of the second degree was in later periods changed to *chü jen* ("recommended man"). An academy of scholars later known as the Hanlin Academy was established for select scholars whom the emperor could call upon for advice and expert opinion on various subjects. Membership in this institution became the highest honour that could be conferred upon those who passed the *chin-shih* degree with distinction. To be appointed a Hanlin scholar was to be recognized as one of the top scholars of the land. Among the services that they rendered were the administration and supervision of examinations and the explanation of difficult texts in literature, classics, and philosophy.

Examinations were given for students of medicine and for military degrees. The study of medicine included acupuncture, massage, and exorcism, as well as the treatment of general diseases of the body and those of eye, ear, throat, and teeth.

The Sung (960–1279). The Sung was another dynasty of cultural brilliance. Landscape painting approached perfection, and cultural achievement was stimulated by the invention of movable type (first made of earthenware, then of wood and metal). This advance from the older method of block printing led to the multiplication of books; the printing of a complete set of the classics was a boon to literary studies in schools.

The rulers of Sung were receptive to new ideas and innovative policies. The outstanding innovator of the dynasty was Wang An-shih, prime minister from 1068 to 1076. He introduced a comprehensive program of re-

form that included important changes in education; more emphasis was subsequently placed on the study of current problems and political economy.

Educational controversy in the Sung era

Wang's reforms met with opposition from conservatives. The controversy was only a phase of a deeper and more far-reaching intellectual debate that made the philosophical contributions of the Sung scholars as significant as those of the Hundred Schools in the Chou dynasty over a millennium earlier. It involved the interpretation of the classics and a re-examination of China's traditional thought. Confucianism and the dominant mode of Chinese thinking had been subject to the challenge of ideas from Legalism, Taoism, and Buddhism, and, despite the resistance of conservatives, the traditional views had to be modified. Outstanding Confucian scholars of conservative bent argued vigorously with aggressive proponents of new concepts of man, of knowledge, and of the universe. The result was Neo-Confucianism, or what some prefer to call rational philosophy. It comprised a reinterpretation of the past, a new formulation of thought, and new directions in scholarship in response to new contacts, fresh ideas, and a bold presentation of opposing views and proposals. The most eminent Neo-Confucianist was Chu Hsi, a Confucian scholar who had studied Taoism and Buddhism. His genius lay in his ability to synthesize ideas from a fresh point of view. Sung scholars distinguished themselves in other fields, too. Ssuma Kuang's *Tzu-chih t'ung-chien* ("Comprehensive Mirror for Aid in Government") was a history of China from the 5th century BC to the 10th century AD. The result of 20 years of painstaking research, it consisted of 1,000 chapters prepared under imperial direction. A volume on architecture was produced that is still used today as a basic reference work, and a treatise on botany contained the most ancient record of varieties of citrus fruits then known in China. No less worthy of mention is an encyclopaedia titled *T'ai-ping yü lan*.

The general pattern of the school system remained essentially the same, with provision for lower schools, higher schools, and technical schools. There was an increase in nonliterary studies and a consequent broadening of the curriculum. A noteworthy development was the rise of a semiprivate institution known as the *shu-yüan*, or academy. With financial support coming from both state grants and private contributions, these academies were managed by noted scholars of the day and attracted many students and lecturers who discoursed on the new ideas of rational philosophy and on subjects not taught in the state schools. Often located in mountain retreats or in the woods, they symbolized the influence of Taoism and Buddhism and a desire to pursue quiet study far away from possible government interference.

The Mongol period (1279–1368). The Mongols were ferocious fighters but inept administrators. Distrustful of the Chinese, they enlisted the services of many nationalities and employed non-Chinese aliens in order to avoid too much dependence on the Chinese. To facilitate the employment of non-Chinese aliens, the civil-service examinations were suspended for a number of years. Later, when a modified form of examinations was in effect, there were special examinations for Mongol candidates to make sure of their admission into high offices.

The Mongols despised the Chinese and placed many limitations on them. Consequently, an aftermath of Mongol rule was a strong antiforeign reaction on the part of the Chinese. The antiforeign reaction was accompanied by an overanxious desire to preserve the Chinese heritage, and the long-range result was an increasing formalism and conservatism that contrasted sharply with the creativity of the Han, the T'ang, and the Sung dynasties.

Despite the setback in Chinese culture under Mongol rule, the period was not devoid of positive cultural development. The increase in foreign contacts as a result of travel to and from China brought new ideas and new knowledge of other lands and other peoples. Mathematics and medicine were further influenced by new ideas from abroad. The classics were translated into the Mongol language, and the Mongol language was taught in schools.

Private schools and the academies of the Sung dynasty became more popular. As a result of a decrease in opportunities for government appointment, scholars were no longer concentrated in the capital city and withdrew into the provinces for study and tutoring. Relieved of the pressure of preparing for the examinations, they applied their talents to the less formal but more popular arts and literary forms, including the drama and the novel. Instead of the classical form, they used the vernacular, or the spoken, language. The significance of this development was not evident until the 20th century, when a "literary revolution" popularized the vernacular tongue.

The Ming period (1368–1644). The Ming dynasty restored Chinese rule. Ming was famous for its ceramics and architecture. There were excellent painters, too, but they were at best the disciples of the T'ang and Sung masters. The outstanding intellectual contribution of the period was the novel. Ming novels continued to be written in the vernacular language and are today recognized as masterpieces of popular vernacular literature. Also of note was the compilation of *Pen-ts'ao kang-mu* ("Great Pharmacopoeia"), a valuable volume on herbs and medicine which was the fruit of 26 years of labour.

The *Yung-lo* encyclopaedia

Of considerable scholarly and educational importance was the *Yung-lo ta-tien* encyclopaedia, which marked a high point in the Chinese encyclopaedic movement. It was a gigantic work resulting from the painstaking efforts of 2,000 scholars over a period of five years. It ran into more than 11,000 volumes, too costly to print, and only two extra copies were made.

The examination system remained basically the same. In the early period of the dynasty, the schools were systematized and regularized. In the latter part of the dynasty, however, the increasing importance of the examination system relegated the schools to a secondary position. The decline of the state-supported schools stimulated the further growth of private education.

The Manchu period (1644–1911). Except for two capable emperors, who ruled for a span of 135 years at the beginning, the Manchu dynasty was weak and undistinguished. Under emperors K'ang-hsi and Ch'ien-lung, learning flourished, but there was little originality. The alien Manchu rulers were so eager to promote Chinese learning that they concentrated on the preservation of what seemed best for stability and the maintenance of the status quo. They ordered new editions of classical and literary works and the preparation of new encyclopaedias. The encyclopaedias, concordances, and dictionaries were the fruits of painstaking labour but were not creative contributions to scholarship.

Distrust of the Chinese by the Manchus and a feeling of insecurity caused the conquerors to erect barriers between themselves and the Chinese. The discriminatory policy was expressed in the administration of the examinations. To assure the appointment of Manchus to government posts, equal quotas were set aside for the Manchus and the Chinese, although the former constituted only about 3 percent of the population. The Chinese thus faced the keenest competition in the examinations, and those who passed tended to be brilliant intellects, whereas the Manchus could be assured of success without great effort.

Schools were encouraged and regulated during the early period of the dynasty. The public school system consisted of schools for nobles, national schools, and provincial schools. Separate schools were maintained for the Manchus, and, for their benefit, Chinese books were translated into the Manchu language. Village and charitable schools were supported by public funds, but they were neglected in later years, so that, by the end of the dynasty, private schools and tutoring had overshadowed them.

At the threshold of the modern era, China had sunk into political weakness and intellectual stagnation. The creativity and originality that had brightened previous periods of history were now absent. Examinations dominated the educational scene, and the content of the examinations was largely literary and classical. Taoism and Buddhism had lost their intellectual vigour, and Confucianism became the unchallenged model of scholarship.

Influence of the examination system

Much could be said for the Chinese examination system at its best. It was instrumental in establishing an intellectual aristocracy whereby the nation could be sure of a cultural unity by entrusting government to scholars reared in a common tradition, nurtured in a common cultural heritage, and dedicated to common ideals of political and social life. It established a tradition of government by civilians and by scholars. It made the scholars the most highly esteemed people of the land. The examinations provided an open road to fame and position. Chinese society was not without classes, but there was a high degree of social mobility, and education provided the opportunity for raising one's position and status. There were no rigid prerequisites and no age limits for taking the examinations. Selection was rigorous, but the examinations were, on the whole, administered with fairness. The names of the candidates did not appear on the examination papers, and the candidates were not permitted to have any outside contacts while writing them.

Nevertheless, the system had serious drawbacks. The content of the examinations became more and more limited in scope. The Confucianist classics constituted the core, and a narrow and rigid interpretation prevailed. In early times Chinese education was broad and liberal, but, by the 19th century, literary ability and knowledge of the classics had become practically the sole concern of the examinations and of education. Art, music, and science had been dropped on the wayside; even arithmetic was not accorded the same importance as reading and writing. Modern science and technology were completely neglected.

Time was when higher criticism was practiced by eminent scholars, but after alien rule by the Mongols the Chinese were so obsessed with the restoration of their heritage that they bent over backward to preserve the past and to avoid deviating from established forms and views. This conservatism was accentuated under Manchu rule and resulted in sterility and stagnation. The creativity and original spirit of classical education was lost. From the standpoint of the new needs of the 19th century, the narrow curriculum was far removed from the pressing problems of changing times. (T.H.C.)

JAPAN

The ancient period to the 12th century. The Japanese nation seems to have formed a unified ancient state in the 4th century AD. Society at that time was composed of *shizoku*, or "clans," each of which served the *chōtei* ("the Imperial court") with its specialized skill or vocation. People sustained themselves by engaging in agriculture, hunting, and fishing, and the chief problem of education was how to convey the knowledge of these activities and provide instruction in the skills useful for these occupations.

Influence of Chinese and Indian civilization

The influence in Japan of the civilizations of China and India had a profound effect on both the spiritual life and the education of the Japanese. Toward the 6th century the assimilation of Chinese civilization became more and more rapid, particularly as a result of the spread of Confucianism. Buddhism was also an important intellectual and spiritual influence. Originating in India and then spreading to China, Buddhism was transmitted to Japan through the Korean peninsula in the mid-6th century.

A monarchic state system with an emperor as its head was established following a coup d'etat in 645. The subsequent Taika (Great Reform) era saw the beginning of many new institutions, most of which were primarily imitations of institutions of the T'ang dynasty of China. In the field of education, a *daigakuryō*, or college house, was established in the capital, and *kokugaku*, or provincial schools, were built in the provinces. The chief aim, however, was to train government officials. The early curriculum was almost identical to that of the T'ang dynasty of China but by the 8th and 9th centuries had been modified considerably to meet internal conditions, particularly as regards the educational needs of the nobility.

Through the Nara and the Heian eras (8th to 12th centuries), the nobility (*kuge*) constituted the ruling class, and learning and culture were the concern primarily of the *kuge* and the Buddhist monks. The nobility lived an artistic life, so that the emphasis of education came to be placed on poetry, music, and calligraphy. Teaching in the *daigakuryō* gradually shifted in emphasis from Confucianism to literature, since the *kuge* set a higher value on artistic refinement than on more spiritual endeavours. Apart from the *daigakuryō*, other institutions were established in which families of influential clans lodged and developed their intellectual lives.

The feudal period (1192-1867). *Education of the warriors.* Toward the mid-12th century political power passed from the nobility to the *buke*, or warrior, class. The ensuing feudal period in Japan dates from the year 1192 (the establishment of the Kamakura shogunate) to 1867 (the decline of the Tokugawa shogunate).

The warrior's way of life was quite unlike that of the nobility, and the aims and content of education in the warrior's society inevitably differed. The warrior had constantly to practice military arts, hardening his body and training his will. Education was based on military training, and a culture characteristic of warriors began to flourish.

Some emphasis, though, was placed on spiritual instruction. The warrior society, founded on firm master–servant relations and centring on the philosophy of Japanese family structure, set the highest value on family reputation and on genealogies. The military arts also proved insufficient to enable warriors to grasp political power and thereby maintain their ruling position. There thus arose a philosophy of *bumbu-kembi*, which asserted the desirability of being proficient in both literary and military arts. The children of warriors attended temples and rigorously trained their minds and wills. Reading and writing were the main subjects.

Temples were the centres of culture and learning and can be said to have been equivalent to universities, in that they provided a meeting place for scholars and students.

Education in the temple, originally aimed at instructing novitiates, gradually changed its character, eventually providing education for children not destined to be monks. Thus, the temples functioned as institutions of primary education.

Education in the Tokugawa era. In 1603 a shogunate was established by a warrior, Tokugawa Ieyasu, in the city of Edo (present Tokyo). The period thence to the year 1867, the Tokugawa, or Edo, era, constitutes the later feudal period in Japan. This era, though also dominated by warriors, differed from former ones in that internal disturbances finally ended, and long-enduring peace ensued. There emerged a merchant class that developed a flourishing commoner's culture. Schools for commoners thus were established.

Temple schools of the Tokugawa period

Representative of such schools were the *terakoya* (temple schools), deriving from the earlier education in the temple. As time passed, some *terakoya* used parts of the houses of commoners as classrooms. As private-run schools designed to provide elementary education, the *terakoya* developed rapidly after the middle years of the Tokugawa era. At the end of the period, such schools flourished not only in the major cities but also in every town and village throughout the country. An increasing number of schools came to teach arithmetic as well as reading and writing and assumed the characteristics of the modern primary school. Pupils counted on a soroban, or abacus, in the arithmetic lesson.

The shogunate supported Confucianism as the highest learning and established shogunate-administered schools. Learning was encouraged among shogunal retainers. *Han*, or feudal clans, following the same policy, built schools in their castle towns for the education of their own retainers.

The warriors' schools were at the apex of the educational system in the Tokugawa era. One of them, the Confucian Academy, was administered directly by the shogunate, becoming a model that the feudal clans throughout Japan followed in establishing their *hankō*, or clan schools. It also functioned as the teachers' training school for *hankō* and survived, until the decline of the shogunate, as the highest seat of learning in the nation.

Feudal clan schools

The heads of the feudal clans invited Confucianists and strategists to their courts, encouraged learning among retainers, and established schools in their castle towns. The *hankō* were directly controlled by the clans and, though few in number at the beginning, gradually spread after about 1750, so that by the end of the era they numbered over 200.

The curriculum in the *hankō* consisted chiefly of *kangaku* (the study of books written in Chinese) and, above all, of Confucianism. Classics of Confucianism, historical works, and anthologies of Chinese poems were used as textbooks. Brush writing, *kokugaku* (study of thought originating in Japan), and medicine were also included. Later, in the last days of the shogunate, *yōgaku* or Western studies and Western medicine were added in several institutions.

Both *hankō* for warriors and *terakoya* for commoners were the typical schools after the middle of the Tokugawa era. Also to be found, however, were *gōgaku*, or provincial schools. They were founded at places of importance in the domain of every clan. Some *gōgaku* were the schools for warriors, but the others were for commoners.

Effect of early Western contacts. The Europeans who first arrived in Japan were the Portuguese, in 1543. In 1549 the Jesuit Francis Xavier visited Japan, and, for the first time, the propagation of Christianity began. Many missionaries began to arrive, Christian schools were built, and European civilization was actively introduced.

In 1633 the shogunate, in apprehension of further Christian infiltration of Japan, banned foreign travel and prohibited the return of overseas Japanese. Further, in 1639, the shogunate banned visits by Europeans. This was the so-called *sakoku*, or period of national isolation. From that time on Christianity was strictly forbidden, and international trade was conducted only with the Chinese and the Dutch.

The *sakoku* policy restricted contact with Europeans to the Dutch only, and, accordingly, Western studies developed as *rangaku*, or learning through the Dutch language.

It is noteworthy that the Tokugawa period laid the foundation of modern Japanese learning. As a result of the development of *hanko* and *terakoya*, Japanese culture and education had developed to such an extent that Japan was able to absorb Western influences and attain modernization at a remarkably rapid pace after the Meiji Restoration of 1868. (A.Na.)

VI. European Renaissance and Reformation

THE CHANNELS OF DEVELOPMENT IN RENAISSANCE EDUCATION

The Arabic influence. The civilization of the Arabs, which flourished from the 7th to the 15th centuries, had a great influence on European Humanism. Early on they had translated works of science and philosophy from Greek, Syriac, ancient Persian, and Hindi. Astronomy came to them from the Greco-Egyptian tradition and mathematics from the Indians, but they then made original contributions to these fields, as well as to medicine, in which their work had far-reaching effects.

In 1056 Constantine the African came from the Muslim city of Carthage in North Africa to Monte Cassino near Rome, where he organized a centre for the translation from Arabic into Latin of significant works in philosophy and the sciences, but particularly in medicine. These texts were then used by the recently established medical school in Salerno, which in turn stimulated research in other European centres, principally in medicine but also in other sciences, thus breaking the scientific monopoly that the Arabs had had for some centuries and paving the way for the new outlook on the world that was later to have its full expression in the Humanistic movement and in the Renaissance.

Humanism

The secular influence. The word humanism comes from *studia humanitatis* ("studies of humanity"). Toward the end of the Middle Ages there was a renewed interest in those studies that stressed the importance of man, his faculties, affairs, worldly aspirations, and well-being. The primacy of theology and otherworldliness was over; the *reductio artium ad theologiam* (freely, "reducing everything to theological argument") was rejected since it no longer expressed the reality of the new situation that was developing in Europe, particularly in Italy. Society had been profoundly transformed, commerce had expanded, and life in the cities had evolved. Economic and political power, previously in the hands of the ecclesiastical hierarchy and the feudal lords, was beginning to be taken over by the city burghers. Use of the vernacular languages was becoming widespread. The new society needed another kind of education and different educational structures; the burghers required new instruments with which to express themselves and found the old medieval universities inadequate.

Humanistic thought did not suddenly develop but was anticipated in the 14th century by such writers as Petrarch and Boccaccio. Similarly, the educational institutions of Humanism had their origin in the schools set up in the free cities at the end of the 13th and the 14th centuries—schools designed to answer to the needs of the new urban population that was beginning to have greater economic importance in society. The pedagogical thought of the Humanists took these transformations of society into account and worked out new theories that often went back to the classical Greek and Latin traditions; it was not, however, a servile imitation of the pedagogical thought and institutions of the classical world.

The Renaissance of the classical world and the educational movements it gave rise to were variously expressed in different parts of Europe and at various times from the 14th to the 17th centuries; there was a connecting thread, but there were also many differences. What the citizens of the Florentine Republic needed was different from what was required by princes in the Renaissance courts of Italy or in other parts of Europe. Common to both, however, was the rejection of the medieval tradition that did not belong in the new society they were creating. Yet the search for a new methodology and a new relation with the ancient world was bitterly opposed by the traditionalists, who did not want renewal that would bring about a profound transformation of society; and, in fact, the educational revolution did not completely abolish existing traditions. The Humanists, for example, were not concerned with extending education to the masses but turned their attention to the sons of princes and rich burghers.

The Humanists had the important and original conception that education was neither completed at school nor limited to the years of one's youth but that it was a continuous process making use of varied instruments: companionship, games, and pleasure were part of education. Rather than suggesting new themes, they wanted to discover the method by which the ancient texts should be studied. For them knowledge of the classical languages meant the possibility of penetrating the thought of the past; grammar and rhetoric were being transformed into philological studies not for the sake of pedantic research but in order to acquire a new historical and critical consciousness. They reconstructed the past in order better to understand themselves and their own time.

THE HUMANISTIC TRADITION IN ITALY

Early influences. One of the most influential of early Humanists was Manuel Chrysoloras, who came to Florence from Constantinople in 1396. He introduced the study of Greek and, among other things, translated Plato's *Republic* into Latin, which were important steps in the development of the Humanistic movement.

Inspired by the ancient Athenian schools, the Platonic Academy established in Florence in the second half of the 15th century became a centre of learning and diffusion of Christian Platonism, a philosophy that conceived of all forms as the creative thoughts of God and that inspired considerable artistic innovation and creativity. Marsilio Ficino and Pico della Mirandola were perhaps the most original of the scholars who taught there. Florence was the first city to have such a centre, but Rome and Naples soon had similar academies, and Padua and Venice also became centres of culture.

A famous early Humanist and professor of rhetoric at Padua was Pietro Paolo Vergerio (1370–1444). He wrote the first significant exclusively pedagogical treatise, *De ingenuis moribus et liberalibus studiis* ("On the Manners of a Gentleman and on Liberal Studies"), which, though not presenting any new techniques, did set out the fundamental principles by which education should be guided. He gave pedagogical expression to the ideal of harmony, or equilibrium, found in all aspects of Humanism, and underlined the importance of the education of the body as well as of the spirit. The liberal arts were emphasized ("liberal" because of the liberation they reputedly brought), but the program outlined by Vergerio included mathematics, astronomy, and natural sciences as well as medicine, law, metaphysics, and theology. This did not mean going into each one of these subjects in depth; Humanism was by its nature against encyclopaedism, but it brought out the relations between the disciplines and enabled a man to know many subjects before making a decision as to which one he should dedicate himself to. Learning was not to be exclusively from books, and emphasis was placed on the advantages of preparing for social life by study and discussion in common. Vergerio felt that education should not be used as a means of entering the lucrative professions; medicine and law, in spite of being considered important fields of study, were looked on with suspicion if one's aim in studying them was that of gaining material advantages.

Emphasis on the liberal arts

Humanistic boarding schools. The theory of Humanistic education, written about in many treatises, was given practical expression in a few boarding schools founded and directed by outstanding Humanists. These schools offered a general nonprofessional education and were attended by young men before going to a university. Among the most famous were those of Gasparino Barzizza, Guarino Veronese, and Vittorino da Feltre.

Gasparino Barzizza lived in Padua from 1408 to 1421, and his school was considered a model by many who came after him. The novelty was not so much in its organization as in the high calibre of the Humanists who taught there. The school had a hard struggle, however, because of resistance to the new methods, particularly at the universities.

Barzizza was followed by Guarino Veronese (1374–1460), who opened first a boarding school at Verona and subsequently another one at Ferrara, having been invited by the lord of Ferrara, Nicolò d'Este, to take charge of the education of his son. Guarino wrote no treatises, but something may be learned about his work from his large correspondence and from *De ordine docendi et studendi* (1485; "On the Order for Teaching and Studying"), written by his son Battista. The courses were organized into three stages: the elementary level, at which reading and pronunciation were primarily taught, followed by the grammatical level, and finally the highest level, concentrating on rhetoric. The education given in his schools was perhaps the best example of the Humanistic ideals, since it underlined the importance of literary studies together with a harmonious development of body and spirit, to the exclusion of any utilitarian purpose.

Humanist de-emphasis of "useful" education

Vittorino Rambaldoni da Feltre was a disciple of both Barzizza and Guarino. He started boarding schools at Padua and Venice and, most importantly, in 1423 one at Mantua, where he had been invited by the reigning lord, Gianfrancesco Gonzaga. This last school, known as La Giocosa (literally, the "jocose" or "humorous"), soon became famous. At La Giocosa only those who had both talent and a modest disposition were accepted; wealth was neither necessary nor sufficient to gain admission; in fact, the school was one of the few efforts made during this period to extend education to a wider public. The program of study at La Giocosa was perhaps closer to the medieval tradition than that of the other boarding schools, but, in any case, the spirit was different. Studies were stimulating; mathematics was taught pleasantly—Vittorino going back to very ancient traditions of practicing mathematics with games. After having studied the seven arts of the trivium (grammar, rhetoric, and logic) and the quadrivium (geometry, arithmetic, music, and astronomy), students completed the cycle by a study of philosophy and then, having mastered this discipline, could go on to higher studies leading to such professions as medicine, law, philosophy, and theology. Italian was completely ignored at Vittorino's school; all instruction was given in Latin, the study of which, together with Greek, reached a high level of excellence. Great importance was given to recreation and physical education; his concern for the health of his students did not come to an end with the scholastic year, for during the summers, when the cities became unhealthy, he would arrange for his students to go to Lake Garda or to the hills outside Verona.

Vittorino's educational philosophy was inspired by a profound religious faith and moral integrity, which contrasted with the general relaxation of standards within the church itself; but, if he was severe with himself, he was very open and tolerant with his pupils. The school continued only for a while after his death because, more than in the case of the other schools, La Giocosa was identified with the personality of the founder.

Nonscholastic traditions. Leon Battista Alberti, one of the most intelligent and original architects of the 15th century, also dedicated a treatise, *Della famiglia* (1435–44; "On the Family"), to methods of education. Alberti felt that the natural place for education was the home and not scholastic institutions. The language in which he wrote was Italian, education being in his view so important in social life that he felt that discussion of it should not be limited to scholars. He stressed the importance of the father in the educational process.

Baldassare Castiglione expressed the transition of Humanism from the city to the Renaissance court. He himself was in the service of some of the most splendid princes, the Gonzagas at Mantua and the Montefeltros at Urbino. Just as in the 15th century the Humanists had been concerned with the education of the city burgher, so in the 16th century they turned their attention to the education of the prince and of those who surrounded him. *Il cortegiano* ("The Courtier") was published in 1528, and within a few years it had been translated into Latin and all the major European languages. The courtier was to be the faithful collaborator of the prince. He had to be beautiful, strong, and agile; he had to know how to fight, play, dance, and make love. But this was not all, since great importance was also attached to the study of the classics and the practice of poetry and oratory; the courtier had to be able to write in rhyme and in prose and have perfect command of the vernacular, which was becoming important in political affairs; but above all he had to have skill at arms.

Education of the courtier

The courtier described by Castiglione, though in the service of necessarily devious princes, had to know how to keep his dignity and his virtue. Castiglione's moral standards, reflecting the spiritual climate at Urbino, completely disappeared, however, in Giovanni della Casa's work, *Galateo* (1551–54), in which considerations of etiquette were placed above all others; the values of Humanism no longer existed, and all that was left was ceremonial.

THE HUMANISTIC TRADITION OF NORTHERN AND WESTERN EUROPE

The economic and social conditions behind the intellectual and cultural revolution of Humanism in Italy were also present, though in different forms, in other parts of Europe. In some states, such as England, France, and Spain, Humanism and educational reforms developed around the courts, where political power was being concentrated; in others, such as the Netherlands, they were brought about by the city burghers, whose power, both economic and political, was increasing. The educational reforms that the Humanists brought about in northern and western Europe developed slowly, but on the whole they were lasting, since they affected a greater number of people than was the case in Italy, where they tended to be restricted to a narrow circle of families. There were close relations between Italian and other European educational Humanists, as there were among English, Dutch,

Courtly and bourgeois forms of Humanism

French, and German Humanists, and, thus, national differences were not so significant.

Dutch Humanism. In the Netherlands the ground for educational reform had already been prepared in the 14th century by the Brethren of the Common Life, a group founded by Gerhard Groote to bring together laymen and religious men. Although their work was not originally in the field of education, education started when they set up hostels for students and exercised some moral direction over these students; this work was extended, and the Brethren eventually set up schools, first at Deventer, then in other cities. Some of the most important Humanists of the Netherlands and Germany attended their schools—among others, Erasmus.

The school at Deventer came to have great prestige under Alexander Hegius, rector from 1465 to 1498 and author of a polemic treatise, *De utilitate Graeci* ("On the Usefulness of Greek"), underlining the importance of studying Greek, and of *De scientia* ("On Knowledge") and *De moribus* ("On Manners"). Hegius had great talent as an organizer and succeeded not only in attracting some of the best scholars of the time but also in giving the school an efficient structure that became a model for many schools in the north.

Desiderius Erasmus was a great educator, and his influence was felt all over Europe. His strong personality earned him the respect and sympathy of Humanists who saw in him, as in few others, the symbol of their ideals and values. Unfortunately, his proposals for reform and greater tolerance were not always accepted in the tortured Europe of the 16th century.

Erasmus was a prolific writer, and part of his work was concerned with education: *De ratione studii* (1511; "On the Right Method of Study"), *De civilitate morum puerilium* (1526; "On the Politeness of Children's Manners"), *Ciceronianus* (1528), *De pueris statim ac liberaliter instituendis* (1529; "On the Liberal Education of Boys from the Beginning"). His educational program was original in many ways but in no sense democratic. The masses could not partake in higher education, since their aim was that of gaining skill in an occupation. He felt that religious instruction should be made available to all but that classical literary studies—the most important of all studies—were for a minority.

The classical and elitist tradition in Humanism

Study of ancient languages and intelligent comprehension of texts formed the basis of Erasmus' system of education; he took a stand against the formalism and dogmatism that were already creeping into the Humanist movement. Erasmus was in favour of acquiring a good general liberal-arts education until the age of 18, being convinced that this would be a preparation for any form of further study. His great love for the classical languages, however, made him neglect the vernacular; he was not interested in local traditions; and he attributed very little importance to science, which he did not think necessary for a cultured man. He was against instruction being imposed without the participation of the student. His optimism about the nature of man and the possibilities of molding him made Erasmus feel that, if adequately educated, any man could learn any discipline. He further sought renewal of the schools and better training for teachers, which he felt should be a public obligation, certainly no less important than military defense. Many of Erasmus' themes were elaborated a century later by Comenius and form the basis of modern education, in particular the effort to understand the child psychologically and to consider education as a process that starts before the school experience and continues beyond it.

Juan Luis Vives. Strongly influenced by Erasmus was Juan Luis Vives, who, though of Spanish origin, spent his life in various parts of Europe—Paris, Louvain, Oxford, London, Bruges. His most significant writings were *De institutione foeminae Christianae* (1523; "On the Education of a Christian Woman"), *De ratione studii puerilis* ("On the Right Method of Instruction for Children"), *De subventione pauperum* (1526; "On Aid for the Poor"), and *De tradendis disciplinis* (1531; "On the Subjects of Study").

Not only was his vision of the organic unity of pedagogy new, but he was the first of the Humanists to emphasize the importance of popular education. He felt that it was the responsibility of the city to provide instruction for the poor and that the craft and merchant guilds had an important contribution to make to education. Unlike other Humanists, moreover, he did not despise the utilitarian aspects of education but on the contrary suggested that his pupils should visit shops and workshops and go out into the country to learn something of real life.

The new social and utilitarian traditions

Just as he felt that education should not be limited to a single social class, so he felt that there should be no exclusion of women, though perhaps they required a different kind of education because of their different functions in life.

Vives worked out a plan to take account of both educational structures and teacher training. In emphasizing the social function of education, he was against schools being run for profit and believed that teachers should be prepared not only in their specific fields but also in psychology so as to understand the child. He also suggested that teachers should meet four times a year to examine together the intellectual capacities of each one of their pupils so that suitable programs of study could be arranged for them. Vives considered that, in teaching, games had psychological value. He favoured use of the mother tongue as being simple for the boys to learn; nevertheless, as a Humanist, he had a passion for Latin and felt that there was no substitute for Latin as a universal language. Classical studies were to be completed by investigation of the modern world, in particular its geography, the horizons having been greatly enlarged by recent discoveries. Vives's method was an inductive one, based not on metaphysical theories but on experiment and exercise.

The early English Humanists. At the end of the 15th century there was a flowering in England of both Humanistic studies and educational institutions, enabling a rapid transition from the medieval tradition to the Renaissance. The English Humanists prepared excellent texts for studying the classical languages, and they started a new type of grammar school, long to be a model. Most important were John Colet and Thomas More. Thomas Linacre, author of *De emendata structura latini sermonis libri sex* ("Six Books on the Flawless Structure of the Latin Language"), should also be remembered, as well as William Lily, author of a Latin syntax, *Absolutissimus de octo orationis partium constructione libellus* (1515; "Comprehensive Study of the Construction of the Eight Parts of Speech"), and director of St. Paul's School in London from 1512 to 1522.

The English grammar schools

John Colet has an important place in English education. As dean of St. Paul's Cathedral he founded St. Paul's School, thus favouring the introduction of Humanism in England and the transformation of the old ecclesiastical medieval schools. He had travelled a great deal in France and Italy and wanted to bring to his country the Humanistic culture that had so fascinated him. In 1510 he started a "grammar school," open to about 150 scholars who had an aptitude for study and had completed elementary school. Colet's personality and energy made his school a lively centre of English Humanism.

Thomas More was both a distinguished Humanist and a statesman. He was interested in pedagogy, to which he dedicated part of his work *De optimo reipublicae statu deque nova insula Utopia* (1516; "On the Highest State of a Republic and on the New Island Utopia"). In his Utopia, More saw the connection between educational, social, and political problems and the influence that society therefore has on education. English Humanists such as More were engaged in a bitter battle because medieval tradition was deeply rooted; they were fierce opponents of a group called the Trojans, who opposed the Greek language and all that the new instruction of that language represented.

EDUCATION IN THE REFORMATION AND COUNTER-REFORMATION

New political and social systems developed in those European countries that, for various reasons and at dif-

ferent times, broke away from the Church of Rome in the 16th century. The religious reforms brought about by Luther, Calvin, Zwingli, and the ruling family of England were both cause and effect of these transformations. Characteristic of all these countries was the importance of the state in the organization of the educational system.

Humanism and the Reformation

The Reformation and European Humanism influenced one another. There were analogies between the flowering of the classical world in the European courts and the reawakening of religious interests; there were similarities in the critical position adopted toward Aristotelianism and in the interest shown toward the study of classical languages, such as Greek and Hebrew. The presuppositions behind the two movements—Humanism and Reformation—were different, however, and sooner or later a clash was inevitable. The most spectacular of these clashes was between Erasmus and Luther, despite the fact that for a long time they had respected each other. It was important for Erasmus and for the Humanists to encourage the development of a world of writers and artists who, free from material preoccupations, could devote their time to literary and artistic pursuits. For the Reformers the situation was different: they did not aim to educate a small minority; unlike Erasmus, Luther had to keep the popular masses in mind, for they had contributed to the success of the religious reforms.

Martin Luther and the German Reformation. Martin Luther specifically wished his humble social origins to be considered a title of nobility. He wanted to create educational institutions that would be open to the sons of peasants and miners, though this did not mean giving them political representation. (The German princes were glad to promote the Reformation on condition that it would not diminish but would, on the contrary, increase their political power.) Luther realized that an educational system open to the masses would have to be public and financed by citizens' councils. His educational programs are set out in *An die Radsherrn aller Stedte deütsches Lands: Das sie christliche Schulen affrichten und hallten sollen* (1524; "Letter to the Mayors and Aldermen of All the Cities in Behalf of Christian Schools"), in *Dass man Kinder zur Schulen halten solle* (1530; "Discourse on the Duty of Sending Children to School"), and in various letters to German princes.

Lutheran emphasis on public schools for all

Pre-Reformation schools in Germany had been influenced by the Humanistic movement and had been restricted to rather a narrow circle. As a social reformer, Luther suggested that sons of those lacking means and for whom work was a necessity should spend one or two hours a day at school and the rest of the time at home learning a trade. Luther was thus in favour of limited democratic reforms of the educational system: even though public schools would be open only a few hours, they would be open to all, girls and boys, without reference to the financial situation of the family.

Humanistic schools, whose upkeep was the responsibility of the princes and the cities, were soon organized—on the premise that a new class of cultivated men must be developed to substitute for the dispossessed monks and priests. In 1543 Maurice of Saxony founded three schools open to the public, supported by estates from the dissolved monasteries. It was more difficult to set up the popular schools, for which there was no tradition. In towns and villages of north Germany Johannes Bugenhagen (1485–1558) set up the earliest schools to teach religion and reading and writing in German, but it was not until 1559 that the public ordinances of Württemberg made explicit reference to German schools in the villages. This example was shortly followed in Saxony.

Whereas Luther combined his interest in education with his work as a religious reformer and politician, another Reformer, Philipp Melanchthon (1497–1560), concentrated almost entirely on education, creating a new educational system and, in particular, setting up a secondary school system. He taught for many years at the University of Wittenberg, which became one of the centres of theological studies in Reformation Germany; and his experience there enabled him to reorganize the old universities and set up new ones, such as Marburg, Königsberg, and Jena. His ideas about secondary education were put into practice in the schools he founded at Eisleben. Scholastic work was divided into three stages, access to each successive stage depending on the ability of the student to master the previous course work; this was a new concept (foretelling the later "grading system"), unknown in the traditional scholastic system. He was convinced that too many subjects should not be imposed on the student. He felt that Latin was important, but not German, Greek, or Hebrew, as had been taught in the Humanistic schools; such variety, he felt, was exhausting and possibly harmful. This opened the door to a new type of formalism, however, a danger that in other spheres the educational reformers had tried to fight.

Growing formalism in German schools

The work of Johannes Sturm (1507–89) illustrates this danger. He founded a grammar school in Strassburg that became a model for German schools. Sturm believed that methods of instruction in elementary schools and, to some degree, in secondary schools should be different from those in the institutes of higher education. Not much autonomy was to be allowed the child, who started learning Latin at the age of six by memorizing. Sturm's love of Latin was even greater than that of his friend Erasmus, who never wanted it to become a mechanical exercise. As a consequence, German was neglected, as was physical instruction, and too much importance was given to form and expression for its own sake.

The work of Valentin Friedland, known as Trotzendorf (1490–1556), was interesting in a different way. He founded a grammar school at Goldberg, where he tried to put into practice Luther's idea that school should be enjoyable. His school did not limit itself to instruction but was run like an autonomous society.

The English Reformation. The separation of the Church of England from the Church of Rome in the 16th century under Henry VIII did not have quite the repercussions in the scholastic field that were experienced by the continental Reformations. The secondary school system in England had been strongly influenced by the Renaissance in the period preceding the reform, and about 300 grammar schools were already in existence. Nevertheless, the situation became precarious, for political reasons, under a succession of sovereigns.

Henry VIII included the schools in his policy of concentration and consolidation of power in the hands of the state. In 1548, under Henry's son Edward VI, the Chantries Act was passed, confiscating the estates of the church expressly for use in education; but the turmoil of the times, under the boy Edward and then his Catholic sister Mary I, allowed the funds allocated to education to be diverted elsewhere. Many primary schools and grammar schools disappeared or retrenched their operations for lack of funds. Elizabeth I, however, succeeding to the throne in 1558, revived Henry VIII's scholastic policy; considerable sums were appropriated for education, even though it was not always possible to enforce the new provisions because of local opposition and some lack of concern on the part of the Anglican clergy.

The growth of a rich and prosperous mercantile class and the spread of Calvinist reforms through the Puritans in England and the Presbyterians in Scotland were also factors in the transformation of English education in the 16th and 17th centuries. Scholastic programs reflected changes in society: importance was given to English, to science, to modern languages (in particular French and Italian), and to sports, as is still the case in England today. The Puritan contribution was thus considerable, though often hindered by the traditional forces of the Anglican Church and the old nobility.

The work of Sir Thomas Elyot and Roger Ascham

Sir Thomas Elyot, in *The Boke Named the Governour* (1531), wrote the first treatise in English that dealt specifically with education. He was interested in those who would have the future economic and political power in their hands. Though their education was to include the classics, it was to be supplemented by the needs of the new mercantile class—the national English language, manual arts, drawing, music, and all forms of sport. Elyot was obviously influenced by Erasmus.

Roger Ascham was close in thought to many of the English Humanists. In *The Scholemaster* (1570) he underlined the importance of the English language (in spite of his being a professor of Greek) and proposed that it should be used in teaching the classical languages. He also believed that physical exercise and sport were important, not only for the nobility and the leisured classes but also for students and teachers. He was aware of the social changes in the country; and, observing with sadness the corruption of the new wealth, he was particularly chagrined to see students going to university not to gain culture but to prepare themselves for high offices of state.

Richard Mulcaster had 30 years of experience as an educator at St. Paul's School and at the Merchant Taylors School, a Latin secondary school maintained by the tailors' guild in London—and most famous of all the "guild schools." Mulcaster was in favour of efficient teacher training and of teachers being adequately paid. In agreement with some of the Lutheran educational reforms, he felt that schools should be open to all, including women, who should moreover have access to higher education. He is particularly remembered for his opposition to Italianate trends: "I love Rome, but London better. I favour Italy, but England more. I know the Latin, but worship the English."

Sir Francis Bacon was interested in education though it was not his main concern—his main concern being the championship of the scientific method and "sense" realism, in opposition to traditional Aristotelianism and Scholasticism. He was opposed to private tutors and felt that boys and youths were better off in schools and that their education should be geared to their social status and future activity. Schooling should aim at preparing statesmen and men of action as well as scholars and thus should include history, modern languages, and politics. Bacon himself had a passion for study not only for its utilitarian purposes but because of its being for him a true source of delight.

The French Reformation. François Rabelais was a great and original interpreter of Humanistic ideals, and his views on education reflected this. He himself studied in various fields, from medicine to letters, and was passionately interested in all of them. His controversy with the Sorbonne, a remaining stronghold of medievalism and Scholasticism, was bitter and satirical. In *Gargantua* the Sorbonistes, or Sorbonagres, had filled Gargantua's head with useless notions, and it took a purge to liberate him from such indigestible knowledge.

Reformers opposed to medieval Aristotelianism and authority

Rabelais's educational philosophy was entirely different from that of the medievalists—his being based on liberty of the pupil, in whom he had maximum faith. In his novel, this cult of liberty was celebrated in the utopian Abbey of Thélème, where all could live according to their own pleasure but where the love of learning was so great that everyone was dedicated to it, getting much better results than those obtained at the medieval universities. And yet in the education of Gargantua and Pantagruel, there were limits placed on liberty: Gargantua's day started at 4 in the morning; he studied all subjects, both literary and scientific; and this was alternated with play and pleasing diversions. The heavy program, however, was not a constriction because of Gargantua's delight in learning. The culture that Rabelais wanted for his two heroes was directly connected with the world in which they lived.

Gargantua and *Pantagruel* were perhaps among the first texts by a Humanist in which not only the quadrivium but scientific studies were enthusiastically proposed. There was nothing arid or abstract in Rabelais's approach to nature, and in this context the classics also had a new flavour: ancient literature, no longer limited to Latin, Greek, and Hebrew but expanded to include Arabic and Chaldaic, could bring to light valuable knowledge that had been accumulated by the classical world.

Petrus Ramus, one of the most bitter critics of French medieval Aristotelianism, was an intelligent reformer of educational methods. His best known treatises are *Aristotelicae animadversiones* ("Animadversions on Aristotle") and *Dialecticae partitiones* ("Divisions of Dialectic"), both condemned by royal decree; he also wrote two discourses on philosophy, *Oratio de studiis philosophiae et eloquentiae conjungendis* (1546; "Speech on Joining the Study of Philosophy with the Art of Speaking") and *Pro philosophica Parisiensis accademiae disciplina oratio* (1551; "Speech in Defense of the Philosophical Discipline of the Parisian Academy"), as well as *Ciceronianus*, published posthumously. In these works his criticism of traditional ways and of the degeneration of Humanistic thought made him hated by all Roman Catholics, though not much better understood by Protestants; he died a Protestant victim of the massacre of St. Bartholomew. His program of study was fairly close to the traditional one, but his method was original, for he was concerned that the teacher should not suffocate the child with too many lessons and considered the child's autonomous activity important. He especially resented any pedagogy that relied on a blind appeal to authority; learning had to be utilitarian and issue from practice.

Michel de Montaigne (1533–92) was much influenced by his personal experience as a student. Though often critical of Humanism, especially when it was misinterpreted and transformed into pedantic studies, he had great admiration for the classics and lacked the scientific interests of Rabelais or Ramus. Montaigne wrote specifically about education in two essays on the upbringing of children and on pedantry. Culture, he felt, had become imitation, often with no trace of originality left, whereas it should be a delight—not something a student is forced to assimilate but something to draw the student's participation. He was in favour of instruction by tutors capable of giving the student individual attention —the ideal tutor being one with a good mind rather than one filled with pedantic notions. He also believed in the importance of physical education and in a boy's being hardened to nature and to danger.

For Montaigne it was important not only to travel to foreign countries but also to stay there for a while, to learn languages but, even more, to learn about foreign customs and thus break out of the narrow limits of one's own province. There were many differences between Montaigne and Erasmus, but both were convinced that for the wise man there could be no geographical boundaries, for, through cultural diffusion, barriers would be broken down.

The Calvinist Reformation. The Protestant Reformer John Calvin was of French origin, but he settled in Geneva and made this Swiss city one of the most prominent centres of the Reformation. Unlike Luther, whose reforms were backed by princes hoping to gain greater political independence, Calvin was supported by the new mercantile class, which needed political and administrative changes for the purposes of its own expansion.

Though he considered popular education important, Calvin concentrated on higher education. The theological academy he founded in Geneva in 1559 became distinguished under the directorship of Theodore Beza, an intelligent Reformer but unfortunately a very intolerant one, at least in theological matters. Calvin's influence on higher education was felt in many of the European universities, even as far as England, where, in spite of Anglican opposition, the Puritans had gained a foothold.

Calvinist emphasis on universal literacy

Calvin was in favour of universal education (the cost to be in large part borne by the community), but "universal" did not mean "democratic." Even if some form of instruction was to be given to everyone (so that everyone might in some measure read the Scriptures for himself, in good Calvinist tradition), very few individuals reached secondary or higher education, and of these only a minute percentage came from the working classes. Documents of the period show the steps taken to achieve the aim of universal education. In the Netherlands, the Calvinist Synod of The Hague in 1586 made provision for setting up schools in the cities, and the Synod of Dort in 1618 decreed that free public schools should be set up in all villages. In Scotland in 1560 John Knox, a disciple of Calvin and the leader of the Scottish Presbyterians, aimed at setting up schools in every community, but the

nobility prevented this from actually being carried out. The major educational contributions of Calvinism were its diffusion to a larger number of people and the development of Protestant education at the university level. Not only Geneva was significant but also the universities of Leiden (1575), Amsterdam (1632), and Utrecht (1636) in the Netherlands and the University of Edinburgh (1582) in Scotland. The Puritan, or English Calvinist, movement was responsible for the founding of Emmanuel College at the University of Cambridge (1584).

The Roman Catholic Counter-Reformation. The religious upheaval, so important in northern Europe, also affected, though less violently, the Latin countries of southern Europe. If the new ferment in the Roman Catholic Church was mainly directed at answering the Protestants, at times it also had something original to suggest. At the Council of Trent (1545–63) the Roman Catholic Church tried to come to terms with the new political and economic realities in Europe.

The growth of seminaries

Education was foremost in the minds of the leaders of the Counter-Reformation. The faithful were to be educated. For this, capable priests were needed, and, thus, seminaries multiplied to prepare the clergy for a more austere life in the service of the church. There was a flowering of utopian ideas, which should be remembered when trying to understand unofficial Catholic thought of the period. Various writings, such as *La città del sole* ("The City of the Sun"), by Campanella, *Repubblica immaginaria* ("The Imaginary Republic"), by Agostini, and *Repubblica di Evandria* ("Evander's Republic"), by Zuccolo, are examples of this new vision of the church and of the duties of Christians. But if in the minds of the utopians this education was to be universal, it was in fact almost entirely directed at the ruling classes.

The Society of Jesus, founded in 1534 by Ignatius Loyola, was not specifically a teaching order, but it was nevertheless very important in this field. The first Jesuit college was opened in Messina, Sicily, in 1548; by 1615 the Jesuits had 372 colleges, and by 1755, just 18 years before the suppression of the order, the number had risen to 728. (The Society was not re-established until 1814.) In *Ratio studiorum*, an elaborate plan of studies issued by the Jesuits in 1599, there is laid out an organization of these institutions down to the smallest details; an authoritarian uniformity was thus the rule in their colleges, and individual initiative was discouraged. The complete course of study took 13 years, divided into three periods: six years that included grammar and rhetoric, three years of philosophy, and four of theology. The teacher was thought of not only as an instructor but also as an educator and often a controller, for he was at the centre of a vast network of controls, in which those students considered promising also took part. Emulation was encouraged in the class, which was often divided into two groups to stimulate competition. These new techniques, as well as the Jesuits' efficient training of teachers, had good results, proof of this being the rapid increase in their colleges, which found greater favour than others started in the same period.

The legacy of the Reformation. The effects on education of a movement as complex and widespread as the Reformation were far-reaching. Perhaps its most original contribution was the extension of education at the elementary level. As a result, the vernacular language took on a new importance, and also the new pedagogy had to take account of the realities of the situation—namely, that the children brought into the new school network could not spend as much time on "useless" books, so that schoolwork had to be combined with learning a practical trade, which had not previously been considered a part of education. (E.Ge.)

VII. European education in the 17th and 18th centuries

THE SOCIAL AND HISTORICAL SETTING

The Renaissance had been the beginning of a new era in history, which culminated in the 17th and 18th centuries in the development of the absolutist state everywhere but in England and Holland (and even in these states the issue was for some time in doubt). France, the Habsburg Empire, England, and Russia became the leading powers in Europe. The absolutist state extended its control beyond the political and into the religious (with the creation of the established church) and into almost all other aspects of human life. Although the High and later Middle Ages had witnessed the growth of middle class forces, the pattern of society still clearly bore the stamp of court life. The concentration of power determined this life, and the citizen and his possessions were more and more at the disposal of the aristocracy. The citizen was subject.

Influence of absolutism on education

Even in an absolutist state, however, education cannot be the sole privilege of the rich or the ruling classes, because an efficient absolutist state requires capable subjects, albeit bound to their social position. Elementary education for the middle classes thus developed in the 17th and 18th centuries, and more and more the state saw as its task the responsibility for establishing and maintaining schools. This tendency toward general education did not stem only from considerations of political expediency; it stemmed also from the desire to improve the world through education—making all areas of life orderly and subordinate to rational leadership. There was not only an inclination toward encyclopaedism and systemization of the sciences but also, in similar fashion, a tendency to set education aright by extensive school regulations.

The 18th century is especially marked as a century for education. Essentially, there were three leading reforms: teaching in the mother language grew in importance, rivalling Latin; the exact sciences were brought into the curriculum; and the correct methods of teaching became a pedagogic question.

If a distinction can be made between the 17th and the 18th centuries, it was this: in the 17th century the aim of education was conceived as a religious and rationalistic one, whereas in the 18th century the ideal of the secular, pragmatic, "enlightened" gentleman prevailed more and more.

The new scientism and rationalism. These social and pedagogic changes were bound up with new tendencies in philosophy. Sir Francis Bacon of England was one who criticized the teachers of his day, saying that they offered nothing but words and that their schools were narrow in thought. He believed that the use of inductive and empirical methods would bring the knowledge that would give man strength and make possible a reorganization of society. Therefore, he demanded that schools should be scientific workplaces in the service of life and that they should put the exact sciences before logic and rhetoric.

Another 17th-century critic of medievalism was René Descartes, but he did not proceed from empirical experience, as did Bacon; for him the only permanence and certainty lay in human reason or thinking (*cogito ergo sum*, "I think; therefore, I am"). The ability to think makes doubt and critical evaluation of the environment possible. A science based only on empiricism fails to achieve any vital, natural explanations but only mathematical, mechanistic ones of doubtful living use. Only what reason (*ratio*) recognizes can be called truth. Thus, education must be concerned with the development of critical rationality and the teaching of any practical discipline that makes a man master and lord of nature.

Knowledge conceived as thinking

Like Descartes, Benedict de Spinoza and Gottfried Wilhelm Leibniz also outlined rationalistic philosophic systems. Decisive for educational theory was their statement that knowledge and experience originate in thinking (not in sense impressions) and that formal thinking categories should form the substance of education. Here the overthrow of encyclopaedism is already hinted at: education is guaranteed not by limitless widening of assimilated facts but by the mastery of thinking and judgmental categories.

The Protestant demand for universal elementary education. The schools that were actually developed fell short of these philosophically based demands. This is especially true of elementary education. In the Middle

Ages, the grammar schools (especially for the education of the clergy) had developed, and the Humanism of the Renaissance had strengthened this tendency; only those who knew Latin and Greek could be considered educated. For basic, popular education there were meagre arrangements. Although schools for basic writing and arithmetic had been established as early as the 13th and 14th centuries, they were almost exclusively in the towns; the rural population had to be content with religious instruction within the framework of the church. An improvement came as a result of Protestantism. John Wycliffe had demanded that everyone become a theologian, and Martin Luther, by translating the Holy Scriptures, made the reading of original works possible. Everyone, he asserted, should have access to the source of belief, and all children should go to school. So it happened that church regulations of the 16th and 17th centuries began to contain items governing schools and the instruction of young people (mainly in reading and religion). The Protestant "bell ringer" schools acquired great importance alongside the town schools. At first, such schools were directed and supported almost entirely by the church. Not until the 18th century, following the general tendency toward secularization, did the state take over the responsibility of supporting the schools.

EDUCATION IN 17TH-CENTURY EUROPE

Central European theories and practices. It was while Europe was being shaken by religious wars and was disintegrating into countless small states that such writers as Tommaso Campanella and Sir Francis Bacon dreamed their utopias (*La Città del sole* and the *New Atlantis*, respectively), where peace and unity would be had through logical and realistic means. To even attempt realizing this dream, however, man needed suitable education.

Realism in education

Both leading representatives of so-called pedagogic realism, Wolfgang Ratke and John Amos Comenius, were motivated by this ideal of world improvement through a comprehensive reform of the school system. Despite this common starting point, however, both were highly distinct personalities and, moreover, had divergent influences on the development of education and schools.

The pedagogy of Wolfgang Ratke. Wolfgang Ratke (1571–1635), a native of Holstein in Germany, journeyed to England, Holland, and through the whole of Germany and to Sweden expounding his ideas to the political authorities and finding considerable support. His plans for progressive reform failed for several reasons. First, political conditions during the Thirty Years' War were understandably not favourable for any kind of planning or reform of schools. Moreover, even when Ratke had opportunities, he demonstrated scant administrative or practical ability in trying to put his plans into execution. Several of his attempts collapsed because of this. Finally, Ratke's ideas were not free of exaggerations and presumption. He promised, for example, to be able to teach ten languages in five years, each language in six months.

His thoughts about the art of teaching are of greater importance for the theory and practice of education. First, he believed that everything must be studied through induction and experiment (as Bacon has espoused). Knowledge is to be gained by experience in dealing with specific problems: only then can the problem be discussed and a rule or "law" be reached regarding it.

Second, everything must follow the order and course of what may be called human nature. In modern terms, one would say that a lesson should be designed with psychological conditions taken into consideration.

Third, he asserted that everything should be taught first in the mother language, the mother language being the natural and practical language for a child and the one that allows him to concentrate wholly on the business at hand. Only when the mother language is fully commanded should the child attempt a foreign language; then special attention should be paid to speaking it and not merely reading it.

Fourth, Ratke emphasized what might now be called a kind of programmed learning. One piece of work should be fully completed before progress is made to the next piece, and there should be constant repetition and practice. The teacher's methods and the textbook program should agree and coincide.

Ratke's teacher-centred education

Fifth, there should be no compulsion. A teacher should not be a taskmaster. To strike a pupil was contrary to nature and did not help him learn. A pupil should be brought to love his teacher, not hate him. On the other hand, all work was the teacher's responsibility. The pupil should listen and sit still. More generally, all children, without exception, should go to school, and no lessons should be cancelled for any reason. There is, of course, a certain paradox in Ratke's views here—at least in modern terms. There was to be no compulsion, and yet pupils were to remain utterly disciplined. The emphasis on order, uniformity, and strictness was common to the times, however, and Ratke was not alone in failing to see the possibility of a pupil working independently.

As for curricula, Ratke suggested reading and writing in the native tongue, singing, basic mathematics, grammar, and, in the higher classes, Latin and Greek. The sciences had not yet appeared in his timetable. His demand that, above all, young people should be given instruction in the affairs of God is typical of the combination of rationalistic and religious education in the 17th century.

The pedagogy of Comenius. John Amos Comenius (1592–1670) was, even more than Ratke, a leading intellect of European educational theory in the 17th century. Born in Moravia, he was forced by the circumstances of the Thirty Years' War to wander constantly from place to place—Germany, Poland, England, Sweden, Hungary, Transylvania, and Holland—and was deprived of his wife, children, and property. He himself said, "My life was one long journey. I never had a homeland."

As a onetime bishop of the Bohemian Brethren, he sought to live according to their motto, "Away from the world towards Heaven"—yet while combining a religious and realistic outlook. Preparation for the hereafter, he felt, comes not through disregarding this world but through *recte vivere*, or "living rightly"—that is, living one's life according to correct factual and moral principles and seeking "learned piety" (*pietas litterata*). Whether his realism was truly realistic or not, Comenius could certainly be said to have had a philosophy that was humanitarian, inspired at least partly by the turmoil of his times. It encompassed three levels: (1) peace on earth exists only when religious factions cease fighting and unite toward realizing God's will; (2) such understanding and cooperation can best be achieved through "pansophia," universal knowledge among all men and all nations, achieved by means of (3) widespread educational reforms in methods and principles of teaching, including a universal language and universal coeducation.

The curriculum of Comenius

An all-embracing school system would serve these ends. During infancy (up to six years of age), the child in his "mother school," or family grouping, would develop his basic physical faculties. During the following period (seven to 12), the child would go to the "vernacular school," which was divided into six classes according to age and could be found in every town. The prime aim of these schools would be to develop the children's imagination and memory through such subjects as religion, ethics, diction, reading, writing, basic mathematics, music, domestic economy, civics, history, geography, and handicraft. This vernacular school formed the final stage of education for technical vocations. After this school would come the grammar school (or Latin school), which the pupils would attend during their youth (13–18) and which would exist in every town of every district. Through progressive courses in language and the exact sciences, the young people would be brought to a deeper understanding of things. Finally, the university (19–24) would be a continuation of this school. Every province ought to have one such university, whose central task would be the formation of willpower and powers of judgment and categorization. Over and above this four-tier school system Comenius also envisaged a "college of light," a kind of academy of the sciences for the centralized pooling of all learning. It is important to note, in this

regard, that it was Comenius' stay in England (1641–42) that initiated discussions leading to the founding of the Royal Society (incorporated 1662). Furthermore, the German philosopher Leibniz, influenced by Comenius, founded the Berlin Academy, and similar societies sprouted elsewhere.

With regard to methodology, Comenius believed that, like nature, education must take care to get the right timing and the right sequence. The lesson must begin with clear, practical facts and enable pupils to engage in independent and self-regulated activity and to have time for revision and practice. Most important, everything should be presented to the child's senses—and to as many senses as possible simultaneously, using pictures, models, workshops, music, and other "objective" means. On the upper age levels, he recommended that language study and other studies be integrated, and indeed he employed this scheme in his *Gate of Languages Unlocked* (1631), a book of Latin and sciences arranged by subjects, which revolutionized Latin teaching and was translated into 16 languages. *Orbis Sensualium Pictus* (1658; "Picture of the Sensual World"), which remained popular in Europe for two centuries, attempted to dramatize Latin through pictures illustrating Latin sentences, accompanied by one or two vernacular translations. His major theories were developed in *The Great Didactic* (1657) and in *Pampädia* (discovered in 1935).

The schools of Gotha. The zeal for reform on the part of such educators as Ratke and Comenius, on the one hand, and the interests of the ruling classes, on the other, led in the years after about 1650 to the publication of school regulations, free of church regulations. The circumstances in the central German principality of Gotha were typical. The duke, Ernst the Pious, commissioned the rector Andreas Reyher to compile a system of school regulations, which appeared in 1642 and is known historically as the *Gothaer Schulmethodus.* This was the first independent civil system of school regulations in Germany and was strongly influenced by Ratke. The most important points of these regulations were compulsory schooling from the age of five; division of the school into lower, middle, and higher classes; extension of the usual subjects (reading, writing, basic arithmetic, singing, and religion) to various other fields (natural history, local history, civics, and domestic economy); the introduction of textbooks (for reading and basic arithmetic), notably the first textbook of exact sciences for elementary schools, Reyher's own *Kurzer Unterricht von natürlichen Dingen* (1657; "Short Course on Natural Things"); and methodical instruction that, above all, emphasized the clarity of the lesson and the activity of the pupils.

Systematizing education

The system of elementary schools in Gotha was an example for many other parts of Germany. The Duke also took an interest in the education and examination of teachers and tried to raise their status. The other German lands for the most part did not keep pace with this latter development.

French theories and practices. In the second half of the 17th century, Germany suffered from the aftereffects of the Thirty Years' War, whereas France under Louis XIV reached the zenith of political and military power. France's leadership was also demonstrated in the cultural field—including education. Some of the most important developments in France included the promotion of courtly education and the involvement of religious orders and congregations in the education of the poor. One of the most important educational theorists of the late 17th century was a French cleric—Fénelon, archbishop of Cambrai.

Views of Montaigne

Courtly education. The rationalistic ideal of French courtly education can be seen foreshadowed in Michel de Montaigne's *Essais* (1580), in which the ideal man was described as having a natural, sensible way of life not deeply affected by the perplexities of the time but admitting of pleasure. He had a "correct" attitude toward the world and people, a certain spiritual freedom, and an independent judgment—all of which, in Montaigne's view, were more important than being steeped in knowledge. "As lamps are extinguished from too much oil, so is the mind from too much studying." Montaigne came from a merchant family that aspired to nobility, and thus there is a certain fashionable elitism in his views; he held, among other things, that courtly education succeeds best when the pupil studies under a private tutor.

This ideal, rather unlike the ideal of the learned and humanistic Renaissance man, became important in 17th-century France, especially after midcentury and the rise of the court of Louis XIV. The education of the would-be versatile and worldly-wise gentleman was furthered not only by the continuation of the institution of private tutoring but also by the establishment of schools and academies for chevaliers and nobles, in which the emphasis was on such subjects as deportment, modern languages, fencing, and riding. It was most emphatically an example of class education, designed for the nobility and higher military and not for any commoners.

The teaching congregations. In the countries, such as France, that remained Catholic, the Roman Church retained control of education, and indeed, as monarchy became more absolute, so largely did the authority of the church in matters of education. In France, practically all schools and universities were controlled by so-called teaching congregations or societies, the most famous and powerful of which during the first half of the 17th century was the Society of Jesus. By midcentury the Jesuits had 14,000 pupils under instruction in Paris alone; and their colleges (not including universities) all over the land numbered 612.

Jesuit education

It was their successful teaching and comparatively mild discipline that caused the Jesuit schools to attract thousands of pupils. "They are so good," said Francis Bacon of the Jesuit teachers in his *Advancement of Learning,* "that I wish they were on our side." The curriculum was purely classical, but importance was attached to spacious, well-adapted buildings and amenities designed to make school life interesting. In general, however, the religious and international conflicts did great harm to education, which suffered much because those kings and religious factions that gained power in France (as elsewhere) used the schools to propagate their cause, discarding teachers not of the approved persuasion. Moreover, the schools continued largely to ignore the new directions of men's minds; in the universities staffed by Jesuit fathers, medieval Scholasticism, though purged of the formalistic excesses that had degraded it, was fully restored. Schools and universities declined for the most part to contemplate any enlargement of the frontiers of knowledge and were too often deeply involved in the religious conflicts of the time. The University of Paris in particular remained distracted throughout the 17th century by theological dissensions—in at least one instance as a result of the rivalry that ensued after the Jesuits had effected a footing at the Collège de Clermont.

Aside from the Jesuits, the most important teaching congregations in France were the Bérullian Oratory, or Oratorians, and the Jansenists of Port Royal. The former, founded in 1611 and soon to open a number of schools and seminaries for young nobles, was composed of priests—but priests more liberal and rationalist than was common for the times. They offered instruction not only in the humanities but also in history, mathematics, the natural sciences, and such genteel accomplishments as dancing and music and, though continuing to use Latin in instruction, promoted also the use of the vernacular French in the initial years of their curriculum. They tended indeed to be drawn to the ideas of Descartes, to a faith based on reason. When in 1764 the Jesuits were banned from France, their teaching positions were largely assumed by Oratorians.

The Jansenists

More famous than the schools of the Oratorians, though enjoying a briefer career, were the Little Schools of Port Royal. Their founder was Jean Duvergier de Haurame, better known as the Abbot of Saint Cyran, who was one of France's chief advocates of Jansenism, a movement opposed to Jesuitry and Scholasticism and favouring bold reforms of the church and a turn to a certain Pietism. About 1635 Saint Cyran, with the help of some wealthy,

influential Parisians, succeeded in gaining control of the convent of Port Royal, near Versailles. There the Jansenist group began about 1637 to educate a few boys, and by 1646 it had established the Little Schools of Port Royal in Paris itself. Their curriculum was similar to that of the Oratorians, though excluding dancing, and was celebrated for its excellence in French language and logic and in foreign languages. The theory of learning adopted was that things should begin with what is familiar to the child: thus, a phonetic system of teaching reading was used; all instruction was in French, not Latin; student compositions were directed toward topics drawing on one's own experiences or toward subjects in one's current reading. Involved in political struggles with the Jesuits, who were still influential at court, the Jansenists were fated to have all their schools closed down by 1660, but their theories and practices were widely adopted and became extremely influential.

Female education. During the century, the education of girls was not entirely neglected, and France was notable for its efforts. Mme de Maintenon, for instance, had been a pupil of the Ursuline nuns in Paris and then a governess at the court of Louis XIV, before she was wedded to the King in 1684. From her royal vantage point, she took upon herself the founding of a school in 1686 at Saint-Cyr near Versailles—a higher school principally for orphan girls descended of noble families. Besides such basic subjects as reading and writing, the girls were prepared for their future lives as wives and mothers or as members of genteel professions. In 1692 this school was taken over by the Augustinian nuns. Another important worker in the field of female education was St. Jane Frances de Chantal, who, together with her father confessor, St. Francis de Sales, founded in 1610 the order of the Visitandines, a group dedicated to charitable work and the religious education of women.

Views of Fénelon

François de Salignac de la Mothe Fénelon, archbishop of Cambrai and noted theologian and writer, is especially known for his views on the education of girls. In his "Traité de l'éducation des filles" (1687), he remarked on the importance of women in improving the morals of society and went on to express his thoughts about girls' education. Because girls, he believed, are meant to fulfill roles as housewives and mothers, they should pursue religious and moral education rather than scholarly learning. They should learn reading and writing, basic mathematics, history, music, needlework, Latin (because it is the church language), but no modern languages, since they tend to moral corruption. Education, he maintained, should make the lady of the house both Christian and accomplished, neither ignorant nor *précieuse*.

English theories and practices. The 17th century in England (up to the Glorious Revolution of 1688–89) was one of argument over religious and political settlements bequeathed by Queen Elizabeth; the period was one characterized by the confrontation of two different world views—on one side the royalist Cavaliers and on the other side the Puritans. The division was reflected in education.

The Puritan reformers. In the Anglo-American world the Reformation came about in the form of Calvinism—"Puritan" being the derisory name for strict Calvinists. Their ideals were sober, practical behaviour, careful management, thrift, asceticism, and the rejection of hedonistic pleasures of life. Many of the educationists who sought this Puritan ideal were followers of the reform plans of Comenius. Samuel Hartlib, a Polish merchant residing in England who was friend, publisher, and patron of Comenius, tried to interest Parliament in the idea of popular education; his tractate entitled "London's Charity Enlarged" (1650) proposed that a grant be made for the education of poor children, all in the interest of general social betterment. The Committee for Advancement of Learning, which he founded in 1653, was the impulse and model for later educational associations. In general, his ideas for reform included the introduction of agricultural schools, the state organization of the educational system, as well as the establishment of general elementary education. The name of John Dury stands close to those of Comenius and Hartlib. In 1651 appeared his book *The Reformed School*, in which he proposed teaching societies in England much like the teaching congregations in France. Indeed, he was particularly insistent that control of education be in the hands not of a regimentizing state but of free educational organizations. He was also concerned about teaching youth the useful arts and sciences so that they might "become profitable instruments of the Commonwealth." From him, too, stemmed the draft of a nursery school; thus, he can be regarded as the first representative of infant teaching in England.

Sir William Petty, the noted economist, carried the idea of education in the useful arts even further. He wanted to turn ordinary schools into trade or handicraft schools and proposed a *gymnasium mechanicum* or "Colledge of Tradesmen of the Advancement of all Mechanical Arts and Manufactures."

The Humanism of John Milton

The most renowned of the Puritan intellectuals, John Milton, was more concerned with the education of "our nobler and our gentler youth" than with the education of common boys. "Of Education" (1644), written at the request of Hartlib, was one of the last in the long line of European expositions of Renaissance Humanism. Milton's aim was the traditional aim, the molding of boys into enlightened, cultivated, responsible citizens and leaders. His proposed academy, which would take the place of both secondary school and college, was to concentrate on instruction in the ancient classics, with due subordination to the Bible and Christian teaching. The traditional role of physical and martial exercise was perhaps heightened by Milton's anxiety over the English Civil War in progress. Interestingly, he also gave a large place to all kinds of science.

Royalist education. Frequently opposed to Puritanism on educational as well as political grounds were the royalists and supporters of the nobility. In education, their views went back to Thomas Elyot and Roger Ascham in the 16th century, who had written so persuasively about the education of gentlemen in the tradition of the so-called courtesy books. Influenced by these few English forerunners and also by Montaigne were James Cleland (*The Institution of a Young Nobleman*, 1607) and Henry Peacham (*The Compleat Gentlemen*, 1622). In the view of the latter, an extreme royalist, "Fashioning him [the pupil] absolute in the most necessary and commendable Qualities concerning Minde and Body to country's glory" was the overriding aim of education; the table of contents of *The Compleat Gentlemen* exhibits the variety of interests of an ideal gentleman or noble—cosmography, geometry, poetry, music, sculpture, drawing, painting, heraldry, and so on. John Gailhard (*The Compleat Gentleman*, 1678), another writer in the same tradition, can be said to have anticipated John Locke's empiricism (see below) when he wrote that "the nature of Youth is like Wax by fire, or a smooth table upon which anything can be written."

The academies. The beginning of academies for the promotion of philosophy, arts, or sciences can be traced to the early Renaissance, particularly in Italy and France. The Platonic Academy in Florence, cited earlier in this article, was one of the most noted of speculative societies. The first scientific academies belong to the 16th century: in 1560, for instance, the Academia Secretorum Naturae was founded in Naples; in 1575 Philip II of Spain founded in Madrid the Academia de Ciencias Matemáticas. Then, in 1617, the first German academy, Fruchtbringende Gesellschaft, was founded at Weimar for the purification of the language and the cultivation of literature. A number of other academies were founded throughout Europe.

Royal Society and Académie des Sciences

It was in the 17th century that the two pre-eminent scientific academies were founded. Both the English Royal Society and the French Académie des Sciences began as informal gatherings of famous men. The "invisible college" of London and Oxford had its first meetings in 1645; it was incorporated as the Royal Society in 1662. In Paris, a group of men including the philosophers Descartes and Pascal started private meetings almost at the same time. In 1666 they were invited by the economic

minister Colbert to meet in the royal library. In 1699 the society was transferred to the Louvre under the name of the Académie des Sciences. The Académie Française also started as a private society of men of letters some five years before its incorporation in 1635 under the patronage of Cardinal de Richelieu. In the 18th century, the fame and achievements of these English and French academies became internationally recognized, and many other European countries started to found their own national academies.

EDUCATION IN 18TH-CENTURY EUROPE

General condition of education during the Enlightenment. In the 18th century the theories and systems of education were influenced by various philosophical and social trends. Among these were realism, which had its origins in Ratke and Comenius, among others, and also Pietism, which derived principally from Philipp Jakob Spener and August Hermann Francke in the late 17th and early 18th centuries (see below). Another trend was the far-reaching spiritual and humanitarian movement of the Enlightenment, best seen in the pedagogical views of John Locke and in the upsurge of philanthropy. Also important was naturalism, of which Rousseau can be regarded as the main representative. Ultimately, the Enlightenment was the most important of these trends, mainly because it most affected the political scene.

Locke's theory of knowledge

John Locke and the ideas of the Enlightenment. For 18th-century education, the late-17th-century philosopher John Locke is significant both for his general theory of knowledge and for his ideas on the education of youth. Philosophically, he was a thorough empiricist; that is, he was suspicious of the view that a thinker could work out in his study by pure reason alone the truth about the universe. Much as he admired Descartes, he was annoyed by this speculative spirit in him, and he despised it in the Scholastic and Jesuitical philosophers. In this sense he rejected metaphysics. Knowledge of the world around one could be gained only through sense perception or experience and by rational reflection on experience. It did not begin in innate knowledge of maxims or general principles.

For education this meant that Locke and like-minded men of the 18th century would emphasize the principles of practice and conditioning. By example, by habituation and practice, through encouragement, respect, and discredit, a child's good habits and intellect could be developed. Experience would be the watchword. "Children," he wrote,

> are not to be taught by rules which will be always slipping out of their memories. What you think necessary for them to do, settle in them by an indispensable practice as often as the occasion returns.

Some Thoughts Concerning Education (1693), in which these words appeared, was actually a compilation of replies written earlier to a friend and country squire who had requested advice about the education of his son. Thus, the work is somewhat random, written to no particular didactic end, though full of commonsense advice on the rearing of a squire's son, who was to be physically, mentally, socially, and morally prepared for life. Moral training was especially emphasized; character development should come first before learning; the educator's aim is to instill virtue, wisdom, and good breeding into the minds of the young. This does not mean ironhandedness, however; young children should be allowed to give vent to their feelings and should be restrained rarely; "the chief art is to make all that they have to do sport and play too." Locke sounds remarkably modern in this; and his writings in any case were enormously influential during the 18th-century Enlightenment.

The condition of the schools and universities. The school system became more and more in the 18th century an ordered concern of the state. Exponents of enlightened absolutism, as well as parliamentarians, recognized that the subject was of more use to the state if he had a school education. Ideally, there was to be compulsory schooling everywhere, but of course in practice the ideal was scarcely reached anywhere. The state also recognized that worthwhile school instruction depended on the standard of education of teachers: thus, the first teachers' colleges were established. But admittedly the standard of education of teachers was fairly poor. The teaching profession still did not provide a living wage, for which reason can be read from a regulation of 1736:

> If the teacher is a workman he can already support himself; if he is not, then he is hereby allowed to go to work for daily wages for 6 weeks at harvest-time (*Principia regulativa,* clause 10).

Ever since the 16th century, the universities had suffered a decline, mainly as a result of religious wars. Progress in the exact sciences was accomplished under government support in the academies of science, not in the universities, which became more and more training institutions for higher civil servants. There was, however, a notable change for the better, at least in Germany.

Halle, the first modern university

The year 1694 saw the foundation of the University of Halle, which has been described as the first real modern university. It originated in a *Ritterschule,* or "knight's school," imitative of the schools for chevaliers in France, and in 1694 the Holy Roman emperor Leopold I granted it a charter. The primary object in founding a university in Halle was to create a centre for the Lutheran party; but its character, under the influence of its two most notable teachers, the philosophers Christian Thomasius and August Hermann Francke, soon expanded beyond the limits of this conception. Thomasius was the first to set the example—soon after followed by all the universities of Germany—of lecturing in the vernacular instead of the customary Latin; this was a declaration of war against Scholasticism. Francke, as the founder of the Pietistic school, exercised great influence. Throughout the whole of the 18th century, Halle was the leader of academic thought and advanced theology in Protestant Germany, although sharing that leadership, after the middle of the century, with the University of Göttingen (founded 1737). With Göttingen, another important contribution was made by the revival of classical studies and the creation of a faculty of philosophy distinct from that of theology. This was designed not only to advance scholarship but also to train teachers. Halle itself established the first chair of educational theory.

The background and influence of Pietism. The dispute over the correct religious dogma, fought for almost 200 years with the utmost strength, controversy, and academic subtlety and reaching its terrible culmination in the Thirty Years' War, led to a certain ill feeling against dogmatically sanctioned religious revelation. There was a widespread trend toward secularization. Everywhere, there was a clear tendency to free belief from dogmatic quarrels. The search for a new belief took generally two different paths. One wanted to base belief in man's reason; the other wanted a godliness of the heart. For one line of thought, belief was a postulate of omnipotent human reason; for the other, man, corrupted by original sin, was to be saved only by simple belief in God's grace. The one path turned to the religious understanding of the Enlightenment; the other followed the subjective, mystical, zealous devoutness of Pietism. Such a movement away from the institutionalized church, away from the established church, and toward an intensified faith was evident in France within Catholicism in the form of Jansenism and Quietism. In England it was clearly evident in certain forms of Puritanism and in Independent movements and Quakerism. In Germany it was evident in Pietism.

Rationalistic versus devotional experience

Pietism was a Protestant movement of renewed faith that became popular from about 1675 to 1740, though it remained residually influential even into the 19th century. Its spiritual centres were in Württemberg, among the Moravian Brethren, and above all in Halle. Pietism was principally opposed to dogmatic Protestant orthodoxy, which usually included impatience and polemics against other beliefs. Pietism, on the contrary, stood for the renewal of importance of the individual prayer and for humility. The experiences of belief were to be based less in the acceptance of fixed conditions of belief and more in a mystical, personal submersion in feelings. Ac-

cording to standard Protestant theory, salvation could be hoped for only by the suppression of the corrupted individuality and by waiting for the grace of God to show one the way. From this came the Pietists' inclination to turn away from the world with its temptations (*e.g.*, the theatre, dancing, games, and other enjoyments). The uneasiness that they felt toward church institutionalization led to their splitting into numerous separatist groups; their subjective certainty about their belief led to a certain arrogance; and finally their seclusion led often to a joyless and moralizing way of life.

Although the founder of German Pietism is considered to be Philipp Jakob Spener, who established several private devotional gatherings (*collegia pietatis*) for Bible study in Frankfurt am Main and elsewhere, he was important for education only in the sense that he fashioned a spirit or concept in which education could be conducted—a concept that would subordinate all education to a simple Christian faith. This concept was realized mainly by his follower August Hermann Francke.

August Hermann Francke. Francke, after service as a grammar school teacher and priest in Leipzig, Lübeck, Hamburg, and Erfurt, was, through Spener's recommendation, given a post at the University of Halle in 1691, at the same time assuming the post of parish priest nearby. Motivated by the sad conditions of neglect in his parish, he quickly devoted himself to practical pastoral duties. In 1695 he instituted a vernacular school for the poor, popularly called the "ragged school," whose purpose was that the children should be led to a living knowledge of God and Christ and to a rightly accomplished Christianity. Through his activity and eloquence Francke quickly won several charitable patrons for his school, and the institution quickly expanded. After the school for the poor came the establishment of an elementary school for children of fee-paying burghers, then an orphanage, and lastly a *Pädagogium*, or boarding school, for the sons of nobility. Because Francke felt a lack of suitable teachers for his schools, he subsequently established two teachers' seminaries, *seminarium praeceptorum* and *seminarium selectum* (for teachers in higher schools). In 1697 there followed a Latin grammar school and in 1698, even if short-lived, a *gynaeceum*, a school for the daughters of nobility. To the whole complex of Halle's institutions (known collectively as the Halle Foundation) there also belonged a book shop with a publishing house and press, a very profitable chemistry laboratory, as well as four agricultural properties, a Bible institution, and an office for sending evangelical missions abroad. These institutions flourished, and about 1750 they were more and more brought under the control of the state. (In 1946 they were incorporated into the Martin Luther-Universität Halle-Wittenberg.)

The schools of Halle

Francke's main concern was ministerial work in the spirit of Pietism and not systematic educational theorizing. His educational aims were religious and at the same time practical. He himself paraphrased it as "true godliness and Christian wisdom"—true godliness meaning a pious, moral, devout life, and Christian wisdom referring to an ability to work hard according to the Protestant ethic. Francke's style of education went along with this aim: the corrupted willfulness of man must be broken; that does not happen through severe punishment but through "loving reproaches," a close supervision of the pupils, and a schooled and regimented care of the spirit. Games and childlike exuberance have no place in the system; thus, education had a joyless and moralizing effect and made heavy demands on the children, perhaps too heavy.

Strictness and realism in Pietistic schools

The harsh demands and regimentation are shown, for instance, in the daily timetable and the syllabus. The children arose at 5:00 AM; there was almost continuous instruction with frequent Bible reading and religious lessons until 7:00 in the evening. The grammar school had lessons in reading, writing, basic mathematics, catechism, the Holy Scriptures, Latin, Greek, Hebrew, optionally another Oriental language, geography, history, mathematics (including astronomy and geometry), botany, zoology, mineralogy, anatomy, and theology, as well as lathework, glass polishing, field trips to observe trades, factory work, horticulture, and so on. These latter subjects were counted as "recreation." The pansophic idea of Comenius was being followed here, in the sense that there was to be an all-encompassing education. It is worth noting that Francke was actually trying to inject realism into education—promoting, as he did, scientific subjects, lessons in manual skills, planned field trips, and even the reading of newspapers in the classroom.

Johann Julius Hecker. Julius Hecker came to Halle shortly before Francke's death in 1727 and became a teacher in the *Pädagogium*. In 1739 he was summoned by Frederick I of Prussia to Berlin, where he established a six-year *Realschule*, or "realist school," designed to prepare youth for the Pietistic and Calvinistic ideal of hard work and, especially, for the new technical and industrial age that was already dawning in countries such as England and France. Godliness was to be combined with a realistic and practical way of life. As early as 1699 Francke had conceived the idea of a school for children who were not meant for scholarship but who could serve usefully in commercial pursuits or administration; and in 1739 one of his teachers, Christoph Semler, published a pamphlet proposing such a "mathematical and mechanical *Realschule*." It was Hecker's fortune to put these plans into realization. His school included, among other things, classes for architecture, building, manufacturing, commerce, and trade. Both the exact sciences and manual skills were in the curriculum. A room for natural-history specimens, geographical maps, and realia was set aside for the illustration of lessons. Schools like Hecker's were gradually opened in other cities. In the 19th century the courses were extended to nine years, and such an institution was renamed *Oberrealschule*, or "higher realist school"; henceforth it was one of the main types of German secondary education.

Realschule, or "realist school"

Hecker is significant for two other, more minor contributions. He worked on a uniform program of education for teachers, *Kurmärkisches Landschullehrerseminar* (1753; "Seminary for Bell Ringers and Schoolmasters"). He was also entrusted by the King with the compilation of the general school regulations (1763), which fashioned the main outlines of the Prussian school system.

The background and influence of naturalism. Pietists emphasized Christian devotion and diligence as paths to the good life; Enlightenment thinkers focussed on reason and clear thinking as the sensible way to happiness. In Jean-Jacques Rousseau one finds a decisive change in the history of ideas; he gave a third answer. Only by reverting back to nature can the real, happy man come into being. For the naturalists of the Rousseau variety, "nature" is not a primitive or raw condition that man must conquer through culture. Nature is itself a governing principle, which should correct the unnatural development of mankind. Human reason can never fully take in the laws of "wise" nature; if man wants to see the play of nature, he must feel it.

Jean-Jacques Rousseau. Of those who thought and wrote about education in the 18th century no one has had so great an influence as Rousseau, whose *Émile* (1762) has often been called the charter of childhood. "The first thing," he wrote in his preface, "is to study your pupils more, for it is very certain that you do not know them." By stressing the importance of helping children to grow and develop naturally Rousseau started a revolution in educational theory that in time completely changed the relationship of teacher and pupil. His insistence that child nature is wholly good helped to destroy the harsh systems of discipline widely prevalent in his day, but it should be noted that it is a view at variance with Christian doctrine and with the findings of modern psychoanalysis. His version of education, so different from the artificial, conventional, formal teaching then usual, inspired generations to come.

The main ideas of Rousseau's education were these: attention to naturalness; development of independent judgment, harmony of needs, and abilities; and the founding of a general humanity and not a training of people to fill

Rousseau's educational ideas in *Émile*

set roles. The child was to be educated to be a man, not to be a priest, civil servant, or soldier. The first book of the educational novel *Émile* describes the period from birth to learning to speak. The most important thing for the healthy and natural development of the child at this age is that he learn to use his physical powers, especially the sense organs. The teacher must pay special attention to distinguishing between the real needs of the child and his whims and fancies. The second book takes in the time from the child's learning to speak to the age of 12. Games and other forms of amusement should be allowed at this age, and the child should by no means be overtaxed by scholarly instruction at too early an age. The child Émile is to learn through experience, not through words; he is to bow not to the commands of man but to necessities. The third book is devoted to the ages from 12 to 15. This is the time of learning, not from books of course but from the "book of the world." Émile must gain knowledge independently without long-winded studying, in concrete situations, which, if necessary, his tutor may provide. In this learning are included practical subjects (a trade, among other things). Not until the age of 15, described in the fourth book, does Émile learn of the world of morals, ethics, and religion. Here for the first time he comes face to face with history, comes into contact with other people, and finally goes to Paris to look for a wife. He is unable to find one, of course, since everything there is corrupt. He finds his Sophia in the country, in simple surroundings.

In Rousseau's view, the education of girls was to be similar with regard to naturalness, but it differed because of sexual differences. A girl cannot be educated as a man. She is more the centre of the family, a housewife, and a mother and should strive to please her husband, concern herself more than he with having a good reputation, and be satisfied with a simple religion of the emotions, since her intellectual education is not of the essence. "Her studies must all be on the practical side."

Even though Rousseau himself never had any success as a practical teacher, his theories certainly had a lasting influence. With his criticism of culture he opposed a naïve optimistic feeling of progress, which equated growth of knowledge with moral growth. By showing the child's right to a natural development relevant to his age he placed the teacher in the role of an adviser, and he also freed the child from the compulsion of having to be a miniature grown-up.

The Sensationists. A group of French writers contemporary with Rousseau and paralleling in some ways the thought of both Rousseau and Locke are known as the Sensationists, or, sometimes, the Sensationist psychologists. One of them was Étienne Bonnot de Condillac, who, along with Voltaire, may be said to have introduced Locke's philosophy to France and established it there.

Views of Condillac

The most important of Condillac's works is the *Traité des sensations* (1754), in which he imagined a statue organized inwardly like a man, animated by a soul that had never received an idea, into which no sense impression had ever penetrated. He then unlocked its senses one by one, beginning with smell. At its first experience of smell the consciousness of the statue was entirely occupied by it; and this occupancy of consciousness was attention. The statue's smell experience would produce pleasure or pain; and pleasure and pain would thereafter be the master principle determining all the interests and operations of the mind. The next stage was memory, which was the lingering impression of the smell experience upon the attention. From memory sprang comparison and from comparison judgment, giving rise to the powerful principle of the association of ideas. From comparison of past and present experiences in respect of their pleasure-giving quality arose desire; it was desire that determined the operation of man's faculties, stimulated the memory and imagination, and gave rise to the passions. Throughout the *Traité* Condillac examined all the senses and their combinations and came to the conclusion that man was nothing but what he had acquired; all notions of innate faculties and ideas were to be rejected. The important point for education is that, in Condillac's view, the teacher was to concentrate on training the child's sense observations and comparative understanding and to do so in a properly graded or progressive fashion. He had also to take cognizance of the pleasure–pain principle involved in experience.

Claude-Adrien Helvétius, a countryman of Condillac's professing much the same philosophy, was perhaps even more insistent that all men lacked any intellectual endowment at birth and, though of differing physical constitution, had the potential for identical passions and ideas. What made men different in later life were differing experiences. Hypothetically, two men brought up with the same chance experiences and education would be exactly the same. From this it followed, in education, that the teacher must attempt to control the environment of the child and guide his instruction step by step. Helvétius also contended that the only effective motivation was self-interest, founded on the love of pleasure and the fear of pain, and teachers should keep this in mind.

The Rousseauists. Rousseau left behind no disciples in the sense of a definite academic community, but hardly a single theorist of the late 18th century or afterward could avoid the influence of his ideas. One of those influenced was the German Johann Bernhard Basedow, who agreed with Rousseau's enthusiasm for nature, with his emphasis on manual and practical skills, and with his demand for practical experience rather than empty verbalism. The teacher, in Basedow's view, should take pains over the clearness of the lesson and make use of the enjoyment of games: "It is possible to arrange nearly all playing of children in an instructive way. . . ." In another respect, however, the contrast between Rousseau and Basedow could not be sharper; Basedow tended to force premature learning and overload a child's capabilities. A foreign language, for instance, was to be learned in six months. He promoted, in general, a pedagogic hothouse atmosphere. Basedow, however, was perhaps influenced by his seven-year-old daughter, who was put forward as a wonder child with extraordinary knowledge. He established an experimental school called a *Philanthropinum*, in Dessau, which lasted from 1774 to 1793.

Kant on pedagogy

The philosopher Immanuel Kant referred to Rousseau's influence on him. Kant dealt specifically with pedagogy only within a lecture he gave as holder of the chair of philosophy in Königsberg; the main features of the lecture were collected in a short work, *Über Pädagogik* (1803). In it he asserted, "A man can only become a man through education. He is nothing more than what education makes him." Education should discipline man and make him cultured and moral; its aim is ultimately the creation of a happier mankind. In general, it may be said that Kant agreed with Rousseau's education according to nature; but, from his ethical posture, he insisted that restraints be put on the child's passionate impulses and that the child even be taught specific maxims of conduct. The child must learn to rule himself and come to terms with the twin necessities of liberty and constraint, the product of which is true freedom.

> Children should be educated, not with reference to the present conditions of things, but rather with regard to a possibly improved state of the human race—that is, according to the ideal of humanity and its entire destiny. (From "On Pedagogy.")

The influence of nationalism. The Enlightenment was cosmopolitan in its effort to spread the light of reason, but from the very beginning of the age there were nationalistic tendencies to be seen in varying shades. Although Rousseau himself was generally concerned with universal man in such works as the *Social Contract* and *Émile*, his *Considérations sur le gouvernement de Pologne* ("Considerations on the Government of Poland") did lay out a proposal for an education with a national basis, and generally his ideas influenced the nationalistic generation of the French Revolution of 1789.

National views of the Philosophes

France. The real starting point generally of national pedagogic movements was in France. It perhaps began with the Philosophes, the rationalists and liberals such as Voltaire and Diderot, who emphasized the development of the individual through state education, not as

a means, of course, of adjusting to the state and its current government but as a means of creating critical, detached, responsible citizens. The Marquis de Condorcet was closely connected with this line of thought. For him man was by nature good and capable of never-ending perfection, and the goal of education should be the "general, gradually increasing perfection of man." He drafted a democratic and liberal but at the same time somewhat Socialist concept of school policy: there should be a uniform structure of public education and equal chances for all; ability and attainment should be the only standards for selection and careers; and private interests should be prevented from having influence in the educational system. An educational concept so rationalistic in its aims and with such a democratic and liberal structure cannot be narrowly nationalistic; it is cosmopolitan. But Condorcet was nationalistic insofar as he wanted "to show the world at last a nation in which freedom and equality for all was an actuality." He was, in fact, a strong supporter of the Revolution.

Many of the Rousseauists were nationalistic in a somewhat different way. They believed in a kind of "moral patriotism." They distrusted state-controlled nationalism and favoured instead a virtuous, patriotic citizen who experienced spontaneous feelings for his nation. Proper development in the family setting and in school would lead to the mastery of everyday situations and would naturally lay the foundations for this true nationalism.

Some of the French revolutionists, particularly Jacobins such as Robespierre and Saint-Just who were associated with the period of the Terror (1793–94), were concerned with an education for the revolutionary state, an education marked by an enmity toward scholarly knowledge, by state control, collectivism, the stressing of absolute equality, and the complete integration of all. What is good is decided by the collective "people." Thus, it could be said that the Jacobins favoured a complete politicalization of educational practice and theory. Connected with this was their support for private and boarding-school education on Spartan lines.

Benevolent despotism and education

National education under enlightened rulers. The absolutism of the 18th century has often been called "benevolent despotism," referring to the rule of such monarchs as Frederick II the Great, of Prussia, Peter I the Great and Catherine II the Great, of Russia, Maria Theresa and Joseph II of Austria, and lesser figures who were presumably sufficiently touched by the ideas of the Enlightenment to pursue social reforms. Their reforms were limited, however, and usually did not include anything likely to upset their sovereignty. Thus, they were often willing to improve education for middle class persons useful in civil service and other areas of state administration, but they were often chary of educating the poor. That risked upsetting the social order.

Frederick the Great, for example, was opposed to general education for all. For reasons of state, though, his minister Freiherr von Zedlitz founded a chair of pedagogy at Halle (1779) and generally planned for the improved education of teachers; he supported the founding of new schools and the centralization of school administration under an *Oberschulkollegium*, or national board of education (1787); and one of his colleagues, Friedrich Gedike, was instrumental in introducing the school-leaving examination for university entrance, the *Abitur*, which still exists. Zedlitz even acknowledged the work of philanthropists in education.

The guarded though increasingly liberal attempt by benevolent despots to nationalize and expand education is well illustrated by the events in Russia. Until the 18th century, schools in Russia were founded by ecclesiastical organizations (monasteries), the clergy (priests, deacons, readers), and private persons (boyars, or lower level aristocrats). Boys were taught reading, writing, arithmetic, singing, and religion. A system of state-owned schools was started by Peter the Great as a state organization for purposes of administration in peace and war and for the development of mining and industry. Peter did not intend to promote the Orthodox faith or formal classical learning, whether Greek, Latin, or Slavonic, or universal education. He created mathematical, navigation, artillery, and engineering schools for utilitarian purposes. In 1725 an Academy of Sciences with a university and a *gimnaziya* (secondary school) was founded at St. Petersburg. These utilitarian, secular, and scientific characteristics of Peter's schools became the dominant features of the Russian system, but, as a result of the many changes of policy of his successors after his death in 1725, they did not develop into a national system of education. A second attempt at building up a national system of schools was made by Catherine II. After many abortive schemes, Catherine issued in 1786 a statute for schools, which can be considered the first Russian education act for the whole country. According to this act, a two-year course in minor schools was to be started in every district town and a five-year course in major schools in every provincial town. Catherinian schools were also to be utilitarian, scientific, and secular. At the end of the 18th century, 254 towns had the new schools, but 250 smaller towns and the rural districts had no schools whatever. The third nationalizing attempt in Russia was made by Alexander I and was influenced by the disintegration of the serf system, by the development of industry and commerce, and by the ideas of the French Revolution. The new statutes (1803 and 1804) maintained the principles of utility and secular scientific instruction. The parochial schools (*prikhodskiye uchilishcha*) in the rural areas were to instruct the peasantry in reading, writing, arithmetic, and elements of agriculture; the district schools of urban areas (*uyezdniye uchilishcha*) and the provincial schools (*gimnazii*) were to give instruction in subjects necessary for civil servants—law, political economy, technology, and commerce. The system was state controlled and free and formed a continuous ladder to the universities. Later conservative reactions, however, tended to blunt or reverse these reforms.

The education act of Catherine II

England. In England the development of a "national" education took a completely different course. It was influenced not by a political but by an industrial revolution. It is true that theorists such as Adam Smith, Thomas Paine, and Thomas Robert Malthus proposed state organization of elementary schooling, but even they wanted to see limited state influence; the state could pay the musicians but not call the tune. Not until 1802 did Parliament intervene in the development of education, when the Health and Morals of Apprentices Act required employers to educate apprentices in basic mathematics, writing, and reading. For the most part this remained only a demand, since the employers were not interested in such education.

The reluctance on the part of the state induced several philanthropists to form educational societies, principally for the education of the poor. In 1796, for example, was founded the Society for Bettering the Conditions of the Poor. A further impulse for elementary education stemmed from the Sunday schools, the first of which was founded in 1780 in Gloucester; by 1785 their numbers had so increased that the Sunday School Society was founded. The lessons in such schools, however, were mainly those of Bible reading.

The monitorial, or Lancasterian, system

Educators Andrew Bell and Joseph Lancaster played a major role in progress toward an elementary school system. They realized that the root of the problem lay in the lack of teachers and in the lack of money to hire assistants. Therefore, first Bell developed, then Lancaster modified, the so-called monitorial system (also called the Lancasterian system), whereby a teacher used his pupils to teach each other. The use of children to teach other children was not new, but Bell and especially Lancaster took the approach and developed it into a systematic plan of education. From 200 to 1,000 children were gathered in one room and seated in rows, usually of ten pupils each. An adult teacher taught the monitors, and then each monitor taught his row of pupils the lesson in reading, writing, arithmetic, spelling, or higher subjects. Besides monitors who taught, there were, in Lancaster's system, monitors to take attendance, give examinations, issue supplies, and so on; school activity was to be directed with military precision; the emphasis was on drill and memori-

zation. The system and the publicity connected with it expanded the efforts toward mass education, even though, pedagogically, the whole process was so routinized and formalized that opportunities for creative thinking or initiative scarcely existed. Also Lancaster's statement that "Any boy who can read can teach" proved to be false. The only valid, long-range answer to teacher shortages would be the expanded training of teachers. (H.-J.I.)

EUROPEAN OFFSHOOTS IN THE NEW WORLD

Spanish and Portuguese America. With the Spanish conquerors of the New World, the conquistadores, came friars and priests who immediately settled down to educate the Indians and convert them. Because no concept of education as a governmental responsibility existed, the Roman Catholic Church assumed complete control of elementary education, and the early Franciscan and Dominican friars were followed by Augustinians, Jesuits, and Mercedarians.

Church control of colonial education

The first elementary school in the New World was organized in Mexico by the Franciscan Pedro de Gante in 1523 in Texcoco, followed in 1525 by a similar school in San Francisco. Because such schools in Mexico were designed for Indian children, the monks learned the native languages and taught reading, writing, simple arithmetic, singing, and the catechism. The schools of the *hospicio* of the bishop Vasco de Quiroga in Michoacán added agriculture, trades, and crafts to their curriculum.

Mestizo children, the issue of Spanish and Indian parents, were often abandoned; thus, special institutions appeared to collect and educate them—for example, the Colegio de Niñas and the Colegio de San Juan de Letrán, founded by Viceroy Mendoza in New Spain, and the Bethlehemite schools of Guatemala and Mexico.

In the beginning, children of Spaniards born in the colonies, called Creoles, had tutors. Eventually, schools promoted by municipal authorities (*cabildos*) emerged, especially in Chile and Argentina (*escuelas del rey*).

During the 18th century the Enlightenment came to Latin America, and with it a more secular and widespread education. Among famous projects were those of Viceroy Vertiz y Salcedo in Argentina and two model schools, free for children of the poor, by Archbishop Francos y Monroy in Guatemala. In New Spain the Colegio de las Vizcaines (1767) became the first all-girl lay institution.

Because of the social structure, riches and administrative privilege were held by an elite, the Creoles. Secondary education was scarce and originally existed only in the monasteries, where the grammar and Latin colleges prepared the clergy and some lay Creoles who wanted to enter the universities. There also existed a few special colleges or secondary institutions for the Indian nobility, the first outstanding one being Santa Cruz de Tlaltelolco (1535–95) in Mexico, which produced teachers of native languages and liberal arts. Similar colleges existed in Bogotá, Quito, Cuzco, Charcas, Santiago, and Chillán. Though the Dominicans had been pioneers in all such education, the Jesuits became the most important teachers. They offered an efficient education, molded to contemporary requirements, in boarding schools, where the elite of the Spaniards born in the Americas studied. When their order was expelled in 1767, education was dealt a severe blow. In Portuguese Brazil, where the expulsion edict had been issued eight years earlier and where they had been the only educators, the royal chancellor was forced to make feeble attempts toward organizing a secular education. The Spanish king, Charles III, also took advantage of the occasion and founded some new institutions—the Academy of San Carlos, the School of Mining in Mexico, the Royal College of San Carlos in Buenos Aires—and modernized others.

The Latin American universities

Traditionally, Spanish universities had been organized on the model either of Paris or of Bologna. The former was a *universitas magistrorum*, governed by professors organized in faculties, whereas the latter, as a *universitas scholarium*, received its corporate authority from the student body organized into "nations" that elected leaders to whom even the professors were subject. In 1551 the Council of the Indies authorized the founding of the first American universities, one in Mexico and one in Lima; academic government was placed in the hands of a *claustro*, or faculty, composed of the rector, the teachers, and the professors. Dedicated to general studies, the universities required a papal as well as a royal authorization.

The Real y Pontificia Universidad de Mexico was the first to open its doors, in 1553. In the Spanish colonies eventually ten major and 15 minor universities came into existence. The latter were actually colleges or secondary institutions—nine Jesuit, four Dominican, one Franciscan, and one Augustinian—which, because they were located far from the closest university (minimally 200 miles), were given special authorization to grant higher degrees. In Brazil no university existed, and Portuguese born in the colony had to go to Portugal for study.

Though in Spain itself law reigned supreme, in the Americas theology became the principal chair. Teaching was in scholastic mode: it began with the reading of a classical text; then the professor explained the thesis or proposition and offered arguments pro and contra so that a conclusion in accord with Catholic dogma would result. (J.V.de K.)

French Québec. Soon after the founding of the Québec colony in 1608, the first organized educational activity began with missionary work among the Indians, carried on mainly by members of the Récollet and Jesuit orders and, from 1639, by Ursuline nuns. The first mission "school" recorded was that of Brother Pacifique du Plessis, established in 1616 in Trois-Rivières (Three Rivers).

Christian efforts among the Indians were only an environmental dimension of the religious purposes that framed educational activity in Old World France. Roman Catholic social philosophy allowed no compromise in the spiritual direction of education, and both in informal socialization patterns and in what formal provisions existed the doctrine and aim of religion coincided with that of education. At the general level, education was intended to produce religious conformation in thought and behaviour; at the higher level, education was to produce a progeny of clerical leadership. The paternalistic authority of church and monarch was carried from the Old to the New World, where it perhaps became even more pervasive, due to the initial absence of alternative institutional developments. In education, the exclusive role of the state (though not insignificant) was confined to financial subsidization. Authority for the institution of education was vested in the bishop.

Most of the nonreligious functions now associated with formal education were, in the 17th and 18th centuries, carried in other institutional sectors: the family, the community, the vocation. Just as there was no sharp break between church and school in formal learning, there was an easy transition between the information and behaviour necessary for work and life as transmitted in the course of various socialization experiences. Thus, the self-sustaining and isolated life of the farmers, the wild and solitary ways of the *coureurs de bois* (the fur traders), the miniature of European manners and customs established in the cities by the gentry—all contained within their own cycle the educative procedures for life in that society. Education as a separate institution was understandably associated with learning not related to the business of life.

The schools of New France

Institutional forms found in French colonial Québec included parish schools, girls' schools, secondary schools, and vocational schools; and literacy records indicate that the provision for education was in sum comparable to that in the Old World. Parish or common schools were irregularly provided to afford the rudiments of literacy and religion. Because of the relative sparseness of educational resources, social classes were frequently mixed in these schools. Girls' schools were established by the Ursulines from 1642 and by the Sisters of the Congrégation de Notre Dame from 1659, with a rudimentary curriculum but including a characteristic "finishing" of social graces appropriate to the French-Canadian girl. Vocational training was probably of least concern in this early period, but specific attempts to institutionalize this educa-

tional area were begun as early as 1668 with the establishment of L'École des Arts et Metiers in Saint Joachim, for instruction in agriculture and certain trades.

Secondary education was offered by the Jesuits from 1636. The Jesuit college, offering early training for eventual entrance into the priesthood, was conducted along characteristically Jesuit lines: militaristic discipline in conduct, unequivocal authority in method, classical curriculum in content. The classical curriculum pattern, comprising basically Latin, Greek, mathematics, philosophy, and theology, was to be essentially preserved in the French-Canadian development of *collèges classiques* for secondary education.

In 1663, Bishop Laval established in the city of Québec the *grande seminaire* as the apex of the educational "system," as the first French-Canadian "university." Shortly thereafter, he also established the preparatory *petit seminaire*.

Following the cession of Québec to Britain in 1763, education fell prey to political and cultural disruption. Although the British military and colonial government attempted to preserve the structure of French civil and religious institutions, the cultural integrity of the system was inevitably broken. Financial grants from France for education discontinued and were not replaced by the British government; recruitment to religious orders was restricted; and educational development was obstructed by the continual association of educational plans with cultural–religious controversies. The end of the 18th century saw French-Canadian education fall backward into neglect. (R.F.L.)

British America. *New England.* The year 1630, chronicled in New England annals as the beginning of the Great Migration, witnessed the founding there of Puritanism as the established religion. Rejecting democracy and toleration as unscriptural, the Puritans put their trust in a theocracy of the elect that brooked no divergence from Puritan orthodoxy. So close was the relation between state and church that an offense against the one was an offense against the other and, in either case, "treason to the Lord Jesus." The early Puritans also put their confidence in centralized church governance; however, geographic reality forced them to settle for a localized, congregational administration, for impossible roads made land travel over any distance onerous and even dangerous, and thus the focal point of social and political life had to be the village. Small and constricted, a place where the vital necessities, sacred and profane, were within walking range of all and where one's conduct was exposed to constant public watch, the New England village was the prime mover of communal life.

Puritan education in New England

In Puritan moral theology the young, like the old, were sinners doomed by almost insurmountable odds to perdition. Not even infants were spared. To God, indeed, they were depraved, unregenerate, and damned. Hence, the sooner the young learned the ground rules of the good society, as revealed in the Bible, the better. The task of teaching them first befell the parents. Later, when they were old enough, the burden was conferred upon the school. And schools were not long in coming. The first was probably at Boston, which, in 1635, sought the services of a schoolmaster. The year following, Charlestown undertook to pay a teacher out of the town purse. Three years thereafter, Dorchester made fiscal provision to maintain a permanent town school. What these towns inaugurated others imitated, so that by the 1640s the idea of town-supported schooling had lost its novelty.

If towns braved the first steps in education, then the Commonwealth of Massachusetts did not trail far behind. In 1642 it ordered parents and masters of apprentices to see to it that their charges were instructed in reading, religion, and the colony's principal laws. Five years later, the General Court reinforced this enactment with yet another. Aimed, as it explained, at the "old deluder Satan," it undertook to thwart him from keeping "men from a knowledge of the Scriptures," by requiring every township of 50 households to commission someone to teach reading and writing. In addition, the law directed towns of 100 families to furnish instruction in Latin grammar so that youth might be "fitted for the university." Finally, the measure required teachers to be paid by "parents or masters . . . or by the inhabitants in general." Nevertheless, the measure was given only a pallid obedience, and directly and indirectly numerous towns evaded its requirements. More pertinent, doubtless, to subsequent times, is its underlying assumption that the state may compel the schooling of its young and that in order to support education it may impose taxes.

The school founded in Boston in 1635 was the Boston Latin School. The first secondary school in colonial America, Boston Latin was modelled on the grammar schools of England, which is to say that it put an overwhelming emphasis on the ancient languages and "humane learning and good literature." Taking its novices at an age of seven, when they could read rudimentary English, the school belaboured them with Latin grammar. Once they were its master, they were put upon reading the salient Latin literati. After six years of such industry they proceeded similarly with Greek. Sometimes, when circumstances were propitious, a bright lad was set to work on Hebrew. The legatee of the classical humanistic tradition but rescored to attune it to the Reformation, the Boston Latin School was the most famous of its kind in New England. But there were several others of high merit, among them Cambridge, Dorchester, and Roxbury, all of which, like Boston, are still in practice.

The founding of Harvard College

The first colonists had scarcely settled when, in 1636, the General Court appropriated £400 "towards a school or college." When, two years later, John Harvard died and left the institution his library and some £800, the grateful founders honoured their school with his name. Designed to train youth for important Puritan places, particularly in the ministry, the college accepted only those who could read, write, and speak Latin in prose and verse, besides knowing Greek nouns and verbs familiarly. Once admitted, the student was lodged at the college, pledged to a blameless behaviour, and put upon a prescribed four-year course of grammar, rhetoric, logic, arithmetic, geometry, astronomy, ethics, ancient history, Greek, and Hebrew. If he weathered these hazards, he was made a bachelor of arts (B.A.), and, if ambition still rowelled him, he could enroll for another three years to become a master of arts (M.A.).

So things sat until the century's passing. Then, swayed by the intellectual breezes of Europe's Enlightenment, Harvard College ventured some earnest renovation. Its texts, cobwebbed with Aristotelianism, were replaced with newer ones by Locke and Newton. In 1718 it added mathematics and sciences to its offerings, and 20 years later it enriched itself with a professorship of mathematics and natural philosophy. There were the usual grumblings from conservatives. In truth, in 1701 a number of Congregational parsons, all Harvard sons, distressed by their alma mater's dalliance in newfangled ideas, inaugurated a grove of their own, the collegiate school of Connecticut, now Yale University. Harvard, however, remained unchastened.

The new academies. Disdainful of the challenging intellectual values, the Latin schools continued in their classical tracks. By the 18th century, however, their tradition was playing out, especially among the rising nabobs of the marketplace. When the old schools failed to respond to their demands for an education calculated to prepare their sons for everyday living, they resorted to private schooling. From such endeavour emerged the academy. The first school of strictly native provenance, it made its advent in 1751 in Philadelphia (the Philadelphia Academy), the work in the main of Benjamin Franklin. What differentiated it from its classical antecedent was its hearty promotion of "useful learning," to wit, the vernacular, modern languages, history, geography, chronology, navigation, mathematics, natural and applied science, and the like.

The first academies addressed themselves solely to boys, but time saw them vouchsafe instruction to girls in a "female department," which in turn gave way to the "female academy." Private ventures always, academies generally were not loath to solicit outside assistance—some, in-

deed, as in New York, enjoyed a public subsidy. Whatever their special character, to their very end they maintained their original purpose of bringing education into closer consonance with "the great and the real business of living," as Phillips Academy of Andover, Massachusetts, phrased it when, in 1778, it held its first sessions. To this purpose they taught anything for which there was a sufficient call, from the subjects proffered by the original Philadelphia Academy to several dozen more.

Diversity of the middle colonies

The middle colonies. The religious uniformity that marked the Puritan theocracy was missing in the middle colonies. From New York through Delaware there flourished a host of sects whose scriptural interpretations were diverse—often, in fact, in collision. Nor was there even the tie of a common language, for the settlers came from many lands. Divergent in religion and language, the bedrock in those times of elementary schooling, the middle colonists could not accommodate themselves, as did the Puritans, to a single school teaching reading and religion to all the children of the neighbourhood. Instead, they depended on parish or parochial schools, each of them free to teach by its own denominational lights. True, for a time New Netherland, with its established Dutch Reformed Church, maintained some town schools, but, after the English seized the colony (renaming it New York), such endeavours ceased. Pennsylvania, linguistically and denominationally the most heterogeneous of the colonies, began its educational history by ordering the erection of public schools and the instruction of children. But the ordinance fell prey to powerful sectarian antagonisms, and in 1701 the colony essayed to make peace by sanctioning the establishment of parochial schools.

Like the New Englanders, the middle colonists aspired to establish colleges, but, with no friendly lawmakers to sustain them, they found their task heavily hobbled, and the mid-1700s were upon them before their hopes materialized with the advent, in 1746, of the College of New Jersey (Princeton). There followed King's College (Columbia) in 1754; the College and Academy of Philadelphia (Pennsylvania) in 1755; and Queen's College (Rutgers) in 1766. Common to these schools was their stress on the ancient languages, metaphysics, and divine science. At the same time, however, one discerns signs of a new liberalism. Both Rutgers and Columbia announced their interdenominationalism. Pennsylvania offered courses in physics, and in 1765 it became the first colonial college to sponsor systematic instruction in medicine.

Effects of plantation life on Southern education

The Southern colonies. Unlike New Englanders, Southerners resided not in villages but on widely scattered plantations. For years, town life was impossible and so, per consequence, were town schools. But even had their establishment been feasible, the odds against them were staggering, since the ruling classes, like their analogues overseas in England, were averse to schooling the young under governmental direction. Instead, they regarded education as a personal concern, the affair of parent and church rather than the state. Left thus to their own devices, Southerners schooled their young to suit their taste, the rich resorting to tutors and private schools and the rest scratching out an education as best they could. Time saw the appearance of a number of free schools serving those who were neither rich nor poor. For the offspring of the low-down and unregarded folk, Virginia enacted its law of 1642. An echo of England's Poor Law, it provided for the "relief of such parents whose poverty extends not to give them [the children] breeding." For this purpose it ordered the creation of a "workhouse school" at James City to which each county was to commit two children of an age of six or over. There, besides being reared as Anglicans, they were to be "instructed in honest and profitable trades and manufactures as also to avoid sloth and idleness." Amended several times, the statute became the model for similar legislation throughout the South.

If Southerners dawdled in schooling the multitude, in the higher learning they moved more quickly. As early as 1619 the Virginia Company reserved £1,500 and several thousand acres of land to found a university at Henrico, but a massacre of nearly all the inhabitants shattered the plan. Several other attempts followed, but not until 1693 did the South obtain a college. Named for its royal sponsors, William and Mary College was chartered to propagate the "Liberal Arts and the Christian Faith," with particular stress on preparing young men for the Anglican pulpit. To accomplish this, it offered preparatory courses in ancient languages, followed by the conventional collegiate curriculum of rhetoric, logic, ethics, mathematics, physics, and metaphysics, besides sacred and moral science, including "the commonplaces of divinity and the controversies with heretics." As the 18th century swept on, the secular interest that had invaded Harvard appeared in Virginia, and there ensued a waning of the earlier religious motivation. In 1779, led by Thomas Jefferson, the college trustees refurbished the school with chairs in medicine, mathematics, physics, moral philosophy, economics, law, and politics. Concurrently, it discontinued the chair in divinity as "incompatible with freedom in a republic." (A.E.M.)

Newfoundland and the Maritime Provinces. Newfoundland was, during most of this period, under British control, and, though there were settlers even before the 17th century, the island was not considered a settlement colony. Other than for naval training and fishing advantages, the British government had no concern for Newfoundland. Thus, policies were constructed with regard to the rights and advantages of English seamen, while, implicitly as well as in overt regulations, settlement was obstructed and restricted. Destruction from the running French–English military conflicts further discouraged development. These conditions of economic and political diminution of the settlement from outside were aggravated by the usurious conduct of merchants and the corruption of officials and by the national and religious divisions among the inhabitants themselves.

With such substantial problems of mere survival in Newfoundland, it is not surprising that the luxury of formal education was almost absent during this period. Some accounts verify that informal, unorganized efforts were made on an occasional basis to convey minimum schooling to settlers' children, but the only organized effort was that of the Society for the Propagation of the Gospel in Foreign Parts (SPGFP). The SPGFP founded or aided a school in Bonavista in 1722 and in St. John's in 1744 and sponsored schools in more than 20 settlements between 1766 and 1824. Religion was undoubtedly more important than education as such to the society, but its provision of reading materials as well as the mere act of establishing some kind of school filled a notable void in the Newfoundland settlement. Other charitable societies, such as the Society for Improving the Condition of the Poor in St. John's, the Benevolent Irish Society, the Newfoundland School Society (later the Colonial and Continental Church Society), the Wesleyan Society, the Sisters of the Presentation, the Sisters of Mercy, and the Irish Christian Brothers, carried the charity-school work into the 19th century and maintained a thread of education through the colonial "dark ages."

Society for the Propagation of the Gospel in Foreign Parts

For a time after 1763, the Maritimes were all one colony—Nova Scotia—but Prince Edward Island was separated in 1769 and New Brunswick in 1784. This area comprised a heterogeneous population of French Acadians, English Protestants and others from Europe, Highland Scots, and loyalists from the United States. Each of these groups carried attitudes more or less favourable to education, and the regionalization of these attitudes, together with other conditions, influenced the differential development of education in the area. At the end of the 18th century, for example, New Brunswick, with a high loyalist population promoting political and educational development, probably ranked highest among the Maritime colonies in educational interest.

The first relatively organized attempt at common schooling in the Maritimes was made by the SPGFP, closely connected to the Church of England. The society opened both weekday and Sunday schools, and it might be said that it fostered teacher training in stipulating qualifications for its teachers. Other than SPGFP schools, education in the Maritime colonies was carried on by itin-

erant teachers and in scattered private-venture schools. Schools for separate ethnic or religious groups were discouraged by the Anglicans, but consistent pressure for such schools did succeed, at least temporarily; for example, in Lunenburg, Nova Scotia, and in Sydney, on Cape Breton Island. An interesting note is the establishment of a school for blacks in Halifax in 1788.

Upper schools were established only toward the end of the 18th century in the Maritimes. As they were established singularly and recruited from a social class rather than from a lower school, there is no clear line of demarcation among the various types as there would be later in an integrated system. Basically, they were Anglican and classical, although the private schools, advertising to as wide a clientele as possible, often included some breadth, extending into practical studies. Probably the most influential of the early attempts were the two Latin grammar schools founded in 1788 and 1789 at Windsor and Halifax, Nova Scotia. The former became associated with King's College, established in Windsor at the same time. Thomas McCulloch's Academy at Pictou, Nova Scotia, and the College of New Brunswick at Fredericton, both founded around the turn of the century, should also be mentioned as early exemplars of higher education.

(R.F.L.)

VIII. Western education in the 19th century

THE SOCIAL AND HISTORICAL SETTING

From the mid-17th century to the closing years of the 18th century, new social, economic, and intellectual forces steadily quickened—forces that in the late 18th and the 19th centuries would weaken and, in many cases, end the old aristocratic absolutism. The European expansion to new worlds overseas had stimulated commercial rivalry. The new trade had increased national wealths and encouraged a sharp rise in the numbers and influence of the middle classes. These social and economic transformations, joined with technological changes involving the steam engine and the factory system, together produced industrialism, urbanization, and the beginnings of mass labour. At the same time, intellectuals and philosophers were assaulting outworn economic abuses, old unjust privileges, misgovernment, and intolerance. Their ideas, which carried a new emphasis on the worth of the individual—the citizen rather than the subject—helped not only to inspire political revolutions, sometimes successful, sometimes unsuccessful, but, more important, to make it impossible for any government, even the most reactionary, to disregard for long the welfare of the common man. Finally, there was a widespread psychological change: man's confidence in his power to use resources, master nature, and structure his own future was heightened beyond anything known before; and this confidence on a national scale—in the form of nationalism—moved all groups to struggle for the freedom to direct their own affairs.

Education as the responsibility of the state

All these trends influenced the progress of education. One of the most significant results was the gradual acceptance of the view that education ought to be the responsibility of the state. Some countries, such as France and Germany, were inspired by a mixture of national aspiration and ideology to begin the establishment of public educational systems early in the 19th century. Others, such as Great Britain and the United States, under the spell of laissez-faire, hesitated longer before allowing the government to intervene in educational affairs. The school reformers in these countries had to combat the prevailing notion that "free schools" were to be provided only for pauper children, if at all; and they had to convince society that general taxation upon the whole community was the only adequate way to provide education for all the children of all the people.

The new social and economic changes also called upon the schools, public and private, to broaden their aims and curricula. Schools were expected not only to promote literacy, mental discipline, and good moral character but also to help prepare children for citizenship, for jobs, and for individual development and success. Although teaching methods remained dominantly oriented toward textbook memorizing and strict discipline, a more sympathetic attitude toward children began to appear. As the numbers of pupils expanded rapidly, individual methods of "hearing recitations" by children began to give way to group methods. The monitorial, or Lancasterian, system became popular because, in the effort to overcome the shortage of teachers during the quick expansion of education, it enabled one teacher to use older children to act as monitors in teaching specific lessons to younger children in groups. Similarly, the practice of dividing children into grades or classes according to their ages—a practice that began in 18th-century Germany—was to spread everywhere as schools grew larger.

THE EARLY REFORM MOVEMENT: THE NEW EDUCATIONAL PHILOSOPHERS

Influence of Rousseau

The late 18th and 19th centuries represent a period of great activity in reformulating educational principles, and there was a ferment of new ideas, some of which in time wrought a transformation in school and classroom. Of those who thought and wrote about education, no one was to have so great an influence as Rousseau (discussed in the preceding section), who, by stressing the importance of helping children to grow and develop naturally, started a revolution in educational theory that in time largely changed the relationship between teacher and pupil. His version of education, so different from the artificial, conventional, formal teaching then usual, inspired others to continue the revolution that he initiated. Among those influenced in the 19th century were Johann Pestalozzi, Friedrich Froebel, and Johann Friedrich Herbart.

Pestalozzianism. The theories of the Swiss reformer Johann Heinrich Pestalozzi laid much of the foundation of modern elementary education. Beginning as a champion of the underprivileged, he established near Zürich in 1774 an orphanage in which he attempted to teach neglected children the rudiments of agriculture and simple trades in order that they might lead productive, self-reliant lives. A few years later the enterprise failed, and Pestalozzi turned to writing, producing his chief work on method, *How Gertrude Teaches Her Children*, in 1801, and then began teaching again. Finally in 1805 he founded at Yverdon his famous boarding school, which flourished for 20 years, was attended by students from every country in Europe, and was visited by many of the most important figures of the time, including the philosopher J.G. Fichte, the educators Friedrich Froebel and Johann Friedrich Herbart, and the geographer Carl Ritter.

The pedagogy of Pestalozzi. In spite of the quantity of his writings, it cannot be said that Pestalozzi ever wrote a complete and systematic account of his principles and methods; an outline of his theories must be deduced from his various writings and his work. The foundation of his doctrine was that education should be organic, meaning that intellectual, moral, and physical education (or, in his words, development of "head, heart, and body") should be integrated and that education should draw upon the faculties or "self-power" inherent in the human being. Education should be literally a drawing out of this self-power, a development of abilities through activity—in the physical field by encouraging manual work and exercises, in the moral field by stimulating the habit of moral actions, and in the intellectual field by eliciting the correct use of the senses in observing concrete things accurately and making judgments upon them. Words, ideas, practices, and morals have meaning only when related to concrete things.

Child-centred education

From these overarching principles there followed certain practical rules of educational method. First, experience must precede symbolism. There must be an emphasis on object lessons that acquaint the child with the realities of life; and the effort to relate experience and abstract thought (and language) must be gradual and progressive. A person gets knowledge by his "own investigation, not by endless talk about the results of art and science." A person "is much more truly educated through that which he does than through that which he

learns" secondhand. Next, the program should be child centred. There must be an allowance for individual differences (whenever possible, grouping children according to ability), and there must be freedom to learn. The teacher's task is to offer "continual benevolent superintendence—giving a helping hand to the instinctive efforts" of the child. "The nature of the child must determine all the details of his education." This also means that the stages of education must be related to the stages of child development. "Each child should be taught that which he has to learn at the time his nature calls for it, for this is proof that his sensibility and power are ready for it." Finally, education should be correlated and well rounded. Intellectual, moral, and physical activities should be as one. All these ideas, which may sound commonplace to today's reader, were extraordinarily innovative for the times.

Self-help for the poor

Much of Pestalozzi's pedagogy, it should be noted, was influenced by the fact that his main work was with children of the poor. Thus, there was a strong emphasis on education in the home. There was a strong emphasis on the development of skills. (The philosopher Fichte accused Pestalozzi of trying to make children learn reading and writing too early but, as an admirer of Pestalozzi, acknowledged that the man was swayed by the need to fit the children for making their own living as soon as possible.) Pestalozzi also placed heavy emphasis on manual training, integrating it with other activities. In effect, whereas the reformers of the Enlightenment and the French Revolution had stressed, and were stressing, the "emancipation" of the lower classes, men like Pestalozzi, in a realistic vein, aimed at helping poor people to help themselves, helping them to get rid of a sense of failing and develop a spirit of enterprise. This was social reform, not social revolution. Indeed, Pestalozzi viewed revolutionaries as "shipwrecked women, [who] saw every straw as a mast by which the republic could reach the shore."

The influence of Pestalozzi. "The art of education," Pestalozzi claimed, "must be significantly raised in all its facets to become a science that is to be built on and proceeds from the deepest knowledge of human nature." By his own efforts in this direction, he stimulated pedagogic theory and practice to an enormous degree in many parts of the Western world. By his philanthropic efforts in behalf of the poor, he inspired new movements toward the reform of philanthropic educational institutions and the pedagogy applied to such institutions; he created a new methodology for elementary education that was introduced not only into schools but also into programs of teacher education in Europe and America; and by his own example he gave teachers a high professional ethos. Against these achievements his weaknesses pale into insignificance: his lack of organizing ability in trying to realize his ideas; some formalistic exaggerations in method. Pestalozzi was a restless, passionate person, self-critical to the point of being self-tormenting, who, like few others at any time, recognized and sincerely tried to alter the misery existing in the world. If the Enlightenment saw its pedagogic mission as the spreading of the light of reason, then Pestalozzi showed that it was not reason alone but above all love that would show a way out of the "mire of the world."

Pestalozzi's influence abroad

It is hardly possible to name all of Pestalozzi's disciples —the Pestalozzians—for almost all the pedagogic figures of his time literally or figuratively went to his school. His influence was most profound in Germany, especially in Prussia and Saxony. Generally speaking, in the first half of the 19th century the English school system was completely under the influence of the disciplinarian monitorial systems of Bell and Lancaster (see preceding section). Pestalozzi for most Englishmen was "a distressing type of the German" and "an idealistic dreamer," as some critics put it. Nevertheless, he exercised some influence in England through James Pierrepont Greaves and the London Infant School Society and through Charles and Elizabeth Mayo and the Home and Colonial School Society. In the United States, Pestalozzianism was introduced by a Philadelphia scientist and philanthropist, William Maclure, one of the sponsors of the utopian colony at New Harmony, Indiana. The movement became extremely popular, extending from New England and New York to St. Louis.

In Switzerland itself, in Hofwil near Bern, Philipp Emanuel von Fellenberg founded an institution for the education of the poor. He tried to build up a kind of pedagogic province or miniature state, in which work was the means of self-help and in which the pedagogic program was the joint responsibility of teachers and pupils. His educational aim was "the harmonious development and education of all the combined talents and abilities in human nature with which God has entrusted each person."

Froebel and the kindergarten movement. Next to Pestalozzi, perhaps the most gifted of early 19th-century educators was Friedrich Froebel, the founder of the kindergarten movement and a theorist on the importance of constructive play and self-activity in early childhood. He was an intensely religious man who tended toward pantheism and has been called a nature mystic. Throughout his life he achieved very little literary fame, partly because of the style of his prose and philosophy, which is so romantically obscurantist as to be difficult to read and sometimes scarcely comprehensible.

In early life, Froebel tried various kinds of employment until in 1805 he met Anton Gruner, a disciple of Pestalozzi and director of the normal school at Frankfurt am Main, who persuaded him to become a teacher. After two years with Gruner, he visited Pestalozzi at Yverdon, studied at Göttingen and Berlin, and eventually determined upon establishing his own school, founded on what he considered to be psychological bases. The result in 1816 was the Universal German Educational Institute at Griesheim, transferred the following year to Keilhau, which constituted a kind of educational community for Froebel, his friends, and their wives and children. To this period belongs *The Education of Man* (1826), his most important treatise, though typical of his obscurantism. In 1831 he was again in Switzerland, where he opened a school, an orphanage, and a teacher-training course. Finally, in 1837, upon returning to Keilhau, he opened his first *Kindergarten*, or "garden of children," in nearby Bad Blankenburg. The experiment attracted wide interest, and other kindergartens were started and flourished, despite some political opposition.

The first kindergarten

The pedagogy of Froebel. Froebel's pedagogic ideas have a mystical and metaphysical context. He viewed man as a child of God, of nature, and of humanity who must learn to understand his own unity, diversity, and individuality, corresponding to this threefold aspect of his being. On the other hand, man must understand the unity of all things (the pantheistic element).

> Education consists of leading man, as a thinking, intelligent being, growing into self-consciousness, to a pure and unsullied, conscious and free representation of the inner law of Divine Unity, and in teaching him ways and means thereto.

Education had two aspects: the teacher was to remove hindrances to the self-development or "self-activity" of the child, but he was also to correct deviations from what man's experience has taught is right and best. Education is thus both "dictating and giving way." This means that ordinarily a teacher should not intervene and impose mandatory education, but when a child, particularly a child of kindergarten age, is restless, tearful, or willful, the teacher must seek the underlying reason and try to eradicate the uncovered hindrance to the child's creative development. Most important, the teacher's dictating and giving way should not flow from the mood and caprices of the teacher. Behaviour should be measured according to a "third force" between teacher and child, a Christian idea of goodness and truth.

School, for Froebel, was not an "establishment for the acquisition of a greater or lesser variety of external knowledge"; actually, he thought children were instructed in things they do not need. School instead should be the place to which the pupil comes to know the "inner relationship of things," "things" meaning God, man, nature, and their unity. The subjects followed from this: religion, language and art, natural history, and the knowl-

edge of form. In all these subjects the lessons should appeal to the pupil's interests. It is clear that, in Froebel's view, the school is to concern itself not primarily with the transmission of knowledge but with the development of character and the provision of the right motivation to learn.

Froebel's emphasis on play

Froebel put great emphasis on play in child education. Just like work and lessons, games or play should serve to realize the child's inner destiny. Games are not idle time wasting; they are "the most important step in the development of a child," and they are to be watched by the teachers as clues to how the child is developing. Froebel was especially interested in the development of toys for children—what he called "gifts," devised to stimulate learning through well-directed play. These gifts, or playthings, included balls, globes, dice, cylinders, collapsible dice, shapes of wood to be put together, paper to be folded, strips of paper, rods, beads, and buttons. The aim was to develop elemental judgment distinguishing form, colour, separation and association, grouping, matching, and so on. The important point is that the children were to play with these things not as they wished but as organized or subtly guided by the teacher. "The child must feel some relationship to you in his activity, as you feel to God in everything you do." This high evaluation of the personal relationship in education is also clear in Froebel's concept of the kindergarten.

The kindergarten movement. The kindergarten was unique for its time. Whereas the first institutions for small children that earlier appeared in Holland, Germany, and England had been welfare nursery schools or day-care centres intended merely for looking after children while parents worked, Froebel stood for the socializing or educational idea of providing, as he put it in founding his kindergarten, "a school for the psychological training of little children by means of play and occupations." The school, that is, was to have a purpose for the children, not the adults. The curriculum consisted chiefly of three types of activities: (1) playing with the "gifts," or toys, and engaging in other occupations designed to familiarize children with inanimate things: (2) playing games and singing songs for the purpose not only of exercising the limbs and voice but also of instilling a spirit of humanity and nature; and (3) gardening and caring for animals in order to induce sympathy for plants and animals. All this was to be systematic activity.

The kindergarten plan to meet the educational needs of children between the ages of four and six or seven through the agency of play thereafter gained widespread acceptance. During the 25 years after Froebel's death in 1852, kindergartens were established in leading cities of Austria, Belgium, Canada, Germany, Great Britain, Holland, Hungary, Japan, Switzerland, and the United States. In Great Britain the term infant school was retained for the kindergarten plan, and in some other countries the term *crèche* has been used (see also PRESCHOOL EDUCATION).

Widespread acceptance of kindergartens

Herbartianism. Johann Friedrich Herbart was a contemporary of Froebel and other German Romanticists, but he can hardly be put into the ranks of such pedagogues. During his lifetime his sober, systematic "philosophical realism" found little approval; and only posthumously, during the latter half of the 19th century, did his work achieve great importance. He is regarded as one of the founders of theoretical pedagogy, injecting both metaphysics and psychology into the study of how people learn.

The psychology and pedagogy of Herbart. As a young man of 18, Herbart had studied at the University of Jena under the Idealist philosopher Fichte. It was a long while before he broke from the spell of Fichte's teachings, which denied the existence of things independent of the consciousness, and turned to philosophical Realism, which asserts that underlying the world of appearances there is a plurality of things or "reals" that have existence apart from the operations of the mind. Change consists simply in the alteration in the relations between these reals, which resist the changed relationships as a matter of self-preservation.

Ideas, as things, likewise always exist and always resist change and seek self-preservation. It is true that some ideas may be driven below the threshold of consciousness; but the excluded ideas continue to exist in an unconscious form and tend, on the removal of obstacles (as through education), to return spontaneously to consciousness. In the consciousness there are ideas attracting other ideas so as to form complex systems. These idea masses correspond to the many "interests" of the individual (such as his home and his hobbies) and to broader philosophical and religious concepts and values. In the course of mental development certain constellations of ideas acquire a permanent dominance that exercises a powerful selective facilitating influence upon the ideas struggling to enter or re-enter the consciousness.

In his systematic account of the nature of education, Herbart conceived the process as beginning with the idea masses that the child has previously acquired from experience and from social intercourse. The teacher creates knowledge from the former and sympathy from the latter. The ultimate objective is the formation of character by the development of an enlightened will, capable of making judgments on right and wrong. Moral judgments (like reals) are absolute, springing from contemplation, incapable of proof and not requiring proof. Ethics, in other words, is the ultimate focus of pedagogy.

In the classroom, it is the aim of the lessons to introduce new conceptions, to bind them together, and to order them. Herbart speaks of "articulation." First the student becomes involved in a particular problem; then he considers its context. Each of these two stages has a phase of rest and of progress, and thus there are four stages of articulation: (1) clarification, or the static contemplation of particular conceptions; (2) association, or the dynamic linking of new conceptions with old ones; (3) systematization, or the static ordering and modification of what in the conceptions are deemed of value; and (4) methodization, or the dynamic application and recognition of what has been learned. Herbart phrased this system of instruction only in very general terms, but his successors tended to turn this framework into a rigid schedule that had to be applied to every lesson. Herbart himself warned:

"Articulation," or the systematization of lessons

> We must be familiar with them [the methods], try them out according to circumstances, alter, find new ones, and extemporize; only we must not be swallowed up in them nor seek the salvation of education there.

Herbart's ethics and psychology are today considered old-fashioned. The tendency toward a schematic, rigid type of instruction, as adopted and simplified by his disciples and widely practiced for decades, was strongly criticised by reformers at the beginning of the 20th century. Indeed, criticism of Herbart sometimes became too negative and overlooked his merits: the establishment of pedagogy as an autonomous science; the discovery of interest and spontaneity as part of the educational process; and the attempt to give instruction a psychologically based rationale.

Herbartians. Herbart's followers, who practiced in the latter half of the 19th century, were mainly concerned with applying the master's thoughts to school practices and teacher training. Among the followers were Tuiskon Ziller in Leipzig (who was the founder of the Association for Scientific Pedagogy, or Verein für wissenschaftliche Pädagogik), Wilhelm Rein in Jena, and Friedrich Wilhelm Dörpfeld in Barmen. Ziller's ideas are representative. He insisted that all parts of the curriculum be closely integrated and unified—history and religion forming the core subjects on which everything else was hinged. The sequence of instruction was to be adjusted to the psychological development of the individual, which was seen as corresponding to the cultural evolution of mankind in stages from primitive savagery to civilization. His main aim in education, like the aim of most Herbartians, was promoting character building, not simply knowledge accumulation.

Other German theorists. In the history of pedagogy there is no period of such fruitfulness as the 19th century in Germany. In addition to Herbart, Froebel, Pestalozzi

(in German Switzerland), and their followers, there were scores of the most important writers, philosophers, and theologians contributing their ideas on education—including Friedrich Schiller, Johann Wolfgang von Goethe, G.W.F. Hegel, Friedrich Ludwig Jahn, Johann Paul Friedrich Richter, Ernst Moritz Arndt, Friedrich Nietzsche, and Julius Langbehn. To list the many ideas and contributions of these figures and others is impossible here, but it is worthwhile to suggest briefly the work of three men—Johann Gottlieb Fichte, Friedrich Schleiermacher, and Wilhelm von Humboldt—representing three divergent views.

When the great heterodox University of Berlin was founded in 1809, Fichte became one of its foremost professors and a year later its second rector, having already achieved fame throughout Germany as an Idealist philosopher and fervent nationalist. At a time when Napoleon had humbled Prussia, Fichte in Berlin delivered the powerful *Addresses to the German Nation* (1807–08), full of practical views on national recovery and glory, including suggestions on the complete reorganization of the German schools along Pestalozzian lines. All children would be educated—and would be educated by the state. Boys and girls would be taught together, receiving virtually the same education. There would be manual training in agriculture and the industrial arts, physical training, and mental training, the aim of which would be not simply the transmission of measures of knowledge but rather the instillation of intellectual curiosity and love and charity toward all men. Unlike Pestalozzi, however, Fichte was wary of the influence of parents and preferred educating children in a "separate and independent community," at least until a new generation of parents had arisen, educated in the new ideas and ideals. Here was an apparent revival of Plato's idea of a strictly ordered, authoritarian state.

Fichte's ordered education

Another of the founders of the University of Berlin (teaching there from 1810 to 1834) was the Protestant theologian Friedrich Schleiermacher, who sounded a very modern note by offering a social interpretation of education. Education, in his view, was an effort on the part of the older generation to "deliver" the younger generation into the four spheres of life—church, state, social life, and science. Education, however, not only assumes its organization in terms of these four areas of life but also serves to develop and influence these areas.

Perhaps more than any other individual, the philologist and diplomat Wilhelm von Humboldt was responsible for the founding of the University of Berlin. Supported by the king of Prussia, Frederick William III, he adopted for it principles that raised it to a foremost place among the universities of the world—the most important principle being that no teacher or student need adhere to any particular creed or school of thought. This academic freedom survived in Germany despite its temporary suspension and Humboldt's dismissal by a reactionary Prussian government in 1819. Philosophically and pedagogically, Humboldt was himself a Humanist—a part of a wave of what were called New Humanists—who reasserted the importance of studying the classical achievements of humanity in language, literature, philosophy, and history. The aim of education in these terms was not the service of society or the state but rather the cultivation of the individual. Though generally commendable, this view sometimes tended toward a certain egoistic detachment from society, a measure of disdain for the work-a-day world, an occasional overemphasis on style of utterance, and a kind of aestheticizing of education.

Humboldt's Humanism

French theorists. At this time there were two men in France who were making their names through the introduction of new methods—Jean-Joseph Jacotot and Édouard Séguin. Jacotot was a high school teacher, politician, and pedagogue, whose main educational interests focussed on the teaching of foreign languages. "You learn a foreign language," he said, "as you learn your mother-language." The pupil is confronted with a foreign language; he learns a text in the language almost by heart, compares it with a text in his own native language, and then tries gradually to free himself from the comparison of texts and to construct new combinations of words. The teacher controls this learning by asking questions. "My method is to learn one book and relate all the others to it." The learning of grammar came later.

Jacotot's method was successful. Its fundamentals are partly old—first the practical side and then the rule, constant repetition, and self-activity on the part of the pupils. His method is also partly questionable. This is especially true of Jacotot's two basic theses: (1) that everyone has the same intelligence, differences in learning success being only a case of differences in industry and stamina; and (2) that everything is in everything: "Tout est dans tout," which suggests that any subject or book is analogous to any other. Jacotot's efforts to extend his method into other subjects were also doubtful.

The doctor and psychologist Édouard Séguin developed a pedagogy for pupils of below-average intelligence. Historically, scientific attempts to educate mentally retarded children had begun with the efforts of a French doctor, Jean-Marc-Gaspard Itard, during the latter part of the 18th century. In his classic book, *The Wild Boy of Aveyron* (1801), Itard related his five-year effort to train and educate a boy found, at about the age of 11, running naked and wild in the woods of Aveyron. Later, Séguin, a student of Itard, devised an educational method using physical and sensory activities to develop mental processes. Limbs and the senses were, in his view, a part of the whole personality, and their development was a part of the whole human education. In his book *Idiocy* he wrote,

Education of the mentally retarded

> From the distinction between a circle and a square up to that between right and wrong, we have followed a continuous path, beginning where the function awakes to the perception of simple notions, finishing where the faculties refuse to soar higher in the atmosphere of idealism.

Séguin's published works influenced Maria Montessori (see the next section on the 20th century), an Italian pediatrician who became an educator and developed methods of training young mentally retarded children in Rome in the 1890s. Her methods, like Séguin's, were based on sensorimotor training. (H.-J.I.)

DEVELOPMENT OF NATIONAL SYSTEMS OF EDUCATION

The great changes in Europe in the 19th century included, among other things, the further consolidation of national states, the spread of modern technology and industrialization, and increasing secularization. These changes had consequences for the design of school systems. National school systems had to be conceived and organized. Alongside the older schools, including elementary schools, Latin, or grammar, secondary schools, and universities, there developed so-called modern schools that stressed the exact sciences and modern languages, reflecting the new technological and commercial age. Vocational schools also appeared in greater numbers. The influence of the church was increasingly repressed, and the influence of the state on the school system correspondingly grew stronger. The ideal of universal education—education for all—became more and more a reality.

Germany. Martin Luther's pronouncements on the educational responsibilities of the individual had no doubt helped create that healthy public opinion that rendered the principle of compulsory school attendance acceptable in Prussia at a much earlier date than elsewhere. State intervention in education was almost coincident with the rise of the Prussian state. In 1717 Frederick William I ordered all children to attend school if schools were available to them. This was followed in 1736 by edicts for the establishment of schools in certain provinces and in 1763 by Frederick II the Great's regulation asserting the principle of compulsory school attendance.

Humboldt's reforms. The schools, however, had established a traditional classical curriculum that ignored the changing needs of life and fields of knowledge. No effective reorganization of the educational system was carried out until after the disaster of the Battle of Jena (1806), during the Napoleonic Wars, which brought about the virtual collapse of Prussia and at the same

Reforms inspired by the Napoleonic disasters

time introduced a period of reform in the government as a whole. One of the first acts of the prime minister Freiherr Karl von Stein in 1807 was to abolish certain semiecclesiastical schools and to place education under the Ministry of the Interior, with Wilhelm von Humboldt at the head of a special section. Humboldt's policy in secondary education was a compromise between the narrow philological pedantry of the old Latin schools and the large demands of the new Humanism that he espoused. The measure introduced by Humboldt in 1810 for the state examination and certification of teachers checked the then-common practice of permitting unqualified theological students to teach in the schools and raised the teaching profession to a high level of dignity and efficiency, placing Prussia in the forefront of educational progress. It was also a result of the initiative of Humboldt that the methods of Pestalozzi were introduced into the teachers' seminaries. To this period also belong the revival, in 1812, of the *Abiturientenexamen* ("the school-leaving examination"), which had fallen into abeyance. About the same time, the local administrations, called *Schulvorstände* for the country and *Schuldeputationen* for the towns, were instituted.

Developments after 1815. The period that succeeded the peace of 1815 was one of political reaction, and not until the 1830s were there further significant reforms. In 1834, for example, an important step was taken in regard to secondary education by making it necessary for candidates for the learned professions, as well as for the civil service and for university studies, to pass the leaving examination of the *Gymnasien*, the classical secondary schools. Thus, through the leaving examination, the state held the key to the liberal careers and was thereby able to impose its own standards upon all secondary schools.

In connection with the so-called Kulturkampf, or struggle between the state and the Roman Catholic Church, the school law of 1872 reasserted the absolute right of the state alone to the supervision of the schools. Nevertheless, the Prussian system remained both for Catholics and for Protestants essentially denominational. On the elementary level, in particular, the mixed school was established only when the creeds were so intermingled that a confessional school was impracticable. In all cases the teachers were appointed with reference to religious faith; religious instruction was given in school hours and was inspected by the clergy.

Types of German secondary schools

The official classification, or grading according to the type of curriculum, of secondary schools in Prussia (and throughout Germany) was very precise. The following were the officially recognized types: (1) the classical nine-year *Gymnasium*, with a curriculum that included Latin, Greek, and a modern language; (2) the semiclassical nine-year *Realgymnasium*, with a more modern curriculum that included, in addition to Latin and modern languages, the natural sciences and mathematics; and (3) the modern six-year *Realschule* or nine-year *Oberrealschule*, with a curriculum of sciences and mathematics. (There were also some variations on these basic types, modifying the length of the courses.)

The differentiation between the types was the result of a natural educational development corresponding to the economic changes that transformed Prussia from an agricultural to an industrial state. The classical schools long retained their social prestige and a definite educational advantage in that only their pupils were admissible to the universities. From the foundation of the German Empire in 1871 the history of secondary education was largely concerned with a struggle for a wider recognition of the work of the newer schools. The movement received a considerable impetus by the action of Emperor William II, who summoned a school conference in 1890 at which he set the keynote: "It is our duty to educate young men to become young Germans and not young Greeks or Romans." New schedules were framed in which the hours devoted to Latin were considerably reduced, and no pupil could obtain a leaving certificate without a satisfactory mark in the mother tongue. The reform lasted only a single school generation. In 1900 equality of privileges was granted to three types of schools, subject to certain reservations: the theological faculties continued to admit only students from classical schools, and the pupils of the *Oberrealschule* were excluded by their lack of Latin from the medical faculties; but insofar as Latin was required for other studies, such as law or history, it could be acquired at the university itself.

Although the official curricula were binding on the schools, their rigidity was not absolute; experiments were possible, but they were carefully supervised. The chief innovation was the postponement of the beginning of Latin to the fourth year and the introduction of French as the first foreign language. This enabled parents to defer their decision as to the form of their son's education until he was about 12 years of age.

A further instance of the willingness of the authorities to sanction reasonable changes was seen in the permission accorded to certain schools to vary the course of study in the top classes, as a preparation for the freedom of choice of the university.

Girls' schools. In Prussia, as elsewhere, the higher education of girls lagged far behind that of boys and received little attention from the state or municipality, except insofar as the services of women teachers were needed in the elementary schools. Thus it came about that in Prussia secondary schools for girls were dealt with administratively as part of the elementary school system. After the establishment of the German Empire in 1871, a conference of directors and teachers of these schools was held at Weimar and put forth a reasoned plea for better organization and improved status. The advocates of reform, however, were not at unity in their aims; some wished to lay stress on ethical, literary, and aesthetic training; others stressed intellectual development and claimed an equal share in all the culture of the age. But even in the schools the women fought an unequal battle, for all the heads and a large part of the staff were men, usually academically trained. The women continually demanded a larger share of the work, and this was secured by the establishment of a new higher examination for women teachers. University study, though not prescribed, was in fact essential, and yet women had not the right of access to the university in Germany. They were allowed to take the leaving examination, for which private institutions prepared them, but their admission to the university rested with the professor. Not until the 20th century were desired changes achieved.

The new German universities. Unquestionably one of the greatest worldwide influences exercised by German education in the 19th century was through its universities, to which students came from all over the world and from which every land drew ideas for the reformation of higher education. To understand this, one must be aware of the state of higher education in most countries in the 19th century. Although the century witnessed a steady expansion of scientific knowledge, the curriculum of the established universities went virtually untouched. Higher education followed a single dimension. This was the century of the scientists Michael Faraday, Hermann von Helmholtz, James Prescott Joule, Charles Darwin, Joseph Lister, Wilhelm Wundt, Louis Pasteur, and Robert Koch. Yet, until the end of the century, most of the significant research was done outside the walls of higher educational institutions. In Great Britain, for instance, it was the Royal Society and other such societies that fostered advanced studies and encouraged research. The basic curriculum of colleges and universities remained nontechnical and nonprofessional. The English cardinal John Henry Newman, lecturing in Dublin on *The Idea of a University* in 1852, stated that the task of the university was broadly to prepare young men "to fill any post with credit, and to master any subject with facility." The university ought not to attempt professional and technical education.

Condition of 18th- and 19th-century higher education

While Newman's words epitomized the views held in most colleges and universities in Europe and America, some of the new universities in Germany were moving toward the expansion of the educational enterprise. In 1807 Fichte had drawn up a plan for the new University

of Berlin, which Humboldt two years later was able to realize in its founding. The school was dedicated to the scientific approach to knowledge, to the combination of research and teaching, and to the proliferation of academic pursuits; and its ideal was adopted in the founding or reconstitution of other universities—Breslau (1811), Bonn (1818), Munich (1826). By the third quarter of the 19th century the influence of German *Lernfreiheit* (freedom of the student to choose his own program) and *Lehrfreiheit* (freedom of the professor to develop the subject and to engage in research) was felt throughout the academic world. The unity of the universities, for better or worse, was more and more dissolved by the fragmentation of subjects into different branches. Some critics would eventually condemn what they considered to be the excesses of the free elective system and the extreme departmentalization of research and curricula. Much of the debate, however, would centre on the general education of undergraduates. The conviction, fathered in Germany, that investigation and research were a main function and responsibility of institutions of higher education seemed secure.

Lernfreiheit and *Lehrfreiheit*

France. In France the Jesuit schools and the schools of other teaching orders created at the time of the Renaissance had reconciled the teaching of the new Humanism with the established doctrines of the Roman Catholic Church and flourished with special brilliance. But, despite the changes brought about by the Renaissance and the attention given to the sciences in the 17th and 18th centuries, it was not until the advent of the French Revolution that the universal right to education was proclaimed (1791).

Unfortunately, the Revolution had little opportunity to put these principles into effect before Napoleon restored the old system, with his own concession—a provision for the study of mathematics. He did, however, reorganize the structure of secondary and higher education in a unified state system, with secondary schools maintained by the *communes*, and with state lycées, universities, and special institutions of higher education. Within this structure the rector of a university headed a teaching body, recruited by the state and supervised by an inspectorate, ranging through various grades up to the university council. Grades of proficiency in studies, from simple certificates to the degrees of *baccalauréat*, *licence*, and doctorate were awarded on the result of examinations, and these tests were made a necessary condition of entry into such professions as medicine, law, and teaching. This structure, despite many modifications, has survived until modern times.

Napoleonic system of education

Development of state education. French educational history in the 19th century is essentially the story of the struggle for the freedom of education, of the introduction at the secondary level of the modern and scientific branches of learning, and, under the Third Republic, of the establishment of primary education, at once secular and compulsory, between the ages of six and 12. There were also a middle education between the ages of 13 and 16 and, finally, a professional and technical education.

Under the restoration of the monarchy in 1814, education fell inevitably under the control of the church; but, during the bourgeois monarchy of Louis Philippe, a law was passed in 1833 that laid the foundations of modern primary instruction, obliging the *communes* to maintain schools and pay the teachers. The higher primary schools that were founded were suppressed by Roman Catholic conservatives in 1850 (their restoration later constituted one of the great positive services rendered by the Third Republic to the cause of popular education). The 1850 law restored the "liberty of teaching" that, in effect, meant free scope for priestly schools, but it also made provision for separate communal schools for girls, for adult classes, and for the technical instruction of apprentices. In 1854 France was divided for purposes of educational administration into 16 (now 17) districts called *académies*, each administered by a rector and each with a university at the apex of the educational structure. The rector not only was made the chief administrator of the university but also was responsible for secondary and higher education within his *académie;* he nominated candidates for administrative positions in his area, appointed examination committees, supervised examination content and procedures, and presided over a personal cabinet, called an academic council. Unlike the political division in some other countries, the *académies* were given little power or authority of their own; rather, they were administrative arms of the national ministry of education. They represented a devolution rather than a delegation of power from the central government.

Centralization of French education

After the Franco-Prussian War, the Third Republic addressed itself to the organization of primary instruction as "compulsory, free, and secular." The law of 1878 imposed on *communes* the duty of providing school buildings and provided grants-in-aid. The national government also henceforth paid salaries throughout the public sector of education. In 1879 a law was passed compelling every department to maintain training colleges for male and female teachers. The law of 1881 abolished fees in all primary schools and training colleges; the law of 1882 established compulsory attendance; and, finally, the law of 1886 enacted that none but lay persons should teach in the public schools and abolished in those schools all distinctively religious teaching.

Secondary education. In European systems of higher education, secondary education was pre-eminently a preparation for the university, with aims and ideals of general culture that differentiated it radically and at the very outset from education of the elementary type. Down to the beginning of the 20th century, the French system could be regarded as a typical and extreme example of the European theory.

The characteristic European organization has been called the "dual plan": elementary and secondary education were distinct types, and only a minority of the elementary school pupils passed on to the secondary schools, generally only if they were bright and could win scholarships through a competitive examination. The secondary schools drew their clientele from private preparatory schools, called the lycées in France. Furthermore, the courses in the lycées were definite and were limited to the study of French language and literature, foreign languages (ancient or modern), mathematics and science, history and geography, some art, music, and physical training. Pupils who did not complete the full course generally entered civil-service or other white-collar occupations. With the development of commerce and industry in the 19th century, France instituted the *écoles primaires supérieures*, or "higher primary schools," for those who did not go on to universities but who needed a better education than the primary schools could give. The curricula of these schools were somewhat more advanced than those of the primary schools; pupils remained longer (up to the age of 16) and were prepared for employment in business as white-collar workers, but generally at a lower level than pupils who came from the lycées. In effect, the different types of schools tended to maintain class cleavages since students of the secondary schools enjoyed higher social and occupational prestige than those of the upper primary schools.

European "dual plan"

The foundation of secondary schools for girls was in its way one of the most notable achievements of the Third Republic. It was inaugurated by the law of December 22, 1880, called after its author the Loi Camille Sée. Until World War II, the curricula were different from those of the boys' schools, and the course of study was only five years. There were no ancient languages, and mathematics was not carried to so high a level as in the boys' lycées.

England. Influenced by doctrines of laissez-faire, England hesitated a long time before allowing the state to intervene in educational affairs. At the beginning of the 19th century, education was regarded as entirely the concern of voluntary or private enterprise, and there was much unsystematic philanthropy. Attempts were made to canalize and concentrate it, and many hoped that the Church of England and the dissenting churches would join in a concerted effort to provide a national system of elementary education on a voluntary basis. But discor-

Laissez-faire and English voluntary education

dant views prevented such cooperation, and two voluntary societies were founded, one representative of the Church of England and the other of dissent. In 1829 the Catholics were emancipated by law from disabilities they had long suffered, and so they also were able to provide voluntary schools. Other religious bodies joined in the effort to meet the growing need for elementary schools, but it was soon evident that voluntary finance would not be equal to this formidable task. In 1833 the government made a small building grant to these societies, and in this modest way state intervention began. Six years later a committee of the Privy Council was established to administer the state grants, now made annually, and to arrange for the inspection of voluntary schools aided from public funds. The work involved led to the establishment of a small central education department, which was the beginning of the ministry of education.

Elementary Education Act. England prolonged its reliance on voluntary initiative for year after year as population increased, and, with the growing industrialization, people crowded into the new towns. At last in 1870 Parliament, after long, acrimonious debates, passed an Elementary Education Act, the foundation upon which the English educational system has been built. Religious teaching and worship were the crucial issues in the debates, and the essentials of the settlement agreed upon were (1) a dual system of voluntary and local-authority schools and (2) careful safeguards to ensure as far as possible that no child would receive religious teaching at variance with his parent's wishes. It was left to the school boards—as these first local education authorities were called—to decide whether to make elementary education compulsory in their districts. In 1880, however, it was made compulsory throughout England and Wales, and in 1891 fees were abolished in all but a few elementary schools.

Secondary and higher education. Secondary education, however, was still left to voluntary and private enterprise. Attention was focussed on the "public" schools (independent secondary schools such as Eton and Harrow, usually for boarders from upper and well-to-do middle class homes), which under the leadership of outstanding headmasters such as Thomas Arnold were thoroughly reformed. Several new universities were founded during the 19th century, and the latter half of it saw the founding of a number of girls' high schools and boarding schools offering an education comparable to that in boys' public schools and grammar schools. Several training colleges for teachers were established by voluntary agencies, and universities and university colleges toward the end of the century undertook the training of postgraduates as teachers in departments of education created for this purpose.

Russia. Influenced by the disintegration of the serf system, the development of industry and commerce, and the ideas of the French Revolution, Tsar Alexander I at the beginning of the 19th century tried to institute new educational reforms. The statutes of 1803 and 1804 maintained the principles of utility and secular scientific instruction fostered by Peter I the Great and Catherine II the Great in the 18th century. The parochial schools (*prikhodskiye uchilishcha*) in the rural areas were to instruct the peasantry in reading, writing, arithmetic, and elements of agriculture; the district schools of urban areas (*uyezdnye uchilishcha*) and the provincial schools (*gimnazii*) were to give instruction in subjects necessary for civil servants (law, political economy, technology, and commerce). The system was state controlled and free, and it formed a continuous ladder leading to the universities. During Alexander's reign the old Catherinian schools were remodelled and new schools founded.

Russian reform and reaction

Nicholas I, coming to the throne in 1825, considered this democratic trend harmful and decreed that

> It is necessary that in every school the subjects of instruction and the very methods of teaching should be in accordance with the future destination of pupils, that nobody should aim to rise above that position in which it is his lot to remain.

A new statute of 1828 decreed that parochial schools were intended for the peasants, the district schools for merchants and other townspeople, and *gimnazii* for children of the gentry and civil servants. Instruction in the *gimnazii* in Latin and Greek was increased. Although the legislation of Nicholas I established a class system, one way for the gentry and another for the rest, the utilitarian character of the whole system remained.

The Russian radical intelligentsia was fiercely opposed to the privileged schools for the gentry and demanded the re-establishment of a democratic system with a more modern curriculum in secondary schools. This was coupled with the demand for the emancipation of the serfs and the equality of women in education. The new tsar in 1855, Alexander II, inaugurated a period of liberal reforms. The serfs were emancipated in 1861, and thus all social restrictions were removed. A new system of local government in rural areas (*zemstvo*) was enacted with a right to found schools for the peasantry, now free. Combined efforts of the government, *zemstva*, and peasant communities produced a growth of schools in the rural areas. The utilitarian trend was evident in the establishment of technical schools with vocational differentiation. The education of women was promoted, and the first higher courses for women were founded in main cities.

The reign of Alexander II, which was later marked by reactionary measures and political oppression, ended in his assassination in 1881 by the terrorist branch of the Narodniki revolutionary organization. A period of reaction followed under his successor, Alexander III. All reforms were suspended, and the growth of educational institutions was interrupted. The chief procurator of the Holy Synod attempted to build up a rival system of parochial schools under the control of the orthodox clergy; and the Minister of Public Instruction tried to return to the class system of Nicholas I. These reactionary measures set back the growth of education. Four-fifths of all children were deprived of education. The result was that at the turn of the century nearly 70 percent of Russia's male population and 90 percent of its female population were illiterate (1897 census.) The aboriginal dwellers of Russia's national outskirts (more than one-half of the country's population) were almost totally illiterate. (Ed.)

The United States. Administered locally everywhere, schooling of the United States's masses in the republic's younger days was immensely diverse. In New England, primary schooling enjoyed public support. In the South, apart from supplying a meagre learning to pauper children, the states abstained from educational responsibility. In the middle states elementary schools were sometimes public; more often they were parochial or philanthropic. Only beyond the Alleghenies in the huge national domain was there any federal provision for education. There, under the Articles of Confederation, the Ordinance of 1787 reserved a plot of land in every prospective township for the support of education. The measure not only laid the groundwork for education in the states of the Ohio Valley and the Great Lakes; it also became a precedent for national educational aid. Thus, in 1862 the Morrill Act granted every state establishing a public agricultural college 30,000 acres (12,000 hectares) of public land for each of its lawmakers in Congress. Since then some 12,000,000 acres (5,000,000 hectares) have been distributed, on which some 70 of the so-called land-grant colleges currently flourish.

American land-grant colleges

Several of the Founding Fathers expressed belief in the necessity of public education, but only Thomas Jefferson undertook to translate his conviction into actuality. Convinced that democracy can be effective only in the hands of an enlightened people, he offered Virginia's lawgivers a plan in 1779 to educate school children at public cost for three years and a few gifted boys beyond that. The proposal elicited some support but encountered resistance not only from the ruling classes, who insisted that education was a private responsibility, but also from the clergy, who regarded instruction of the common child as an ecclesiastical prerogative. Hence, Jefferson's plan was rejected. Some 40 years later he submitted another proposal for public schools, but again the legislators voted it down. Only one of Jefferson's dreams fructified

—the University of Virginia. When it opened in 1825, it was the most up-to-date institution of its sort, the first frankly secular university in America and the closest to a modern-day conception of a state university.

The educational awakening. When Jefferson died in 1826, the nation stood on the threshold of a stupendous transformation. During the ensuing quarter century it expanded enormously in space and population. Old cities grew larger and new ones more numerous. The era saw the coming of the steamboat and the railroad. Commerce flourished, and so did agriculture. The age witnessed the rise of the common man with the right to vote and hold office. It was a time of overflowing optimism, of dreams of perpetual progress, moral uplift, and social betterment.

The common school

Such was the climate that engendered the common school. Open freely to every child and upheld by public funds, it was to be a lay institution under the sovereignty of the state, the archfather, in short, of the present-day American public school. Bringing the common school into being was not easy. Against it bulked the doctrine that elementary school was essentially for pious ends and that any education that excluded religious instruction—as all state-maintained schools were legally compelled to do—was godless. Nor had there been any great recession of the contention that education was not a proper governmental function, and for a state to engage therein was an intrusion into parental privilege. Still more distasteful was the fact that public schooling would occasion a rise in taxes.

Yet the common school also mustered some formidable support, and finally, in 1837, liberal Massachusetts lawmakers successfully carried through a campaign for a state board of education. It is especially to Horace Mann, the board's first secretary, that Massachusetts credits its educational regeneration. Something of a dreamer, but a hardened pragmatist withal, Mann combined a lucrative legal practice with the labours of a legislator in the Massachusetts senate. There he supported a host of liberal causes, from more humane jails to religious freedom and better schools. To gather data on educational conditions in Massachusetts, Mann roved the entire commonwealth. He lectured and wrote reports, depicting his dire findings with unsparing candour. There were outcries against him, but when Mann resigned, after 12 years, he could take pride in an extraordinary achievement. During his incumbency, school appropriations almost doubled. Teachers were awarded larger wages; in return they were to render better service. To help them Massachusetts established three state normal schools, the first in America. Supervision was made professional. The school year was extended. Public high schools were augmented. Finally, the common school, under the authority of the state, though still beset by difficulties, slowly became the rule.

What Mann accomplished in Massachusetts, Henry Barnard (1811–1900) achieved in Connecticut and Rhode Island. More reserved than his Massachusetts counterpart, Barnard has come down the ages as the "scholar of the educational awakening." He is seen in this capacity as a founder of the Association for the Advancement of Education. He became its first president and editor of its *American Journal of Education*, in whose 30 volumes he discussed virtually every important pedagogical idea of the 19th century.

Although New England led the field in the drive to modernize education, similar campaigns were under way in other states. Though the pattern of reform was similar, there were differences nonetheless. In Pennsylvania the assault centred on the pauper school; in New York it was against sectarianism. Aggravating the contest everywhere, however, was the controversy over the state's right to assume educational authority, and especially its power to levy school taxes.

Rise of the United States high school

Secondary education. The soil that rooted the common school became the seedbed for the high school. Some of its essence is discernible in the 18th-century academy with its accent on modern and useful learning. But the academy catered to a paying clientele and was thus offensive to connoisseurs of democracy. Once the common school was solidly entrenched, the scant opportunity afforded the lower classes for more than a rudimentary education fell under increasing challenge. If it was right to order children to learn reading, writing, and arithmetic and to offer them free tax-supported schooling, some reasoned, then it was also right to accommodate those desiring advanced instruction. Needless to say, the idea ran into opposition; yet it persisted. Before long, indeed, a few common schools, yielding to parental insistence, introduced courses beyond the elementary level. Such was the germ of the high school in the U.S.

The United States's first high school opened in Boston in 1821 as the English Classical School, a designation that soon was changed to English High School. Designed for the sons of the "mercantile and mechanic classes," it provided three years of free instruction in English, mathematics, surveying, navigation, geography, history, logic, ethics, and civics. Shortly after, high schools appeared in Portland, Maine, and Worcester, Massachusetts. In 1825 New York City inaugurated the first high school outside New England. The next year Boston braved free secondary education for girls, judiciously diluted and restricted to 130. When the number of applicants vastly exceeded this figure, the city fathers abandoned the project.

The high school movement was spurred less by these diffuse developments than by legislation by Massachusetts in 1827 that ordered towns of 500 families to furnish public instruction in American history, algebra, geometry, and bookkeeping, in addition to the common primary subjects. Furthermore, towns of 4,000 were to offer courses in history, logic, rhetoric, Latin, and Greek. Unhappily, the measure lacked public backing, and a generation passed before its boons became an actuality. Nevertheless, it set the guideposts for similar legislation elsewhere. Although high schools steadily increased, the contention that government had no right to finance them remained an issue until the 1870s, when Michigan's supreme court, finding for the city of Kalamazoo in litigation brought by a taxpayer, declared the high school to be a necessary part of the state's system of public instruction.

Education for females. Though the common school vouchsafed instruction to girls, their chances to attend high school—not to say college—were slight. For daughters of the wealthy there were "female academies," but they were not numerous, and often they were finishing schools rather than cultivators of the intellect. The truth is that as late as the 1840s, when the lowliest man could vote and hold office, women were haltered by taboos of every sort. But as America advanced industrially, and more and more women flocked to the mill and the office, their yearning for a fuller life ignited, and with it their desire for greater educational opportunity. As in the struggle for the common school, the cause of women's education bred leaders, many of whom founded schools in the Northeast. Not till 1861, however, when Vassar was founded, could women point to a truly first-rate college. Meanwhile, in 1833, Oberlin College in Ohio had hazarded coeducation, and 20 years later Antioch College, also in Ohio, followed suit. Beyond the Mississippi every state university, except that of Missouri, was coeducational from its beginning. The East moved more warily; Cornell University was the first Eastern school to become coeducational, in 1872.

Co-education

Higher education. While women were crusading for a more spacious educational opportunity, the college itself was undergoing alteration. It had begun as a cradle of divinity, but, as the 18th century waned, it was displaying a mounting secularity, with chairs in science, mathematics, law, and politics. In the course of the 19th century, not only did colleges surge in number, but some of the more enterprising of them undertook to reshape their purpose. Some (for instance, William and Mary), though still calling themselves colleges, engaged in university work. Soon after its opening in 1885, Bryn Mawr College in Pennsylvania announced courses for the master's and doctor's degrees. Inspired by the scholarly accomplishments of German universities, Johns Hopkins Uni-

versity in Baltimore, founded in 1867, put its weight on research. Twenty years later Clark University in Massachusetts opened as a purely graduate school. Two years later Catholic University of America in Washington, D.C., did the same. Soon the graduate trend invaded some of the older schools, such as Harvard University.

The early normal schools, or teacher-training schools, were primitive; often they were merely higher elementary schools, rehearsing their students for a year in basic reading and arithmetic, rectitude and piety, some history, mathematics, and physiology, and, if they survived, a rudimentary pedagogy. It was not until the 1860s, under the spell of Pestalozzi and later on, Froebel and Herbart, that teacher training in America began to gather power and momentum. (A.E.M.)

The British dominions. *Canada.* Canadian educational developments in the 19th century can be divided into early, middle, and late periods—in the east running to about 1840, 1870, and 1900, respectively. In the west, parallel developments began later, so that the second and third periods telescope into the early decades of the 20th century. These developments culminated in educational systems similar to those of the present.

In the early period of the 19th century, schooling was much the same as it was in England; it was provided through the efforts of religious and philanthropic organizations and dominated by the Church of England. Although there was overlap among types of schools (identified historically), there are records of parish schools, charity schools, Sunday schools, and monitorial schools for the common people. The instructional fare was a rudimentary combination of religious instruction and literacy skills, perhaps supplemented by some practical work.

More advanced education was limited to the upper social classes and was given in Latin grammar schools or in private schools with various curricular extensions on the classical base. Academies, largely supported by the middle class of nonconformist groups, presented a broad curriculum of liberal arts that spanned the secondary and higher levels of education. In general, instruction relied on a simple chain concept of "transmission–absorption–mental storage," which was kept going by direct application of reward or punishment.

In the middle period, public systems of education emerged, accommodating religious interests in a state framework. Public support was won for the common school, leading toward universal elementary education. Secondary and higher education began to assume a public character. The principle of local responsibility under central provincial authority was elaborated in the respective provinces.

Church and state relations in Canada

Of central importance in the development of Canadian education is the kind of agreement reached on church–state relations in education during this period. At one extreme is the arrangement made in Newfoundland from 1836 to accommodate all numerically represented denominations separately within a loose system (not until 1920 was a unified system of education developed, which still works through five denominational subsystems); at the other extreme are the arrangements made in British Columbia, which became decisive when it entered the Canadian Confederation, to establish and maintain a free, unified, centralized nonsectarian system. Other provinces eventually developed patterns that represented compromises. The Nova Scotia–New Brunswick pattern, for instance, provided a unified system that in principle was nonsectarian but that allowed the grouping of Roman Catholic children for education, thus legalizing sectarian schools within the system. Ontario placed separate Catholic schools within a unified school system. And Québec provided separate administrative machinery for Protestant and Roman Catholic schools and a complete dual system of schools, from common school to university.

The British North America Act of 1867, Canada's constitution, lodged authority for education in the provinces, at the same time guaranteeing denominational rights (in the "minority-school protective clause") if such rights existed at the time of entry into confederation. These two provisions established the pluralistic nature of Canadian education, and the union of the provinces and the entrance of western provinces gave Canada, by 1880, a national base on which to build the Canadian institution of education.

The final years of the 19th century were years of structural formalization of the educational foundations developed in the productive middle period. In this, Ontario's leadership was evident, especially as it affected the model of education evolving in the western territories. Although the North West Territories Act of 1875 established a principle of duality in the west, the Territorial Ordinance of 1892 changed the pattern to that of Ontario, a unified system with minority rights in education. After Alberta and Saskatchewan were admitted as provinces in 1905, some divergence from Ontario took place: notably, both provinces required that Roman Catholic taxes go to separate Catholic schools (the decision in Ontario was based on free choice), and Alberta allowed separate school privileges through the secondary level.

Toward the end of the 19th century, elementary schooling, by then established, was becoming compulsory. The cost of secondary education was diminishing, and the distinction in level and curriculum between the secondary and the elementary school was sharpened in order to define the structural parts and connections in a system of public schools. Communities were responsible for maintaining schools through a combination of local taxes and provincial grants, while provincial departments standardized the conduct of schooling through inspections, examinations, and prescription of course content and materials.

Changes in instructional theory, taking place during the latter part of the 19th century throughout the Western world, revolutionized the classroom. One major shift was from the imposition of knowledge on the mind of the learner to an emphasis on the learner's activity of perception and comprehension of knowledge. The impact of science on the higher school curriculum was matched by its impact on educational theory and, consequently, on teacher training. Both scientific disciplines (such as educational psychology) and scientific methods of teaching became necessary to the training of teachers who were to operate in a new setting of teacher–pupil and subject-matter relations.

Australia. The development of Australian education through the 19th century was affected by a pervasive British influence, by a continuous economic struggle against harsh environmental conditions, and by the tendency for population to be concentrated in centres that accrued and extended political authority over the region. The particular historical thread around which educational developments took place was the question of denominational schools.

Church and state relations in Australia

From the first immigrant landing in 1788 through the early decades of the 19th century, education was provided on an occasional and rather haphazard basis, by the most expedient means available. In general, the assumption and the practice was that schooling would be provided by the church or by church organizations, such as the Society for the Propagation of the Gospel in Foreign Parts, and colonial governments made small grants to aid such provision. It was also assumed that the Church of England would dominate the religious-educational scene, and a Church and School Corporation was set up in 1826 to administer endowments for Church of England efforts. Even at this early stage, however, the resistance of Nonconformists, especially Presbyterians and Roman Catholics, shortly defeated the attempt to "establish" Church of England institutions. The only early organized attempt at mass education was through monitorial systems.

In 1833 the governor of New South Wales asserted government responsibility for education by proposing the introduction of a nondenominational system that would reduce religion in schools to reading commonly approved scriptures and to providing release time for sectarian in-

struction by clergymen. The importance of the proposal lay in its spirit of religious compromise and its initiation of state responsibility for education, both of which were predictive of future development.

Because of sectarian resistance, mainly from Anglican and Catholic groups, so-called national schools were introduced alongside denominational schools in 1848 as a dual system, administered by two corresponding boards. Through the middle period of the century, similar sectarian compromises were found in other Australian colonies—such as Van Diemen's Land (Tasmania), Western Australia, and South Australia—leading to the establishment of state systems. Educational efforts were, however, seriously impeded by the extremity of the struggle for survival in hostile geographic conditions.

The trend toward sectarian and church–state compromise continued with a series of education acts passed between 1850 and 1871. In New South Wales there was a vigorous campaign for a unified state system of schools. With the support of one Anglican wing and of the other Protestant groups, a Public Schools Bill was passed in 1866, creating a single Council of Education. State aid to denominational schools was continued, but now under conditions stipulated by the state.

Victoria became a separate colony in 1850 and was initially fraught with particular problems occasioned by the arrival of a migrant gold-rush population. Little was accomplished in education, other than increased assistance to religious denominations, until 1856. After that the move for a state system gained impetus, and a Common Schools Bill was passed in 1862, establishing a system similar to that accepted in New South Wales. Soon after separation, Queensland's Primary Education Bill was passed in 1860, subordinating denominational schools and reinforcing the principle of common-school development in Australia. South Australia held to a continuous development of a general system based on common Christianity, but Western Australia's Elementary Education Bill of 1871 returned to dual support for both government and voluntary schools.

The support for state educational systems increased during the 1860s and 1870s as an alternative to interdenominational conflict was sought. In this development the Protestants, gradually and sometimes reluctantly, acquiesced. Catholic resistance was never overcome, and the consequent evolution of a separate Roman Catholic school system did not diminish Catholic dissatisfaction with the movement to state schools. The dilemma of Catholic citizens with regard to nonsectarian public education was universal: as citizens they were financially obligated for the public schools; as Roman Catholics they were committed to education in schools of their own faith.

Education in the later 19th century also took other directions. The intention to educate all children and to raise the quality of instruction in common schools required governmental actions that could transform voluntary, exclusive, uneven provisions into uniform public standards. In Australia, particular motivating factors were the dramatic increases in population and economic growth and the recognized inadequacy of existing schools. Although the full move to secular, free, compulsory public education was not completed until the 20th century, the establishment of secular public school systems under central government control was made unequivocal through the passage of legislation between 1872 and 1895 (Victoria, 1872; Queensland, 1875; South Australia, 1875; New South Wales, 1880; Tasmania, 1885; Western Australia, 1893 and 1895). These bills did not abolish general Christian instruction, nor did they generally refuse release time for sectarian instruction. They did disallow sectarian claims for financial support and for a place in public education. The decision was for the operation of schools for all children, undertaken by the one agency that could act on behalf of the whole society, the central government.

New Zealand. In New Zealand's early colonial period, between 1840 and 1852, certain provisions were made for endowments for religious and educational purposes, but education was considered, in accordance with prevailing views in England, a private or voluntary matter. Corresponding to general social distinctions, academic education was relegated to denominational, fee-charging schools, and common education was provided as a charitable service. Religious preference was avoided as much as possible, with the aim of minimizing sectarian conflict.

Church and state relations in New Zealand

Secular opposition to religious bias, even on a pluralistic basis, was, however, already evident. In 1852 New Zealand was granted self-government under the Constitution Act, and responsibility for education was placed in the councils of the six provinces. Although each province acted independently and somewhat according to the traditions of the dominant cultural group, the general sentiment moved in the next 20 years toward the establishment of public school systems.

The province of Nelson, from a highly mixed and Nonconformist popular base, led the way in establishing nondenominational schools, administered by a board, supported through tax levies, and reaching relatively high attendance proportions. Otago, under Scottish Presbyterian influence, set up well-organized administrative procedures and a complete array of school types, though resisting the equalizing tendency of free schools. Canterbury, an Anglican preserve, held to denominational grants initially, but by 1863 this support was channelled through a public board of education, and thereafter the development of a primary system proceeded rapidly, becoming secular and free by 1873. Wellington's Common School Act of 1855 set a legal basis for public schools. There was little alteration to private initiative and school fees until 1871, however, when the Education Act established a secular basis and started the province toward free, compulsory schooling. In Auckland the denominational principle was retained until 1869; between then and 1872, provisions of the Nelson system were adopted, and primary education was made secular, free, and compulsory. By 1876, then, when the provincial governments were abolished, the people of New Zealand, through varying regional decisions, had accepted governmental responsibility for education, had opted for nonsectarian schools, and had started on the path to free, compulsory common schooling.

The basic national legislation was passed in 1877. The Education Act provided for public elementary education that would be secular, free to age 15, and compulsory to age 13. Because of enforcement difficulties and legal exceptions, the compulsory clause was rather loose, but it instituted the rule. It was strengthened between 1885 and 1898, and bold steps in 1901 and 1910 led to steadily increasing high school enrollments after 1911. The act of 1877 also revised the administrative structure under a central (national) ministerial Department of Education. Initially, the central department was little more than a funding source, while critical control was vested in regional boards elected by local school committees. The power of the new boards was enhanced by their coincidence to the old provincial boards, but the central principle had been accepted, and increasing recognition of regional inequalities led to augmentation of the department's role before the end of the century. In the competitive struggle between the department and the regional boards that waxed and waned well into the 20th century, neither gained the exclusive dominance sometimes sought. The primary position of the central authority in educational administration was confirmed in the reform period between 1899 and 1914, however, when control of inspectors, effective control of primary teacher appointment and promotion, and stipulative control in fund granting went to the Department of Education. These developments, together with curriculum and examination reforms, marked a new beginning in New Zealand education. (R.F.L.)

SPREAD OF WESTERN EDUCATIONAL PRACTICES TO ASIAN COUNTRIES

India. Originally the British came to India as tradesmen, but gradually they became the rulers of the coun-

try. On December 31, 1600, the East India Company was established, and, like all commercial bodies, its main objective was trade. Gradually during the 18th century the pendulum swung from commerce to administration; the deterioration of Mughal power in India, the final expulsion of French rivals in the Seven Years' War, and the virtual appropriation of Bengal and Bihār in a treaty of 1765 had all made the Company a ruling power. In spite of this, the Company did not recognize the promotion of education among the natives of India as a part of its duty or obligation. For a long time the British at home were greatly opposed to any system of public instruction for the Indians, as they were for their own people.

Education under the East India Company

The feelings of the public authorities in England were first tested in the year 1793, when William Wilberforce, the famous British philanthropist, proposed to add two clauses to the company's charter act of that year for sending out schoolmasters to India. This encountered the greatest opposition in the council of directors, and it was found necessary to withdraw the clauses. For 20 years thereafter, the ruling authorities in England refused to accept responsibility for the education of Indian people. It was only in 1813, when the company's charter was renewed, that a clause was inserted requiring the governor general to devote not less than 100,000 rupees annually to the education of Indians.

Some organization was required in order to disburse the educational grant. A General Committee of Public Instruction, constituted in Calcutta in 1823, started its work with an Orientalist policy, rather than a Western-oriented one, since the majority of the members were Orientalists. The money available was spent mainly on the teaching of Sanskrit and Arabic and on the translation of English works into these languages. Some encouragement was also given to the production of books in English.

Missionary education and the "semirationalist" movement

Meanwhile, a new impetus was given to education from two sources of different character. One was from the Christian missionaries and the other from a "semirationalist" movement. The Christian missionaries had started their educational activities as early as 1542, upon the arrival of St. Francis Xavier. Afterward the movement spread throughout the land and exercised a lasting influence on Indian education. It gave a new direction to elementary education through the introduction of instruction at regular and fixed hours, a broad curriculum, and a clear-cut class system. By printing books in different vernaculars, the missionaries gave a great fillip to the development of Indian languages. But hand in hand with the study of the vernaculars went the teaching of Western subjects through the medium of English, called in India "English education."

Besides the missionaries, there were men in Bengal who, though admitting the value of Oriental learning for the advancement of civilization, thought that better things could be achieved through a knowledge of English. In 1817 these semirationalists, led by Rammohan Roy, the celebrated Indian reformer, founded the Hindu College in Calcutta, the alumni of which established large numbers of English schools all over Bengal. The demand for English education in Bengal thus preceded by 20 years any government action in that direction.

In the meantime the influence of the Orientalists was waning in the General Committee, as younger members with more radical views joined it. They challenged the policy of patronizing Oriental learning and advocated the need for spreading Western knowledge through the medium of English. Thus arose the controversy as to whether educational grants should be used to promote Oriental learning or Western knowledge. The controversy between the Orientalists and the Anglicists was decided in favour of the latter by the famous Minute on Education of 1835 submitted by Thomas Babington Macaulay, the then legal member of the governor general's executive council. His recommendations were accepted by Lord William Bentinck, the governor general. The decision was announced on March 7, 1835, in a brief resolution that determined the character of Indian education for the ensuing century. Although the schools for Oriental learning were maintained for some years, the translation of English books into Sanskrit and Arabic was immediately discontinued. Thus the system of "English education" was adopted by the government. It should be noted, though, that primary education did not attract any attention at all.

Orientalists versus Anglicists

Bentinck's resolution was followed by some more enactments accelerating the growth of English education in India. The first was the Freedom of Press Act (1835), which encouraged the printing and publication of books and made English books available to the people at low cost. Two years later, Persian was abolished as the language of record and the courts (to the dismay of the Muslims) and was replaced by English and Indian languages in higher and lower courts, respectively. Finally, Lord Hardinge, as governor general, issued a resolution on October 10, 1844, declaring that for all government appointments preference would be given to the knowledge of English. These measures strengthened the position of English in India, and the lingering prejudices against learning English vanished forever.

It may be noted, however, that, although English education held its ground in Bengal, it was found that in Bombay, Madras, and the North-Western Provinces there was as yet little effective demand for English, and the tendency was to lay the main stress on Indian languages. Bombay adopted the policy of encouraging primary education and spreading Western science and knowledge through the mother tongue. This was done under the able guidance of Mountstuart Elphinstone, then the governor, even though the government also conducted an English school in almost every district in the province. Between 1845 and 1848 a bitter controversy arose regarding the medium of instruction, but the issue was between the mother tongue and English, and not between a classical language and English as it was in Bengal. The controversy gathered strength every day; and, in those days of centralization, the matter had to be referred to the Bengal government, which advised the Bombay government to concentrate its attention on English education alone, thus throttling the growth of education through the mother tongue in Bombay. Meanwhile, the Madras government was biding its time, leaving the field of positive effort open to Christian missionaries; as a result of this missionary initiative, English education in the Madras presidency was more extensively imparted than in Bombay. The Bengal government, however, did not neglect vernacular education altogether; in 1844 it established 101 vernacular schools for the main centres of population throughout the province.

A laudable experiment in the field of vernacular education was tried by a famous lieutenant governor, James Thomason, in the North-Western Provinces. He obtained government permission to set up one well-equipped and well-staffed school in every *tahsil*, or local division, as a model for other schools in the neighbourhood, but, although the scheme was based on a good principle, it did not make much headway. He tried another scheme to bring primary education within easy reach of the common people—a scheme known as the *ḥalqabandī* system. In each *ḥalqah* (circuit) of villages, a school was established in the most central village so that all the villagers within a radius of two miles might avail themselves of it. For the maintenance of these schools the village landholders agreed to contribute at the rate of 1 percent of their land income. The experiment proved successful, and in ten years Thomason opened 897 schools and provided elementary education for 23,688 children.

Ḥalqabandī system

The next step in the history of Indian education is marked by Sir Charles Wood's epoch-making Dispatch of 1854, which led to (1) the creation of a separate department for the administration of education in each province, (2) the founding of the universities of Calcutta, Bombay, and Madras in 1857, and (3) the introduction of a system of grants-in-aid. Even when the administration of India passed from the East India Company into the hands of the British crown in 1858, Britain's secretary

of state for India confirmed the educational policy of Wood's Dispatch.

It may be noted that the newly established universities did not initially undertake any teaching responsibilities but were merely examining bodies. Their expenses were confined to administration and could be met from the fees paid by the candidates for their degrees and certificates. Although the establishment of the universities did result in a rapid expansion of college education and although the products of the new learning displayed keen scholarship, the value of learning nevertheless soon decayed. In such circumstances it was ironical for the Indian Education Commission of 1882 to declare, "The university degree has become an accepted object of ambition, a passport to distinction in public services and in the learned professions." Another undesirable practice was the domination of the universities over secondary education through their entrance examinations. University policies regarding curricula, examination systems, mediums of instruction, and other vital problems began to be chalked out by university teachers who had very little experience in schoolteaching and who kept the administrative needs and requirements of colleges in the forefront. Thus, secondary schools began to aim at college education and did not prepare their students for life.

Excessive emphasis on higher education

The new system also became top-heavy. It must be stated that the Commission of 1882 made a very valuable recommendation that the "elementary education of the masses, its provision, extension and improvement requires strenuous efforts of the state in a still larger measure than heretofore." It also desired to check the wild race for academic distinction and "to divert some part of the rapidly swelling stream of students into channels of a more practical character." Despite this warning, however, alternative courses in commerce, agriculture, and technical subjects that were offered in a number of selected schools did not prove popular. The educated classes could not be diverted from their conventional path. The press and the public did demand more and wider facilities for technical education at the higher level, but the government proved indifferent.

In a general view of education during the last two decades of the 19th century, drift was more apparent than government resolve. Elementary education was starved and undernourished, and secondary education was suffering from want of proper supervision. There was an unplanned growth of high schools and colleges since the Education Commission had given a free charter to private enterprise. Many of these private institutions were "coaching institutions rather than places of learning." The universities had no control over them, and state control was negligible because the government had adopted a laissez-faire policy.

The second half of the 19th century is, nonetheless, of great significance to the country because modern India may indeed be said to be a creation of this period. It brought about a renaissance by breaking down geographical barriers and bringing different regions and long-separated Indian communities into close contact with one another. The blind admiration for Western culture was gradually passing away, and a new vision and reorientation in thought were coming about. A feeling of dissatisfaction also developed toward the existing governmental and missionary institutions. It was felt by some of the Indian patriots that the character of Indian youths could be built by Indians themselves. This led to the establishment of a few notable institutions aiming at imparting sound education to Indian youth on national lines—institutions such as the Anglo-Mohammedan Oriental College in Alīgarh (1875), the D.A.V. College in Lahore (1886), and the Central Hindu College in Banares (1898). The politically minded classes of the country had also come to regard education as a national need. They were critical of the government's educational policy and resented any innovation that might restrain the pace of educational advance or diminish liberty. (S.N.M.)

Japan. *The Meiji Restoration and the assimilation of Western civilization.* In the early 19th century, Japan still lived under a feudal system headed by the Tokugawa (Edo) shogunate, a dynasty of military rulers who held the real power while the emperors remained mere figureheads. It was not until 1868 that the shogunate was overthrown, the feudal system abolished, and the Imperial authority of the Meiji dynasty restored, leading to drastic reforms of the social system. This process has been called the Meiji Restoration.

The Tokugawa heritage

In the following generation Japan quickly adopted Western industry and culture and developed rapidly as a modern state. But the rapid modernization of Japan could not be accounted for without the long-enduring peace and cultural achievements of the Tokugawa era. It had boasted a high level of Oriental civilization, especially centring on Confucianism, Shintōism, and Buddhism. The ruling samurai, or warriors, had studied literature and Confucianism at their *hankō* (clan schools); the commoners had learned reading, writing, and arithmetic at numerous *terakoya* (temple schools). This cultural heritage helped equip Japan with a formidable potential for rapid westernization. Indeed, some elements of Western civilization had been gradually introduced into Japan even during the Tokugawa era. The shogunate, notwithstanding its isolationist policy, permitted trade with the Dutch, who conveyed modern Western sciences and arts to Japan. Only after 1853, however, did Japan open its door equally to Western countries, as a result of pressures exerted by the United States Navy under Adm. Matthew C. Perry. Thenceforth Japan was forced to cope with all kinds of Western information and communications, spoken and written in English, German, French, and other languages.

Western studies (yōgaku), especially English-language studies, became increasingly popular after the Meiji Restoration; Western culture flooded into Japan. The Meiji government dispatched study commissions and students to Europe and to the United States, and the so-called westernizers defeated the conservatives who tried in vain to maintain allegiance to traditional learning.

Establishment of a national system of education. In September 1871 Japan's first true Ministry of Education was established. It assigned commissioners, many of whom were *yōgakusha* (students of Western learning), to design the school system, and in September 1872 the *gakusei*, or government order of education, was promulgated. It decreed that all youth, regardless of class or sex, should receive education. The country was initially divided into eight university districts, which were further divided into 32 middle school districts and 210 primary school districts. Each respective district would have one university (*daigaku*), one middle school (*chūgakkō*), and one primary school (*shōgakkō*). Whereas Tokugawa education had separated warriors and commoners, *gakusei* envisioned a unified, classless system of modern national education, designed on a ladder plan. Although the district system was said to have been borrowed from France, the new Japanese education was not merely imitative of a single model; it was based on the study of Western education and assimilated the essences of education in all advanced countries. Curricula and methods of education, for instance, were drawn primarily from the United States.

Gakusei, or government order of education

At the time, however, Japan was not sufficiently prepared to carry out completely the plans of *gakusei*. Districts were marked, and primary schools were opened, but the real conditions were far short of the ideal. No university was erected at all. Soon after the promulgation of *gakusei*, the government, adapting itself to the need for speedy westernization, established professional schools and foreign-language schools, in which foreigners taught in foreign tongues. Thus, the school system of Japan split into dual lines: one consisted of the primary schools for commoners, and the other consisted of the professional schools for leading classes.

Under the *gakusei* system of education, the Ministry of Education, together with local officials, managed with difficulty to set up primary schools and to make children attend them. The number of primary schools amounted to 24,000 in 1875, with 45,000 teachers and 1,928,000 pupils. The ratio of school attendance reached 35 per-

cent of all who could attend (ages six to 14); even those children who commenced schooling tended to drop out after a year or so. In 1873 an educator from the United States, David Murray, was invited to Japan as an adviser to the Ministry of Education; another, M.M. Scott, assumed direction of teachers' training and introduced modern methods and curricula at the *shihangakkō* (normal school), established under the direct control of the ministry. The *shihangakkō* designed curricula for the primary schools, modelled after those of the United States, and compiled primary textbooks involving new teaching methods. These methods and curricula, radically different from that of *terakoya*, or temple schools, of the Tokugawa era, spread gradually into the elementary schools of almost all regions.

Work of educators and teachers from abroad

Many teachers from abroad were engaged at large salaries in professional schools and foreign-language schools. Foreign missionaries built their own institutions. Private foreign-language schools (mainly English) also helped introduce Western culture. In 1877 the University of Tokyo—Japan's first university in a modern sense—was founded, and it too employed many European and American instructors.

The conservative reaction. Japan's education experienced a turning point with the Satsuma Rebellion in 1877. Following this last and greatest of the samurai uprisings, the nation again forged ahead toward political unity, but there was an increasing trend of antigovernment protest from below, which has usually been called the Free People's Rights Movement. Because of the Satsuma rebellion, the government was in heavy financial difficulties. Also, with the people's inclination toward Western ideas fading away, a conservative reaction began to emerge, emphasizing the forgotten traditions closely associated with Confucian and Shintō ideology and favouring a return to local control of education.

There had been mounting discontent among the rural people against the old law of 1872, mainly because it had imposed upon them the financial burdens of establishing schools and yet had not lived up to expectations. Tanaka Fujimaro, an executive of the Ministry of Education, just returning from an inspection tour in the United States, thought that the government should transfer its authority over education to the local governments, as in the United States. In September 1879 the government nullified the former *gakusei* and put into force the *kyoiku-rei* (Education Law), which made for rather less centralization. Not only did the new law abolish the district system that had divided the country into districts for *daigaku*s, *chūgakkō*s, and *shōgakkō*s, but also it reduced central control over school administration, including the power to establish schools and regulate the duration of attendance. The law was intended to encourage the initiatives of towns and villages. Such a drastic educational reform, however, led to a decline in the efficiency of education in some regions, and criticism arose among those prefectural governors who had been striving to enforce *gakusei* in their regions.

As a countermeasure, the government then again changed the law in 1880, revising the rights of prefectures. Thereafter, the prefectural government would provide regulations within the limits of criteria set by the Ministry of Education; some measure of educational unity was thus reached on the prefectural level, and the school system received some needed adjustment. Yet, because of economic stagnation, the attendance ratio remained low.

Re-emphasis on Confucian and Shintōist values

Reacting to the increasing conservative trend, the Ministry of Education reintroduced Confucian and Shintōist values. In the primary schools, *shūshin* (ethics) was made the most important core of the curricula, and the ministry compiled a textbook with overtones of Confucian morality.

Establishment of nationalistic education systems. With the installation of the cabinet system in 1885, the government made further efforts to pave the way for a modern state. The promulgation of the Meiji constitution, the constitution of the Empire of Japan, in 1889 established a balance of Imperial power and parliamentary forms. The new minister of education, Mori Arinori, acted as a central figure in enforcing a nationalistic educational policy and worked out a vast revision of the school system, and this set a foundation for the nationalistic educational system that developed during the following period in Japan. Japanese education thereafter, in the Prussian manner, tended to be autocratic. Later, in 1903, the Ministry of Education was even confirmed in its power to alter, add, or delete information in national textbooks to accord with political currents.

Based on policies advocated by the educational minister, a series of new acts and orders were published one after another. The first was the Imperial University Order of 1886, which changed the name of the University of Tokyo to Tokyo Imperial University (*Tokyo Teikoku Daigaku*), thus, as the name indicates, rendering the university a servant of the state, to train high officials and elites in various fields. The ministry also carried out sweeping revisions of the normal school, which adopted militaristic education and trained teachers to obey strictly the order of their seniors.

Together with these reforms, the Imperial Rescript on Education (*kyoiku-chokugo*) of 1890 played a major role in affirming the nationalistic system. By re-emphasizing the traditional Confucian and Shintō values and redefining the courses in ethics (*shūshin*), it was to place morality and education on a foundation of Imperial authority. It would provide the guiding principle for Japan's education until the end of World War II.

Promotion of industrial education. Ever since the Meiji Restoration in 1868, the national target had been *fukoku-kyōhei* ("wealth accumulation and military strength") and industrialization. From the outset the Meiji government had been busy introducing science and technology from Europe and America but nevertheless had difficulties in realizing such goals.

Inoue Kowashi, who became minister of education in 1893, was convinced that modern industries would be the most vital element in the future development of Japan and thus gave priority to industrial and vocational education. In 1894 the Subsidy Act for Technical Education was published, followed by the Technical Teachers' Training Regulations and the Apprentice School Regulations. The system of industrial education was in general consolidated and integrated. These measures contributed to the training of many of the human resources required for the subsequent development of modern industry in Japan. (A.Na.)

IX. Education in the 20th century

SOCIAL AND HISTORICAL BACKGROUND

International wars, together with an intensification of internal stresses and conflicts among social, racial, and ideological groups, are characteristic of the 20th century and have had profound effects on education. Rapidly spreading prosperity and affluence, immense increases in population, an explosion of knowledge, the establishment of many new professions, the growth of large-scale science-based industry, the enormous influence of mass media such as television, the increasing power of organized labour, a trend toward Socialism, and widespread skepticism about all accepted values, including those supported by religion, are other factors that have also had far-reaching effects.

The two world wars weakened the military and political might of the larger European powers and thus helped to liquidate colonialism. In Africa and Asia, almost every nation became independent, one consequence of which has been a great increase in the quantity of education provided. Attempts, not always successful, have been made to eradicate illiteracy. Colleges and schools have been built everywhere, and the growth in the number of pupils and students has been astonishing.

The establishment, after revolution and civil strife, of Communist and Socialist governments in many parts of the world has resulted in the promotion and expansion of education by governments that have seen education as a key to industrial development. The growing affluence of masses of the population in high-income areas in North

America and Europe has brought about, particularly since World War II, a tremendous demand for secondary and higher education. Most children stay at school till 16, 17, or even 18 years of age, and a substantial fraction spend two, three, or four years at college. The number of universities in many countries has doubled or trebled, and the number of students has increased four times between 1950 and 1970. This explosive growth is sustained by the fact that many more workers need to study in lecture room and laboratory after their school years are completed, if they are to use skillfully the instruments, materials, and methods put at their disposal by modern scientific technology.

Demands of industry for technical and continuing education

Employers and managers, as well as governments, realize very clearly that modern industry has a built-in obsolescence. New methods, processes, and machines are continually introduced. Old skills become irrelevant; new industries spring up. In addition, the amount of scientific, as distinct from merely technical, knowledge grows continually. More and more research workers, design specialists, and high-level professionals are called for. The reaction has been to encourage the establishment of technical and technological colleges; to promote adult education at all levels; to turn attention to part-time and evening courses; and to provide more training and education within the industrial enterprises themselves.

The adoption of modern methods of food production has diminished the need for agricultural workers, who have headed for the cities. Urbanization, however, brings problems: city centres decay, and there is a trend toward violence. The poorest remain in these centres, and it becomes difficult to provide adequate education.

SUMMARY OF MAJOR TRENDS AND PROBLEMS

In general, the chief trends in education since 1900 may be summarized thus: (1) There has everywhere been a very great increase in the number of pupils and students, at all levels. Secondary education is being universalized. (2) The tertiary level—higher education—has become more important. Universities are changing from selective to mass institutions. (3) The relation between education and material well-being has become clearer. In consequence, increasing attention is paid to technical and technological training. (4) The development of new science-based industries, the decay of old industries, and the very rapid obsolescence of modes of production have convinced administrators, managers, and workers that facilities for retraining workers and for in-service education must be provided. This draws attention to the importance of adult education and of permanent, lifelong, continuing education. A major difficulty here is that no one knows what the future for which people are being prepared will be like. All that is certain is that it will be very different from the present. Hence, there is a stress on adaptability, on the readiness to accept and welcome change and novelty. (5) There is a general acceptance of the notion that equality of educational opportunity is an ideal that must be accepted. In consequence there is everywhere a growing acceptance of the "common" or "community" or "comprehensive" school and college. Elitism and selectivity are being abandoned, even rejected, almost everywhere.

Problems of all kinds arise as a consequence of these trends. Among them the following are important: (1) Education is expensive, requiring vast resources in materials and manpower. How can these be best mobilized? (2) Mass systems, replacing selective systems, call for new modes of organization and administration. How should the secondary and tertiary levels be structured? What kinds of schools and colleges are needed? (3) What should be taught at each level of education and how? How should the problem of curriculum and method be approached? What are the relative claims of general and vocational education? (4) What are the best ways of selecting pupils for the various levels? How should students be allocated at the third level among the different professional paths available? (5) Is it possible to use modern methods of planning to make education more productive and less costly? (6) Can new inventions such as television, colour photography, tape recorders, and computers be used to teach students more cheaply and more efficiently? (7) How can systems of permanent, lifelong education be established and how should they be organized?

The above are no more than examples of the kind of problems being studied in the last third of the 20th century. The way in which the discussions proceed is influenced, as should be expected, by the intellectual and cultural climate of the times. There is widespread skepticism about the real value of intuitive and contemplative philosophy. Science has enormous prestige. Mathematics and statistics are revered, as are all quantitative and measurable considerations. The result is that growing importance is accorded to the findings of the social sciences, such as economics and sociology, and less to the arguments and proposals of philosophers, historians, and humanists in general. Positivism, pragmatism, and social, cultural, and moral relativism are dominant, even triumphant.

MAJOR INTELLECTUAL MOVEMENTS

Influence of psychology on education. The attempt to apply scientific method to the study of education dates back to the German philosopher Johann Friedrich Herbart, who called for the application of psychology to the art of teaching. But not until the end of the 19th century, when the German psychologist Wilhelm Max Wundt established the first psychological laboratory at the University of Leipzig in 1879, were serious efforts made to separate psychology from philosophy. From Wundt's monumental *Grundzüge der physiologischen Psychologie* ("Fundamentals of Physiological Psychology") flowed the various movements that had significant effects on education in the 20th century.

Influence of William James

Among Wundt's pupils was the American psychologist and philosopher William James, often considered the father of American psychology of education. In 1878 he established the first course in psychology in the United States and in 1890 published the famous *Principles of Psychology*, in which he argued that man should be thought of as a living organism with instinctive tendencies to react with his environment. A child's mind, therefore, is that aspect of his being that enables him to adapt to the world, and the purpose of education is to organize the child's powers of conduct so as to fit him to his social and physical environment. Interests must be awakened and broadened as the natural starting points of instruction. James's *Principles* and *Talks to Teachers on Psychology* cast aside the older notions of psychology in favour of an essentially behaviourist outlook; they asked the teacher to help educate heroic individuals who would project daring visions of the future and work courageously to realize them.

Both William James and the American philosopher John Dewey are often thought of as defending "functionalism," a body of psychological theory that has had great influence on education during the 20th century. The functionalists viewed man as a living organism, a product of both heredity and environment, and were interested not merely in what the phenomena of consciousness are but also in what they do. They helped to clear away many of the fantastic notions once held about the successive steps involved in the development of mental functions from birth to maturity. Their methodology yielded valuable clues to the understanding of the psychology of learning, and they were the first to call attention to the part played by motivation and goal seeking.

Influence of Freudianism

Freudianism was one of the intellectual mainstreams that influenced educational theory in the 1920s. Freudians began to apply Freudian doctrines to education soon after World War I, but they made a direct impact upon only a limited circle of educators. The powerful and pervasive influence of Freudianism came indirectly through gradual public acceptance of the psychoanalytical image of the child. It accelerated changes in the emotional climate and in the authority pattern of the classroom.

The new psychology that developed during the early decades of the 20th century represented a scientific and

experimental approach to the biological growth of the human animal and to the learning process. Each school of psychology added more to the knowledge of human behaviour, and each had its effect on education. Schools and educators could not return to the concepts of education that had predominated at the turn of the century.

Traditional movements. *Essentialist education.* The name essentialist is derived from the obvious stress on "essentials," by which is meant those human experiences past and present that—so some believe—are indispensable to people living today or at any time. They hold that it is the main function of organized education to transmit these essentials to the young. The leading exponent of the essentialist doctrine was the American educationist William Chandler Bagley. Bagley published *Education and Emergent Man* (1934), a powerful refutation of pragmatism. He believed that it is possible to accept a thoroughgoing evolutionary theory that links mankind with the entire range of life and yet recognize in mankind a potential nobility, all this without bringing in a metaphysical element.

The essentialists saw in the progressive movement the abandonment of standards of scholastic achievement, the neglect of system and sequence, indiscriminate use of activity, overemphasis on social studies, irresponsibility in curriculum revision, and an attempt to use elementary schools to establish a new social order. Bagley spoke for many when he wrote,

> It is true that the world of today is a different world from the world of 1913 and from the world of 1929 . . . but this does not mean that everything has changed. . . . The winds that blow still follow the laws of storms; *Huckleberry Finn* and *Treasure Island* still delight youth; and the Sistine Madonna is just as beautiful as of yore. (From *Education and Emergent Man*, 1934.)

The movement culminated in the formation of The Essentialist Committee for the Advancement of American Education, which aimed at combatting falling standards of achievement in American schools. The essentialists demanded a return "to the exact and exciting studies." They favoured the "mental disciplines" and gave their blessing to Latin, algebra, and geometry. In the matter of method and content, Essentialists put effort above interest, subjects above activities, collective experience above that of the individual, logical organization above the psychological, and the teacher's initiative above that of the learner.

Humanistic, or liberal, education. The chief proponents of "humanistic," or "liberal," education were Robert M. Hutchins, formerly chancellor of the University of Chicago, and Mortimer J. Adler, professor of the philosophy of law at the same institution. Adler argued for the restoration of an Aristotelian viewpoint in education. Maintaining that there are unchanging verities, he sought a return to education fixed in content and aim. He claimed to prove by philosophical inquiry that

The work of Adler and Hutchins

> education is the process by which those powers (abilities, capacities) of men that are susceptible to habituation, are perfected by good habits, through means artistically contrived, and employed by any man to help another or himself achieve the end in view (*i.e.*, good habits). (From National Society for the Study of Education, *41st Yearbook*, Part 1, p. 246, 1942.)

Hutchins denounced American higher education for its vocationalism and "anti-intellectualism," as well as for its delight in minute and isolated specialization. Anti-intellectualism in American education, Hutchins argued, leads to sentimentalism. Suspicion of the intellect leads to the emphasis on knowing what one wants rather than why one wants it. Thus, "unlimited acquisition" becomes the human law, and "success" becomes the measurement of education.

Hutchins and his colleagues urged a return to the cultivation of the intellect. Everyone between the ages of 16 and 20, except the obviously unfit, should be given the same education. The first steps in the training of the intellect would be through grammar, rhetoric, and logic. Through these the learner can cultivate the arts of reading, writing, and speaking correctly. He would be required to study theoretical mathematics, which "exemplifies reasoning in its clearest and most precise form."

Religious and classical education. Opposed to the fundamental tenets of pragmatism is the philosophy that underlies all Catholic education. Theocentric in its viewpoint, Catholic scholasticism has God as its unchanging basis of action. It insists that without such a basis there can be no real aim to any type of living, and hence there can be no real purpose in any system of education. It teaches that man is a personal creation of God; that he is body (material) and soul (spiritual); and that this dual elemental composition functions as an integrated unity. The supernatural is the essence of the Catholic system. It pervades its theology, its ethics, its practice, its view of life, and its education. The church's

The aim of Catholic education

> whole educational aim is to restore the sons of Adam to their high position as children of God. [It insists that] education must prepare man for what he should do here below in order to attain the sublime end for which he was created . . . The true Christian does not renounce the activities of this life, he does not stunt his natural faculties; but he develops and perfects them, by coordinating them with the supernatural. He thus enables what is merely natural in life and secures for it new strength in the material and temporal order, no less than in the spiritual and eternal. (From Pius XI, encyclical on the "Christian Education of Youth," December 31, 1929.)

Such is the ultimate end of Catholic education, and to it every Catholic educational institution, from infant school to the graduate school, must address itself. Everything in education—content, method, discipline—must lead in the direction of man's supernatural destiny. Every agency playing a part in education—the school, the family, and the state—must heed this ultimate end.

> It is necessary not only that religious instruction be given to the young at certain fixed times, but also that every other subject taught be permeated with Christian piety. (From Leo XIII, encyclical "Militantes Ecclesiae," August 1, 1897.)

Experimental movements. *Child-centred education.* Proponents of the child-centred approach to education have typically argued that the school should be fitted to the needs of the child, and not the child to the school. One of the forerunners, called the father of progressive education by John Dewey, was the American educationist Francis W. Parker. In his view, what mattered in a school was not the learning of subject matter, neatly and logically arranged, but the child's own development. Parker became superintendent of schools at Quincy, Massachusetts, in 1875, where he introduced his ideas and practices. The old-fashioned, stiff, unnatural order was banned. An activity program was introduced, and subjects were interrelated around a central core instead of being taught for their own sake. In 1883 he moved to Cook County Normal School, Chicago, surrounding himself with excellent teachers. There he continued along the path he had started at Quincy. He assailed the mechanical, assembly-line methods of traditional schools. He stressed "quality teaching," by which he meant such things as activity, creative self-expression, excursions, understanding the individual, and the development of personality. Parker stressed the child's need for art, music, craft training, science, and physical training. Through the "Quincy method" and the Cook County Normal School, the first steps were taken toward more modern educational methods.

A different approach to child-centred education arose as a result of the study and care of the physically and mentally handicapped. Teachers had to invent their own methods and special apparatus to meet the needs of such children, because the ordinary schools did not and could not supply them. When these methods proved successful with handicapped children, the question arose whether they might not yield even better results with ordinary children. During the first decade of the 20th century, the educationists Maria Montessori of Rome and Ovide Decroly of Brussels both went on to apply their educational inventions in schools for ordinary boys and girls.

The Montessori method

The Montessori method's underlying assumption was the child's need to escape from the domination of parent and teacher. According to Maria Montessori, there is per-

manent conflict between adult and child. Children are the unhappy victims of adult suppression, who have been compelled to adopt defensive measures foreign to their real nature in the struggle to hold their own. The first move toward the reform of education, thought Maria Montessori, should be directed toward the adult educator: to enlighten his conscience, to remove his perception of superiority, and to make him humble and passive in his attitudes toward the young. The next move should be to provide a new environment in which the child has a chance to live a life of his own, free from frustrations. In the Montessori method, the special senses are separately trained by means of apparatus calculated to enlist spontaneous interest at the successive stages of mental growth. By similar self-educative devices, the child is led to individual mastery of the basic skills of everyday life, and then to schoolwork in arithmetic and grammar. There are no teachers but only kindly directors whose function it is to see that things go according to a master plan.

The Decroly method's essential feature was a program of work based on centres of interest and the individualism of learning by means of educative games. The basic feature was the workshop-classroom, at once a workshop, a study, a museum, a zoo, a library, in which children could go freely about their own occupations, planning, talking, looking on, cooperating as the occasion might require. Behind the complex of individual activities there was a carefully organized scheme of work based on an analysis of the fundamental needs of the child. The teachers were required to follow an orderly method in guiding the children. The principle of giving priority to wholes rather than to parts was emphasized in teaching children to read, write, and count, and care was taken to reach a comprehensive view of the experiences of life.

The Montessori method, promoted by the Association Montessori Internationale, and the Decroly method have gradually spread throughout the world, and their success has done much to stimulate the development of other aspects of modern education. Partly as a result of a renewed emphasis on individual learning, educators—often not aware of the source of their ideas—have begun once more to look for better methods of teaching classes that give more consideration to individual difference in the needs and abilities of pupils.

Pragmatist and progressive education. The progressive education movement was part and parcel of a broader social and political reform called the Progressive movement, which dates to the last decades of the 19th century and the early decades of the 20th. Elementary education had spread throughout the Western world, largely doing away with illiteracy and raising the level of social understanding. Yet, despite this progress, the schools had failed to keep pace with the tremendous social changes that had been going on. The increasing mechanization of life and work gradually reduced the need and opportunity for the exercise of personal skills and activities, and in the schools the children were expected to sit immobile, absorbing passively.

Experimental schools were established during the last decade of the 19th century and in the early 20th century by several educational reformers who wished to put their ideas into practice. The development of a new philosophy of education and of new methods was a natural outgrowth of such schools. They comprised three groups: the country boarding schools, the American experimental group, and experimental schools of the Montessori type. There were also other educational reformers who did not establish experimental schools but contributed by their writings.

Some early experimental schools

What is considered to be the first modern experimental school, Abbotsholme, Derbyshire, was founded by the English educator Cecil Reddie in 1889. Reddie's aim was to develop personality through freedom, his motto being "liberty is obedience through law." The aim of instruction was not only to impart knowledge but also to develop character and encourage reason and judgment. The methods of teaching varied, but a favourite one was group discussion. One of the innovations was a system of self-government by the pupils. John Haden Badley, an associate of Reddie at Abbotsholme, founded Bedales School, in Hampshire, in 1893. Essentially the same as Abbotsholme, there was one great difference: Bedales was coeducational and thus highly unconventional. The country boarding school was soon introduced in Germany, France, and Switzerland.

The principal experimental schools in America until 1914 were the Laboratory Schools of the University of Chicago, founded in 1896 and directed by John Dewey; the Francis W. Parker School, founded in 1901 in Chicago; the School of Organic Education at Fairhope, Alabama, founded by Marietta Johnson in 1907; and the experimental elementary school at the University of Missouri (Columbia), founded in 1904 by Professor Junius L. Meriam. The common goal of all was to eliminate the school's traditional stiffness and to break down hard and fast subject-matter lines. Each school adopted an activity program. Each operated on the assumption that education was something that should not be imposed from without but should draw forth the latent possibilities from within the child. And each believed in the democratic concept of individual worth.

Dewey, whose writings and lectures influenced educators throughout the world, laid the foundations of a new philosophy that affected and continues to affect the whole structure of education, particularly at the elementary level. His theories were expounded in *School and Society* (1899), *The Child and the Curriculum* (1902), and *Democracy and Education* (1916).

The views of John Dewey

For Dewey, philosophy and education render service to each other. The one indicates what social values are desirable; the other seeks to promote them. Education becomes the laboratory of philosophy. Society should be interpreted to the child through his daily living in the classroom, which acts as a miniature society. Education leads to no final end; it is something continuous, "a reconstruction of accumulated experience," which must be directed toward social efficiency. Education *is* life and not merely a preparation for life.

The influence of progressive education advanced slowly during the first decades of the 20th century. The experimental schools touched only a small part of the population, and World War I interrupted their increase. A number of progressive schools, however, were established, such as the Play School and the Walden School in New York City, the Shady Hill School in Cambridge, Massachusetts, the Elementary School of the University of Iowa, and the Oak Lane Day School in Philadelphia.

Experimental schools sponsored by universities found few imitators, but the idea of a democratic education through individual activities, which was a fundamental principle in their experimentation, led progressive-minded teachers to try out new ways of their own in their schools. Dewey had given them a philosophy of education for general guidance, but for the most part their approach was not doctrinaire. They wanted a freer, more individual kind of schooling. It was for them to find out by practical trial how to achieve this result under ordinary conditions.

The spread of progressive education was more rapid in the third decade of the 20th century. The movement was not confined to any particular country. In the United States the progressives felt that it was necessary to band together, with the result that the Progressive Education Association (PEA) was formed in 1919. The PEA did much to further the cause of progressive education until it ended, as an organization, in 1955. As in the United States, Europe's leading progressives found it desirable to band together, and in 1921 the New Education Fellowship was formed. During World War II the activities of the New Education Fellowship were curtailed, but it managed to keep going and eventually acquired national sections in 25 countries under the new name of the World Education Fellowship.

Although the name progressive education has fallen somewhat into disrepute since the 1950s, the notions it expressed have had deep influence on public school systems everywhere. Activity programs, imaginative writ-

ing, projects, trips, movable desks, and space in every classroom for dramatics and informal activities are found in almost all school systems. The grading system has been done away with in many schools, and the results of individual research are often pooled and reported to the class. Children are given more responsibility for the social life of their school and so are practiced in citizenship as they never were before. And these developments are common, in varying degree, all over the world.

Scientific-realist education. The scientific-realist education movement began with an appeal by the ladies who had been put in charge of special schools opened for backward and abnormal children in Geneva, Switzerland (1900). Their appeal was made to a young doctor, Édouard Claparède, then an assistant in the Psychological Laboratory of the University of Geneva. He responded to their appeal, and this experience brought him to realize some of the defects of ordinary schools. In a lecture to the Medical Society of Geneva, he suggested the need for *l'école sur mesure* ("the school made to measure"). Not as much thought is given, he argued, to the minds of children as to their feet. Their shoes are of different sizes and shapes, made to fit their feet. When shall we have schools to measure? The psychological principles needed to adapt education to individual children were expounded in his *Psychologie de l'enfant et pédagogie experimentale* (1909). Later Claparède took a leading part in the creation of the Institut J.-J. Rousseau in Geneva, a school of educational sciences to which came students from all over the world.

Jean Piaget's studies of child development

Theorists such as Claparède hoped to provide a scientific basis for the whole of the educational enterprise, an aim that was furthered by the Swiss psychologist Jean Piaget, who studied in a philosophical and psychological manner the intellectual development of children. Piaget argued, on the basis of his observations, that development of intelligence exhibits four chief stages and that the sequence is everywhere the same, although the ages in the stages of development may vary from culture to culture. He starts by defining thinking as an action, a sort of manipulation. It begins as an operation upon actual and material things and objects—upon water, or air, or a ball. These operations are then interiorized. They become part of the living organism itself, mind as well as body, and they are gradually generalized and formalized.

The first stage in the development takes place at the infant stage, when the child, even before he learns to speak, puts objects together (addition), then separates them (subtraction). He perceives them as collections, rings, networks, groups. By the age of two or three, a basis has been laid. Kinetic muscular intelligence has developed to some degree—the infant thinks with his fingers, his hands, his body. Aided by language, the capacity for symbolic thinking slowly develops. This constitutes the second stage. Up to the age of seven or eight, some of the fundamental categories of adult thinking are still absent: there is seldom any notion, for instance, of cause and effect relationships.

The third stage is that of concrete operation. The child has begun to know how to deal with mental symbols and acquires abstract notions such as "responsibility." But he operates only when in the presence of concrete objects that can be manipulated; he still needs the crutch of experience. Pure abstract thinking is still too difficult for him. Nevertheless, notions of classification, series, one-to-one correspondence, whole numbers, fractions, space, speed, time, and chance gradually become clear—clear but always limited and bounded. At this stage the teaching must be exceedingly concrete and active; purely verbal teaching is out of place.

Only after about 12 years of age, with the onset of adolescence, does the power develop to deal with formal mental operations not immediately attached to objects. Only then do theories begin to acquire real significance, and only then can purely verbal teaching be used.

The child's *total* development, particularly his emotional and social growth, also concerned educational reformers. They pointed out that it is wrong to assume that incentives to mental effort that are effective with adults are equally effective with the young. The English philosopher Alfred N. Whitehead, in his doctrine of the "Cycle of Interests," put forward a theory very much in line with the ideas of the reformers. Romance, precision, and generalization, said Whitehead, are the stages through which, rhythmically, mental growth proceeds.

> Education should consist in a continual repetition of such cycles. Each lesson in a minor way should form an eddy cycle issuing in its own subordinate process.

Whitehead believed that any scheme of education must be judged by the extent to which it stimulates a child to think. His warning to teachers is:

> In training a child to activity of thought, above all things we must beware of what I will call 'inert ideas,'—that is to say, ideas that are merely received into the mind without being utilised, or tested, or thrown into fresh combinations.

From the beginning of his education, the child should experience the joy of discovery. The discovery he has to make is that general ideas give an understanding of the stream of events that shape his life.

Social-reconstructionist education. Social-reconstructionist education is based on the theory that society can be changed and reconstructed through the complete control of education. The objective is to change society to conform to the basic ideals of the political party or government in power or to create a utopian society through education. This movement has progressed from the experimental stage to a system of education that controls a large fraction of the world's population.

Communist education, a variety of social-reconstructionism, is based upon the philosophy propounded by Karl Marx, Friedrich Engels, and Lenin. M.I. Kalinin, formerly president of the U.S.S.R., expressed the general aims of Communist education in a short book, *On Communist Education* (1950), as follows:

The aims of the Soviet education system

> Communist principles, taken in their elementary form, are the principles of highly educated, honest, advanced people; they are love for one's socialist motherland, friendship, comradeship, humanity, honesty, love for socialist labour and a great many other universally understood lofty qualities. The nurturing, the cultivation of these attributes, of these lofty qualities, is the most important component part of communist education.

Side by side with his general and technical education, the Communist pupil is constantly reminded of Marxist principles and their applications.

> To learn Marxism means to know after mastering the Marxist method how to approach all other problems connected with your work. If, let us say, the sphere of your future work is agriculture, will it be of advantage to employ the Marxist method? Of course it will. But to employ the Marxist method, you have to study agriculture, too, you have to be an agricultural expert. Otherwise nothing will come of your attempt to apply Marxism in agriculture. This should not be forgotten if you wish to apply Marxism in practice, if you wish to be men of action, and not textmongers of Marxism. But what does being a Marxist mean? It means being able to adopt the correct line. But to be able to adopt the correct, Marxist line, you also need to be a first-rate expert in your particular sphere of activity. (*Ibid.*, p. 16.)

Communist education is strictly controlled as to both curricula and methods. No religious instruction is permitted. From the 1950s onward, much attention has been paid to the ideal of "polytechnization." Man, so the argument runs, is not simply *Homo sapiens* but rather *Homo faber*, the constructor and builder. He attains full mental, moral, and spiritual development through entering into social relations with others, particularly in cooperative efforts to produce material, artistic, and spiritual goods and achievements. The school should prepare pupils for such productive activities—for instance, by studying and, if possible, sharing in the work done in field, farm, or factory. Stress should be laid in the curriculum on science, which is the theory of production, and on mathematics, which is the language of science. The pursuit of these polytechnical aims has led to interesting and valuable experiments and has undoubtedly contributed to rapid industrialization.

Educational aims of the Nazis

The Nazi social-reconstructionist education system in Germany (1933–45) was an attempt to insure that the totalitarian government had complete control over the populace. Even before the National Socialists came to power, Hitler in *Mein Kampf* hinted at far-reaching educational changes. The Ministry of Public Enlightenment and Propaganda exercised control over virtually every form of expression—radio, theatre, cinema, the fine arts, the press, churches, and schools. The control of the schools began in March 1933 with the issuing of the first educational decree, which held that "German culture must be treated thoroughly."

The Nazi government reorganized the school curriculum in accordance with Nazi beliefs. *Rassenkunde* ("race science") was given first place in the school curriculum. A major part of biology became race science, and health education and physical training did not escape the racial stress. Geography became geopolitics, the study of the fatherland being fundamental. Physical training was made compulsory for all: the body came first, character next, and the intellect last.

The Nazi education system was designed to control completely the education of the young, from the primary schools through the universities, and in their everyday life. The secondary schools were run as National Socialist *Erziehungsstatten* (places of education), and youth labour service was made compulsory. An integral part of the Nazi education were the Hitler Youth organizations, in which the program of indoctrination so relentlessly carried out in the school was supplemented along military and political lines.

Reconstructionist education in Israel

Another social-reconstructionist movement, and a very different one, is that of the kibbutzim (collective farms) of Israel. The most striking feature of kibbutz education is that the parents forgo their natural desire to rear and educate their offspring and instead hand them over to the "Children's House" to be educated by professionals (specially appointed nurses and educators). What makes this more astonishing is the fact that children are usually withdrawn from the care of their parents at a very early age and, in some kibbutzim, immediately after birth.

The kibbutzim type of education developed for both practical and economic reasons. It released mothers from the exacting duty of looking after and rearing their own children, thus making it possible for them to work side by side with the men. The labour force was thus augmented, since only a small staff is required to look after the children. It also enabled the members of the group to offer their children conditions of material comfort, which they as individuals could not afford to give them. With the development and consolidation of the kibbutz, the economic motive behind collective education gradually receded, and ideological and educational considerations gained prominence. These were: (1) that the kibbutz way of life makes for complete equality of the sexes; (2) that the education of children in special children's houses is the best way of perpetuating the kibbutz way of life; (3) that collective education is more "scientific" than education within the family, inasmuch as children are reared and trained by experts (*i.e.*, qualified nurses, kindergarten teachers, and other educators), in an atmosphere free of the tensions engendered by family relationships; and (4) that collective education is more democratic than the traditional education within the family, and more in keeping with the spirit of cooperative living. It is, indeed, the practical realization of the ideal of educational equality, which is in turn the corollary of social and economic equality.

The system of education adopted in the kibbutzim tended to be influenced by the ideas of Sigmund Freud. The literature produced by the ideologists and teachers in the schools shows that they accept the Freudian conception of the main task of education: the problems of child management and discipline, the legitimate use of reward and punishment, the educator's attitude to the child's impulses and bodily desires and the attempt to sublimate them—all of these figure quite prominently. The problems are dealt with in the light of Freud's warning that an outright authoritarian approach on the part of the educator may be harmful to the child. The kibbutz teacher and educator constantly strive for new methods to replace the authoritarian approach of the traditional educator.

Kibbutz education has, at least in theory, been on the whole permissive and opposed to unmotivated prohibitions. It does not inculcate a consciousness of sin, nor does it promote taboos. It does not regard any instinctive behaviour as intrinsically evil or wicked and rejects corporal punishment as morally wrong and harmful to the child. Kibbutz educators view punishment of any sort as undesirable.

In the beginning the kibbutz was exclusively an agricultural unit, but now it is often mixed: industrial as well as agricultural. The former prejudice against academic studies is decreasing. This change is evident in the variety of subjects studied by young people. It is understood that at the completion of his studies the student will return to the kibbutz, which is essentially a democratic institution, the educational policy of which will always be influenced by the wishes of the parents.

WESTERN PATTERNS OF EDUCATION

United Kingdom. *Early 19th to early 20th centuries.* English education has been less consciously nationalist than that of continental European countries, but it has been deeply influenced by social class structure. Traditionally, the English have held that the activity of the government should be restricted to essential matters such as the defense of property and should not interfere in education, which was the concern of family and church. The growth of a national education system throughout the 19th century continued without a clear plan or a national decision. The cornerstone of the modern system was laid by the Elementary Education Act of 1870, which accepted the principle that the establishment of a system of elementary schools should be the responsibility of the state. The act divided the country into districts. Where voluntary schools were insufficient, school boards were elected to levy taxes, maintain elementary schools, and pay the fees of needy children. The act of 1870 did not eliminate the traditional prominence of voluntary agencies in the sphere of English education. Nor did it provide for secondary education, which was conducted largely by voluntary fee-charging grammar schools and "public" schools. These public schools were usually boarding schools charging rather high fees. Their tradition was aristocratic, exclusive, formal, and classical. Only a small proportion of the population was served by them. Their main goal was to develop "leaders" for service in public life. In 1900, one child in 70 could expect to enter a secondary school of some kind. The grammar schools copied the curriculum of the public schools, so that only the intellectual and social elite were able to attend.

The Elementary Education Act of 1870

In 1899 an advance was made toward the development of a national system encompassing both elementary and secondary education by creating a Board of Education as the central authority for education. The next step was the passing of the Education Act of 1902, known, after its sponsor, as the Balfour Act. This established a comprehensive system of local government for both secondary and elementary education. It created new local education authorities and empowered them to provide secondary schools and develop technical education. Under the act the county and county-borough councils became the Local Education Authority. School boards were abolished. All councils under the act were required to appoint an education committee with a proportion of co-opted members with experience of education, of whom some were to be women. Local education authorities were "to supply or aid the supply of education other than elementary." Taking advantage of this, local education authorities provided many new secondary schools that were mainly of one type. In regulations issued in 1904, Sir Robert Morant, the first secretary of the Board of Education, disregarding the opportunities offered by the wide terms of the act, set a pattern for their development based on the academic tradition of grammar and public schools.

It was the voluntary-school issue that made the act one of the most fiercely contested measures of the 20th century, especially the decision to give voluntary schools the benefits of rate (local-tax) aid. Voluntary schools (former board schools) were also brought under the general control of the new local education authorities. But the managers of voluntary schools retained the power of appointing teachers, subject to the consent of the local authority, which could not be withheld, except on educational grounds. The managing body was to consist of four foundation managers and two local-authority representatives, and the religious teaching was to be in accordance with the "trust deed" of the school. As a result, the dual system agreed upon in 1870 was not only confirmed but strengthened, while at the same time unified administration of all elementary schools under the new local education authority opened the way to general advance in this important branch of education.

Provisions of the act of 1918

The Education Act of 1918 (The Fisher Act), re-enacted with surviving earlier legislation in consolidated form in the Education Act of 1921, aimed at the establishment of a "national system of public education available for all persons capable of profiting thereby." Local authorities were called upon to prepare plans for the orderly and progressive development of education. The school-leaving age was raised to 14, and power was given to local authorities to extend it to 15.

Education Act of 1944. The Education Act of 1944 involved a thorough recasting of the educational system. The Board of Education, with its president empowered only to superintend, was replaced by a minister who was to direct and control the local education authorities. He could overrule local authorities, governors, and managers if they appeared to him to be acting unreasonably. Vested with such authority, a minister could do much to assure a more even standard of educational opportunity throughout England and Wales and could determine priorities and take the lead in planning. The act provided two promising planning devices. First, every local education authority was required to submit for the minister's approval a development plan for primary and secondary education and a plan for further education in its area. Second, two central advisory councils were constituted, one for England, another for Wales. Unlike the former consultative committee, they had power, in addition to dealing with problems set by the minister, to tender advice on their own initiative. The total number of education authorities in England and Wales was reduced from 315 to 146. To enable county education committees to keep in touch with local circumstances, provision was made for a considerable delegation of power to divisional executives, who with the various bodies of school governors and managers do much to ensure local interest and initiative.

The complexity of the education system in England and Wales arises in part from the pioneer work done in the past by voluntary bodies and a desire to retain the voluntary element in the state system. The act of 1944 continued the religious compromise expressed in the acts of 1870 and 1902 but elaborated and modified it after much consultation with the parties concerned. The act required that, in every state-aided primary and secondary school, the day should begin with collective worship on the part of all pupils and that religious instruction should be given in every such school. As in earlier legislation there was, however, a conscience clause and another to ensure that no teacher should suffer because of his religious convictions.

Structural changes under the 1944 act

Two fundamental reforms in the act were the requirement of secondary education for all, a requirement that meant that no school fees could be charged in any school maintained by public authority; and replacement of the former distinction between elementary and higher education by a new classification of "three progressive stages to be known as primary education, secondary education, and further education." Thus, the primary stage, compulsory from age five, covers the period from two years old to 11–12 years old; the secondary stage, compulsory to 15, covers from 11–12 years of age to 18–19, while further education is officially described as the "post-secondary stage of education, comprising all vocational and non-vocational provision made for young people who have left school, or for adults." Further education thus embraces the vast range of university, technical, commercial, and art education and the wide field of adult education and also, in the words of the act, facilities for

> leisure-time occupation, in such organized cultural training and recreative activities as are suited to their requirements, for any persons over compulsory school age who are able and willing to profit by the facilities provided for that purpose.

The Ministry of Education's interpretation of the type of secondary education to be provided under the act was the touchstone applied to every local development plan submitted. On this interpretation, to provide an adequate secondary education in accordance with "age, ability, and aptitude," it was necessary to establish separate schools with different programs, to which children were recruited at 11 by means of selection tests. If, in exceptional circumstances, such provisions were made in a single school, then the school must be large enough to comprise the three separate schools under one roof.

The tripartite system of grammar, technical, and secondary modern schools did not, in fact, flourish. The ministry had never been specific about the proportion of "technically minded" children in the population, but, in terms of school places provided in practice, it was about 5 percent. Since, on the average, grammar school places were available to 20 percent, this left 75 percent of the child population to be directed to the secondary modern schools for which the ministry advocated ad hoc courses not designed to lead to any form of qualification.

The comprehensive movement. Selection procedures at the age of 11 proved to be the Achilles' heel of the grammar school–secondary modern system. Stress was laid on the fact that there was no longer a "competitive examination" but only an "objective process of allocation." "Intelligence" tests were generally described as scientific measures of a child's inherent powers, so exact as to permit an accurate forecast of what he would achieve in later school life and indeed life itself. Various developments, however, contributed to the downfall of selection at 11: first, the examination successes of the secondary modern schoolchildren; second, the failure of a significant proportion of the children so carefully selected for grammar schools; third, the report of a committee appointed by the British Psychological Society to examine the validity of "intelligence" tests. This report supported arguments that education itself promotes intellectual development and that "intelligence" tests do not in fact measure genetic endowment but rather educational achievement. The National Foundation for Educational Research in 1957 issued a report in which it was estimated that some 12 percent of children were allocated to the wrong schools at the age of 11 and that not even the most stringent methods of allocation could reduce the error to below 10 percent. In the light of such findings, the bipartite structure of secondary education began to be modified, especially in its selective aspects, but with many local variations.

The main issue in the 1950s and 1960s was whether or not the grammar schools should be retained with selection at 11 plus. One of the main arguments used was that the right of "parental choice" must be upheld. Another was that it was in the "English tradition" to retain a selective system. But gradually the number of comprehensive (nonselective) schools increased until, by 1970, nearly half the children were being taught in one or another variant of such schools. There was widespread agreement either that selection should be abandoned altogether or, at least, that the age at which selection operated should be raised to 13 or 14.

Policies of Labour and Conservative governments

The Labour Party during the election of 1964 promised to promote the establishment of the comprehensive school and to abolish selection at 11 plus. Indeed, one of the first acts of the Labour government on taking office was to present plans for reorganizing secondary education. But, instead of legislating, it issued a circular in

the belief that this would enlist local support and encourage local initiative. The result was conflict between national policy and local policy in areas in which the education committees had a temporary majority to vote down the national policy. Although many local authorities (in 1970, 26 out of 162) completely abandoned selection at 11 plus and 104 declared their intention to do likewise, the course that secondary education will take in the future has not been settled. The Conservative government elected in 1970 has declared its intention of leaving decisions about reorganization to the local authorities. Some would prefer to maintain both grammar and comprehensive schools. Others think that the coexistence of both is not possible.

The future course of the public schools (fee-paying private schools), with their attendant private preparatory schools, has yet to be determined. It is far from clear whether they can, in fact, be fully integrated into the national system.

Further education. It is in this sector of education, which is concerned with education beyond the normal school-leaving ages of 15, 16, or 18, that the most complex problems have arisen, particularly since the end of World War II. Many of them are the outcome of the astonishing growth in the number of students, this growth itself being the result of the growth of population and the growth of aspirations. In addition, the explosion of knowledge and the creation of many new professions raise difficulties about the content and structure of courses and the requirements for certificates, diplomas, and degrees.

Many new postschool technical colleges were founded in the early years of the century. The Fisher Act of 1918 empowered the local authorities to levy a rate (tax) to finance such colleges. The universities, on the other hand, received funds from the central government through the University Grants Committee, established in 1911 and reorganized in 1920, after World War I. It was empowered to inquire into the financial needs of university education and to advise the government as to the application of any grants that might be made by the government. Student numbers rose between the war years to more than 50,000 but dropped below 40,000 during World War II. By 1970 the number had risen to well over 300,000.

Organization of higher education

Higher education in Britain is divided into two main branches: technical and university. Technical education operates under a system in which technical colleges are divided into four groups—local, area, regional, and colleges of advanced technology—depending on the level of the courses they offer. The colleges of advanced technology were granted the right to issue a new graduate equivalent award, the Diploma in Technology, and a supplementary postgraduate qualification: membership of the College of Technologists.

In the 19th century the dominance of Oxford and Cambridge was challenged by the rise of the civic universities, such as London, Manchester, and Birmingham. Following the lead of the 18th-century German universities and responding to a public demand for increased opportunity for higher education, Britain's new civic universities quickly acquired recognition—not only in technological fields but also in the fine and liberal arts.

The financing of higher education in the universities is through the University Grants Committee, which provided two-thirds of the income for the universities in 1947. During the 1960s many new universities were founded: their number increased from 23 to 44. Several of them have a strong scientific and technological bias.

A new type of technical college was established in the 1960s—the polytechnic, which provides mainly technological courses of university level as well as courses of a general kind in the arts and sciences. Polytechnics will be chartered to award degrees awarded by a Council for National Academic Awards.

Scotland and Northern Ireland. The educational systems of Scotland and Northern Ireland are separate and distinct from that of England and Wales, although there are close links between them. The essential features of the Education Act of 1944 of England and Wales were reproduced in the Education Act of 1945 in Scotland and in the Education Act of 1947 in Northern Ireland. There were such adaptations in each country as were required by local traditions and environment. In both countries compulsory attendance is from five to 15 and will ultimately be 16. The system in each consists of primary schools from five to 11 years, compulsory secondary education for all to 15 and to 18 on a voluntary basis, and further education from 15 to 18. In Scotland the schools that provide further education are called junior colleges. Secondary education in both countries must be appropriate to the abilities and aptitudes of pupils and is given either in the same school or, as in Scotland, in separate schools. By the 1970s approximately 40 percent of the Scottish schools had become comprehensive. In Northern Ireland secondary education is given in intermediate schools, technical intermediate schools, and grammar schools. As in England and Wales, the local education authorities have been requested to prepare reorganization schemes on comprehensive lines for submission to the secretary of state in Scotland and to the minister of education in Northern Ireland.

Religious instruction is made a regular part of the school curriculum, and in Northern Ireland a period of collective worship each week is required by the Education Bill, rights of conscience being safeguarded.

Germany. *Imperial Germany.* The formation of the German Empire in 1871 saw the beginning of centralized political control over the whole country. What was desired was a system of education that would make the great bulk of the population not only literate but also pious and moral, vocationally and economically efficient, and above all obedient and disciplined. A special type of education was reserved for the minority of citizens in the upper social and economic strata. By separating and differentiating the education of the masses from that of the upper class, the state assured itself, on the one hand, of a vast number of trained and subservient followers and, on the other hand, of a small elite corps equipped to fill the higher positions in the *Reich*.

The German school structure

The education system was structured to accomplish this goal. During the last decade of the 19th century, it was re-examined and overhauled, but the changes did not remove class privileges. Fees, for example, continued to be charged in secondary schools, and no ladder of opportunity was provided to make possible upward mobility.

The *Volksschule* was universal, free, and compulsory. The fundamental subjects were taught along with gymnastics and religion, which held important places in the curriculum. Girls and boys were taught in separate schools except when it was uneconomical to do so. Boys usually received training in manual work, and girls were introduced to domestic science. Graduates of the *Volksschule* found it almost impossible to enter the secondary school, as they graduated at 14 and the entry age to the secondary school was nine. The young graduates of the *Volksschule* had no other destiny than work and continuation school. The *Vorschule* was a private elementary school that charged fees for its three-year program and prepared its pupils for the secondary schools. This was the school attended by the children of the well-to-do.

The secondary schools, which were designed for the intellectual and social elite, were of several kinds. For boys there was the *Gymnasium*, which stressed Latin and Greek. The *Realgymnasium* offered a curriculum that was a compromise between the humanities and modern subjects. It required Latin but not Greek. The *Oberrealschule* stressed modern subjects, particularly sciences and mathematics, and, instead of Latin and Greek, modern languages were taught. A boy about to enter secondary school at the age of nine thus had to decide on one of the three types of curricula.

Although Kaiser William II threw his influence on the side of the modernists in 1890, the *Gymnasium* continued to overshadow the other two schools.

The education of girls was not strongly encouraged. Secondary schools for them were recognized by Prussia in 1872 and were extended and improved in 1894 and

again in 1908. These schools were fee paying and were thus available chiefly to the upper social and economic strata. The course of instruction lasted ten years, from six to 16. This ten-year school was called the *Lyzeum*, the first three years being preparatory. Beyond it was the *Oberlyzeum*, which was divided into two courses: the *Frauenschule*, which offered a two-year general course; and the *Lehrerinnenseminar*, which offered a four-year course for prospective elementary school teachers. Girls who wanted a secondary school education similar to that of the boys transferred at the age of 13 to the *Studienanstalt*.

The continuation school (*Fortbildungsschule*) made rapid strides during the early decades of the 20th century. It was a form of trade school for the graduates of the *Volksschule*.

Weimar Republic. In no sphere of public activity did the establishment of the Weimar Republic after 1919 cause more far-reaching changes than in that of education. The elementary school, which under the old regime was a class school, became a national institution. The concept of the *Einheitsschule* (common school) became an accomplished fact through the establishment of the four-year *Grundschule*, which was free and compulsory for all children. The *Vorschule* was abolished.

The *Grundschule* provided basic training. Besides the elementary subjects and religion, the child was instructed in drawing, singing, physical training, and manual work. The *Oberstufe*, the four upper classes of the elementary school, combined with the *Grundschule*, formed a complete whole. Most elementary schools thus provided an eight-year course of study.

Intermediate schools (*Mittelschulen*) were established for children who wished a longer and more advanced elementary school course and were able to pay modest fees. There were also Latin schools, or *Rektoralschulen*, which were lower secondary fee-paying schools maintained in the smaller localities. The first three years were common to all, but the last three years offered differentiated programs. Some intermediate schools were designed to lead to secondary schools and included a second foreign language in their curricula. The Weimar Constitution preserved the religious tradition, which had been an essential part of the school curriculum in Germany since the Reformation. No pupil, however, could be compelled to study religion, and no teacher could be forced to teach it. Communities were accorded the right to establish schools in accordance with the particular religious beliefs of the pupils.

Secondary schools in the Weimar period

As regards secondary education, the Weimar Republic kept the prewar division of *Gymnasium*, *Realgymnasium*, and *Oberrealschule*. In addition, there was established the *Aufbauschule*, which was a six-year school following completion of the seventh year of the elementary school, and the *Deutsche Oberschule*, a nine-year school that required two modern foreign languages. True to its name, it stressed German culture.

There were several types of secondary schools for girls. The school comparable to the *Gymnasium* was the *Gymnasiale Studienanstalt*. The *Oberlyzeum* gave more attention to the modern languages. The school comparable to the *Oberrealschule* was the *Oberlyzeum der Oberrealschulrichtung*, which focussed on the sciences (for an account of educational policies during the Nazi period, see above *Social-reconstructionist education*).

Changes after World War II. The basic law of the Federal Republic of Germany (West Germany) of May 1949 granted local autonomy in educational matters, and the 11 *Länder* (provinces) and city-states passed their own laws. In 1949 the Standing Conference of Ministers of Education was established to discuss problems common to all; it is advised by a federal educational council created in 1953.

Education was made compulsory from the ages of six to 18 years. Of this period eight or nine years are spent in a full-time school of general education. The pupil leaving it at the age of 14 or 15 must spend three more years in a part-time vocational school. Tuition fees and supplies are free in all the systems. Full-time attendance to the age of 15 was made compulsory in Berlin, Bremen, Hamburg, Hesse, Lower Saxony (now Niedersachsen), Saarland, and Schleswig-Holstein. In Baden-Württemberg it is compulsory up to that age only in some of the larger communities and districts. The other *Länder* have proposed the introduction of a ninth compulsory year as soon as possible. There are only privately supported preschools and kindergartens. There is a four-year *Grundschule* course in the *Länder* and a six-year course in the city-states. This is followed by four years in the upper division of the elementary school or by three, four, or five years in an intermediate school (*Mittelschule*) or six, eight, or nine years in an academic, technical, or general secondary school leading to the university. In the city-states the *Einheitsschule* (common-school) plan—combined elementary and secondary—has been introduced. It has a uniform basic grade and an upper grade consisting of practical, technical, and economic branches. The transition from the elementary school to secondary education in the *Länder* is still selective, in that admission requires the passing of an entrance examination.

In order to provide able young people, already at work, with an opportunity of preparing for the maturity (matriculation) examination in full-time secondary courses, the *Kolleg* (institute preparing for the maturity examination) has been established. The *Kolleg* usually covers three years and gives access to higher education.

France. *The Third Republic.* The establishment of the Third Republic (1870) brought about the overhaul and renovation of the French schools, in the process of which education became a national enterprise. In 1882 primary education was made compulsory for all children between the ages of six and 13. In 1886 members of the clergy were forbidden to teach in the schools, and in 1904 the teaching congregations were suppressed. France had thus established a national free, compulsory, and secularized system of elementary schooling.

The secondary schools charged fees. Their purpose was to bestow a liberal education and to select an elite for administration and leadership. Two types of secondary schools existed—the *écoles secondaires communales* and the *lycées*. Classical studies were emphasized, and, although this tradition was heavily criticized, very few modifications were made. One of the first attempts was a reform of 1902 that promoted the study of modern languages and of science.

Centralized administration in French education

The administration of education in France is highly centralized and is concerned with every phase of national education. At the head of the system is the minister of national education, a member of the Cabinet and usually a member of the legislature. The minister is advised and assisted in the execution of his duties by a hierarchy of administrators, inspectors, and other officials. France is divided into about 40 administrative areas, each known as an "academy." The chief education officer is the rector, the minister's most important representative, who administers the laws and regulations. The largest and most important body is the Higher Council for National Education, which reports to the minister on questions of administration, studies, textbooks, and discipline.

Changes after World War II. In 1947 a commission established to examine the educational system recommended a thorough overhauling of the entire school system. Education was to be compulsory from the ages of six to 18. Schooling was to be divided into three successive stages: (1) six to 11, aimed at mastery of the basic skills and knowledge that enable the child to understand and to be understood; (2) 11 to 15, a period of guidance to discover aptitudes; (3) 15 to 18, a stage during which education was to be diversified and specialized.

Preschool education is given in the *école maternelle*, in which attendance is voluntary from the age of two to six years. Schools are restricted to 150 students. The curriculum is varied, consisting of physical and breathing exercises, games and movement with songs, manual work, and drawing. There are also exercises in observation; exercises for the development of moral habits; language and recitation; and introduction to reading,

writing, and arithmetic. French educationists are justly proud of the quality of their preschool education.

Education is compulsory between six and 16 years of age and is free. In every community of over 500, separate schools for boys and girls must be maintained, unless special permission is given to operate a mixed school. The program is the same for all children between the ages of six and ten. Reading and writing, arithmetic and metric system, history and geography, civic education, first notions of natural science and other scientific subjects, drawing, singing, manual work, and physical training are taught.

Secondary schools (*lycées*) provide either four-year or seven-year courses and tend to specialize in either academic or technical studies. It is usual to speak of a first cycle (11–15) and a second cycle (15–18). The first cycle (11–15) provides general education and is the same in all schools. During this cycle the student is observed by the teacher and a guidance council and, at the age of 15, enters into the second cycle. In the second cycle (15–18), the student is given a choice between occupational training or academic training, depending on his preferences and ability. The occupational training is divided into three courses. The first course involves full-time training over a two-year span, which is given in colleges of secondary education. The curriculum is not specialized and prepares pupils for industrial, commercial, or administrative occupations. At the end the students sit for the diploma of professional studies (*brevet d'études professionnelles*—B.E.P.). The second option consists of two years of either direct apprenticeship or training courses with a narrower form of specialization. Students sit for the certificate of professional aptitude (*certificat d'aptitude professionnelle*—C.A.P.) at the end of the second year. The third possibility is full-time training over a one-year period. This is designed for students who go to work without training, once compulsory schooling is over at age 16. It leads to the certificate of professional training (*certificat de formation professionnelle*—C.F.P.).

Those choosing academic training enter one of four streams: classical, economic, scientific, or technical. In the second year a third foreign language may be offered. Students study for the *baccalauréat*, which gives admission to the university. Students whose marks are not sufficient to obtain the *baccalauréat* may qualify for the certificate of completion of secondary studies.

Higher education is offered in universities, in institutes attached to a university, and in the *grandes écoles*. Students attend for two to five years and sit either for a diploma or, in certain establishments, for university degrees or for a competitive examination such as the *agrégation*.

Reform in the French universities

The existing universities are being decentralized by the creation of colleges, institutes, or schools, each with much autonomy. The entire university system is being drastically reformed, allowing for a measure of student participation in the running of universities, encouraging diversification of courses, and diminishing the importance of examinations. Undergraduate courses last for three or four years, depending on whether the B.A. or M.A. degree is sought.

The higher doctorates are considerably more difficult to obtain and represent solid research or scholarly achievement. They are essential to the award of a full professorship.

Other European countries. *Italy.* Education in Italy up to 1923 was governed by the Casati Law, passed in 1859, when the country was being unified. The Casati Law organized the school system on the French plan of centralized control. School attendance was compulsory only for the years six to nine. Religious education was abolished in the schools by the Coppino Act of 1877, which also extended compulsory school attendance to four years. The Orlando Law of 1904 extended compulsory education, at least in the larger communities, to the age of 12.

In 1923 the entire national school system was reformed. The principle of state supremacy was reinforced by introducing at the end of each main course of studies a state examination to be taken by pupils both from public and private schools. In 1935 Cesare de Vecchi, a violent Fascist leader, became minister of education. He dissolved consultative educational bodies and vested all power in the minister himself. Confusion and chaos followed, and reforms were commenced in 1939. They were accompanied by a reorganization of the youth movements (started in 1937), which played an important role in all schools.

Since World War II, attendance at an elementary school for five years has been compulsory. Elementary schools educate pupils from the ages of six to 11. Infant schools, which are not compulsory, take children between the ages of three and five. These include nursery schools, kindergartens, and similar institutions.

Nonstate schools are those whose operation is only in part or not at all the responsibility of the state, and they are divided into officially recognized schools (*scuole parificate*), private schools (*scuole private*), and subsidized schools (*scuole sussidiate*).

From 11 to 14, young people also compulsorily attend secondary school (*scuola media*). The *scuola media* opens the way to art institutes, professional institutes, and technical schools (which are not obligatory) and offers courses of varying duration, leading to a diploma. Because this diploma does not allow the pupil to continue his studies, these schools have gradually declined in importance.

The *scuola media* also leads to schools that in turn give entry to the universities. These schools—technical institutes and grammar schools (*licei*)—offer courses lasting five years, except for those of the teacher-training institutes (*istituto magistrale*) and art schools (*liceo artistico*), which last only four years. Technical institutes can specialize in agriculture, crafts and commerce, tourism, surveying, industry, and nautical studies. There are also technical schools for girls only. There are also two types of *licei*, one tending to specialize in classics and the other in science. The technical institutes, the *licei*, and the teacher-training institutes open the way to university education.

Italian university education

In Italy there are many state universities and some independent ones. Italian universities are made up of a varied number of faculties, each of which can award one or more degrees. One characteristic of the university in Italy is the large number of *fuori corso* (students who have not managed to complete their studies in the normal period of time, which ranges from four to six years, depending on the faculty). University studies require the passing of a variable number of examinations, at the end of which the students sit for a degree (*laurea*), which gives them the title of *dottore*. To be able to exercise any profession, such as that of lawyer, doctor, or business consultant, one must take a state examination.

Netherlands. The first modern school law in the Netherlands was passed in 1801, when the government laid down the principle that each parish had the right to open and maintain schools. A debate between the proponents of denominational and nondenominational schools went on during the 19th century. The controversy was closed by a law of 1920, which declared that denominational schools were fully equal with state schools, both types being eligible for public funds. The resultant decentralization is unique in western Europe. The private school, in return for public funds, must provide a curriculum equivalent to that offered by the public schools.

The Ministry of Education and Science supervises the quality of education, while the local authorities in the provinces administer it. Under the minister is the secretary general for education, who acts as the permanent chief of the ministry. The Education Council advises the ministry on all important general questions that are of direct concern to the ministry.

Education is compulsory from the age of seven to 15, although most children start at age six. Elementary education is from six to eight years in duration. Higher elementary schools offer an additional two years of work in practical subjects for those students who do not go

to secondary school. Secondary education consists of the *hoogere burgerschool*, a three- to four-year school leading to lower administrative posts or to the university, the *gymnasium*, which leads to the university, the *lyceum*, which combines the above two, the commercial school, and the girls' high school. There are technical schools at the primary and secondary level and institutions for the training of teachers.

Switzerland. The Swiss constitution of 1874 provided that each canton or semicanton must organize and maintain free and compulsory elementary schools. The federal government exercises no educational function below the university level, except to help finance the municipal and cantonal schools. The federal government administers only the Federal Institute of Technology in Zürich.

The Swiss school system thus consists of 25 cantonal systems, each having its own department of education, which sets up its own school regulations. Private schools for children under the age of compulsory attendance are controlled by the local school authorities. Compulsory school starts at six or seven years of age and continues until 14, 15, or 16, according to the canton. Preschool, from four to six years, is followed by a primary school attendance of six to nine years' duration. Secondary education includes schools of various types, which prepare students for university, higher commercial schools, and secondary technical schools. Subjects taught are diversified, and the pupil's future professional life is a decisive factor in the selection of courses. In addition to the routine courses, there are classes in carpentry, cooking, sewing, wood carving, and other manual skills. Physical education is also part of the curriculum. Vocational training is compulsory for boys and girls during apprenticeship and serves to maintain the skill and high quality of Swiss workmanship.

Agricultural schools, schools of applied arts, technical schools with specialized courses in watchmaking, fine mechanics, ceramics, metalwork, carpentry, silk weaving and the various skills of the hotel and tourist trade are found in Swiss cities. There are seven universities, all under the control of the cantons. A serious attempt is being made to achieve some educational unity out of the great diversity of the Swiss education system.

(J.A.L./R.L.Sw.)

United States. As the United States entered the 20th century, the principles that underlie its present educational enterprise were already set. Educational sovereignty rested in the states. Education was free, compulsory, universal, and articulated from kindergarten to university, though the amount of free schooling varied from state to state, as did the age of required school attendance. Although a state could order parents to put their children to their books, however, it could not compel them to send them to a public school. Parents with sectarian persuasions could send their offspring to religious schools. In principle there was to be equal educational opportunity; unhappily, the practice even now lags behind the precept.

American principle of equal educational opportunity

Expansion of American education. Though such principles have remained the basis of America's educational endeavour, yet that endeavour, like America, has undergone a vast evolution. As the century began, American youth attended an eight-year elementary school, whereupon those who continued went to a four-year high school. Known as the "eight–four system," this organization wholly prevailed till about 1910, when the "six–three–three system" made a modest beginning. Under the rearrangement, the pupil studied six years in the elementary and three in the junior and senior high schools, respectively. In 1900 only a handful of the lower school's alumni—some 500,000—advanced into the high school. Fifty years later, the process had swollen tenfold. Of those who took their high school diploma during the earlier period, some three out of every four entered college. Today, though more than ever throng the college, the ratio is almost reversed, and, for every four disgorged by the high school, only one makes tracks for the higher learning.

There are several reasons for these developments. To begin with, America's secondary school, reflecting the nation's expanding utilitarianism, has attuned itself more and more to preparing the young for everyday living. Consequently, though it still serves prospective collegians the time-honoured academic fare, it has gone to great lengths to accommodate the generality of young America with such useful and practical enlightenment as automobile driving, home nursing, cookery, carpentry, plumbing, writing, acting, and the like. Doubtless, too, national prosperity and, even more, the inarguable cash value that a secondary diploma bestows upon its owner have enhanced the high school's present flourishing state. Finally, there stands the fact that several states now require their young to attend school until their 16th, and sometimes even their 17th, birthday—which makes their attendance at high school almost unavoidable.

Federal involvement in local education

Although the U.S. Constitution has delegated educational authority to the states, this does not mean a lack of federal concern or, more critically, of federal counsel and assistance. Actually, national educational aid is older than the Constitution, having been initiated in 1787 in the form of land grants. Seventy-five years later the Morrill Act disbursed many thousands of acres to enable the states to promote a "liberal and practical education." Soon thereafter, the government created the federal Department of Education under the Department of the Interior. Now the Office of Education in the Department of Health, Education, and Welfare, this agency gathers educational information and broadcasts it in a variety of publications. At the request of state and local governments, it has assisted in several nationwide educational surveys.

Since the passage of the Morrill Act in 1862, the flow of federal aid has steadily increased. Beginning with the Smith–Lever Act of 1914, Congress has legislated measure upon measure to develop vocational education in schools below the college plane. A new trail was opened in 1944, when the lawgivers financed the first "GI Bill of Rights" to enable veterans to continue their education in school or college. At length, in 1965, under the Elementary and Secondary Education Act, the legislature appropriated $775,000,000 to subsidize various worthy projects. Among the law's beneficiaries was the nursery school, a rarity at the dawn of the century but serving some 300,000 pupils by the 1970s. The nursery school is a prekindergarten school. Taking its protégés soon after they can walk, it instructs and entertains them to their sixth birthday. Much of their time is given to play. But they are also introduced to some of life's sober affairs, such as learning to live together, keeping themselves clean and tidy, and the like. Under supervision, they occasionally fare forth into shops, cinemas, zoos, and similar landmarks of civilized life.

Programs of special education

Although, patently, not all men are created equal, yet in the American democracy everyone, regardless of condition, is expected to have his fair chance. Such is the tenet that underlay the establishment of the free, tax-supported common school and high school. As science pointed the way, the effort to bridge the gulf between the haves and have-nots presently extended to those with physical and mental handicaps—the blind, the deaf, the dumb, the crippled, and the mentally defective. Of modest proportion at the century's beginning, teaching the handicapped has since become part of the American way. The federal government, for example, subsidizes the publication of books in Braille for use in state schools. Most states and many cities have long since undertaken programs to teach the handicapped, though financially the going has been difficult. Finally, in 1958, Congress appropriated $1,000,000 to help prepare teachers of mentally retarded children. Thenceforward, federal aid for the handicapped has steadily increased.

If the handicapped child has been an object of educational concern, then the gifted child has generally been neglected. The truth is that, until psychologists and sociologists started to apply their science to the superior child, apart from being occasionally viewed as somewhat freakish, he was not suspected of entertaining any particular problems. Eventually, however, augmented

with federal, state, and sometimes foundation money, one city after another embarked on educational programs for the bright child. The upshot was that, by the mid-1960s, some three out of every four municipalities were reported in one way or another to be thus engaged.

Experiments in education. Besides the development of education for exceptional children, there developed numerous "experiments" or "plans" to bring newer learning theories into practice. Among them there was the Dalton Plan, introduced in 1920 at Dalton, Massachusetts, by Helen Parkhurst. Designed to enable a learner to study at his own pace, it discarded the traditional recitation and instead put the pupil upon a mimeographed monthly assignment, or "contract," including readings, questions, written work, and so forth. He distributed his time to suit his need, labouring more intensely on his weak subjects and easing up on his stronger ones. He might dispose of them at an even daily gait; or he might gallop through them in a few days and then do as he chose. Though the Parkhurst system has disappeared, its shift from short, daily assignments to the study of broad topics over a period of time—nowadays spoken of as "units"—has lingered.

Another attempt to individualize learning was the Winnetka Plan. Introduced in 1919 by Carleton Washburne at Winnetka, Illinois, it, too, banished the recitation. Instead of being held to day-by-day achievements, the Winnetka pupil was to achieve certain planned "goals." Guided by simple directions, he was free to proceed as he pleased. When the task was done, he underwent a self-administered test, and, if he fared well, he was allowed to grapple with the next task, and so on from year to year until he secured his diploma. Because the plan sought to adjust the child's capacity to his rate of learning, and because it viewed learning as a continuous rather than a segmented process, failure to be promoted was nonexistent.

Different from the Dalton and Winnetka schemes was the Gary Plan. Instituted in 1908 by William Wirt at Gary, Indiana, it put its stress on stretching the taxpayer's dollars to effect an exemplary all-round education embracing work, study, and play. To this end Wirt conjured up a "complete school," comprising "playground, garden, workshop, social center, and traditional school under one management," and frequented by children from kindergarten through high school. Something of an efficiency wizard, Wirt grouped the students into several "platoons," striving thereby to effect a maximum use of the school's varied facilities, day and night, 12 months a year.

Progressive education

Besides Dalton and other pedagogic innovations, the 20th century saw the development of "progressive education," some of the principles of which were first proclaimed by the 18th-century French philosopher Jean-Jacques Rousseau and elaborated upon by the 19th-century figures J.H. Pestalozzi and Friedrich Froebel. What is currently called progressive education is essentially of American provenance, largely the pioneering work of Francis W. Parker and John Dewey. Parker laid its groundwork, first at Quincy, Massachusetts, then at Chicago, especially in the Cook County Normal School. Dewey, through his work in the Laboratory School of the University of Chicago, but more through his writing and teaching at Columbia University, prepared the philosophic seedbed from which much of the progressive movement sprouted and grew. In any case, the progressive believes, with Rousseau, that the heart of education is the child; that he has a nature all his own that demands respect and understanding; that interest is the motive of successful learning and without it learning is sterile; that young and old learn what they *live* or experience and translate what they learn into their behaviour; that the school should be a "place of good living" providing the interests and activities that the child's nature requires; and, finally, that the school should be a living society helping the child to participate effectively within a social group. Progressivism reached its apogee in the 1920s. Its support came chiefly from the elementary school and from theory-and-practice departments of teachers' colleges. Nevertheless, it never lacked opposition. It was censured especially for its "soft pedagogy" and the alleged downward slide of academic standards. Such charges were never proved, however, and meticulous studies have tended to discredit them.

Changes in higher education. The pedagogic experimentalism that marked America's elementary learning during the century's first quarter was less robust in the high school and feebler still in the college. The first venture of any consequence into collegiate progressivism was undertaken in 1921 at Antioch College, in Ohio, under its president, Arthur E. Morgan. Seeking to fuse a liberal college education, vocational training, and "apprenticeship for life," Antioch required a student to divide his time between his study of the traditional subjects and the extramural world, for which, every five weeks or so, he forsook the classroom to work at a full-time job. Not intended as vocational preparation or even a device to aid the student economically, the Antioch plan sought to amalgamate its liberal learning with "actual living in actual society."

Experiments at Antioch, Bennington, Chicago, and Harvard

While Antioch took the midpath between tradition and innovation, in 1932 Bennington College for women, in Vermont, strode boldly toward progressive ends. Putting a high value on student freedom, self-expression, and creative work, and less on intellectualism, Bennington essayed to weld its curricular and extracurricular segments into an undifferentiated entity. To this purpose it staffed its faculty largely with successful artists, writers, musicians, and other creative persons, rather than Ph.D.'s. And to the same purpose it granted students a large say in making the rules under which they lived.

Such developments in America's higher learning incited gusty blasts from Robert M. Hutchins, president and then chancellor of the University of Chicago from 1929 to 1951. He excoriated Bennington for its vocationalism and anti-intellectualism, its debasement of pure learning with courses in Ping-Pong teaching, cafeteria management, and the like. To arrest such deterioration, Hutchins recommended a mandatory study of grammar, rhetoric, logic, mathematics, and Aristotelian metaphysics. One consummation of the Hutchins prescription is the study of some 100 "great books," wherein reside the unalterable first principles that Hutchins insists are the same for all men always and everywhere. In the groves of higher learning, the Hutchins proposal bore no lasting fruit—save at Saint Johns College at Annapolis, Maryland. There, all courses are required, and for four years all students are made to ponder the First Principles in the Great Books.

The vocationalism that Hutchins deplored was taken to task by several others, but with quite different results—notably by Harvard in its report on *General Education in a Free Society* (1945). Declaring against the high school's heavy vocational leaning, it urged the adoption of a general curriculum in English, science, mathematics, and social science. At Harvard, moreover, the brakes were put on all specialization until the student had completed a two-year selection of courses in the humanities, natural science, and social science. Rather than offer a single "core" course, however, the faculty gave the student a choice among four or five offerings in each of these fields, each prepared by outstanding scholars especially for the instruction of nonspecialists. Optional advanced courses also were offered under the same program. In this way, it was felt, all students would gain a common appreciation of the Western heritage.

The junior college

The desire to enable a greater number of youth to savour learning beyond the high school propagated the junior college (sometimes called the community college). Conceived early in the century by William Rainey Harper, president of the University of Chicago, it proposed to separate the four-year college into an upper and lower half, the one designated as the "university college" and the other as the "academic college." The junior college is a two-year school, sometimes private but commonly public. In the latter event, its tuition toll is slight—often, indeed, it is almost nil. As a dispenser of higher learning, the junior college offers first- and second-year

college work, sometimes with vocational extras. In other cases it may be simply an advanced secondary school. Graduates from the former may continue as juniors in a four-year college or university.

Professional organizations. Americans, as is well-known, like to organize, and in this respect educators have been no different from the generality. They were indulging their proclivity as early as 1743, when the American Philosophical Society was founded, and they have been at it increasingly ever since. Not a few of their organizations, such as the American Historical Association, the Modern Language Association of America, and the American Home Economics Association, are for the advancement of some specialty. Others are more concerned with the interests of the general educational practitioner. Of these the National Education Association (NEA) is the oldest. Founded in 1857, it undertook "to elevate the character and advance the interest of the teaching profession." Despite its high mission, it threw off no sparks, and it was not until after 1870 that it began to grow and prosper. With headquarters in Washington, D.C., the NEA conducts its enormous enterprise through a brigade of commissions and councils, some of them quite large, as witness the Educational Policies Commission, the Legislative Commission for the Defense of Democracy through Education, the National Commission on Safety Education, the National Commission on Teacher Education and Professional Standards, the National Council of Education, and the National Council on Teacher Retirement.

A youngster by comparison, but growing rapidly, is the American Federation of Teachers, an affiliate of the AFL-CIO. Formed in 1916, the federation concerns itself with bettering the teacher's lot. Through collective bargaining and teachers' strikes, it has worked successfully for better wages, pensions, sick leaves, academic freedom, and other benefits.

Although teaching in America is spoken of as a profession, it is only in the recent past that some of its practitioners, like the doctors and lawyers, have undertaken to bring order and dignity to their occupation. Toward such an achievement such bodies as the American Association of Colleges for Teacher Education, the American Association of University Professors, and particularly the National Commission on Teacher Education and Professional Standards (TEPS) and the National Council for Accreditation of Teacher Education (NCATE) have laboured industriously and even with a fair success. (A.E.M.)

Elder members of the British Commonwealth. *Canada.* Although a Canadian nation had been formed by the end of the 19th century, separate political, economic, and geographic influences continued through the 20th century to restrain unified educational development. The historical principle of maintaining minority rights resulted in a truly pluralistic cultural concept, recognized to some extent in religious and linguistic concessions in schools. Each provincial system developed unilaterally, thus producing separately centralized educational units; and, even within a province, the evolving principle of local responsibility and the sparseness of settlement in many areas of Canada challenged the effectiveness of simple control principles. Different production emphases and differential advantages of territorial acquisition after confederation in 1867 had the effect of dividing Canada into "rich" and "poor" provinces, which can support schools accordingly. Finally, the European basis of education was never reconciled philosophically with that evolving out of the North American environment, with the consequence that educational actions were often taken without regard for their integration in a Canadian foundation.

These divisions are characteristic of Canadian education in the 20th century. It is nevertheless possible, because of a common national experience and because of the communication stimulated by national development, to describe education in national terms. Educational movements afoot in the early 20th century and associated with "progressive education" (such as child study, kindergarten development, and curriculum integration) had a relatively mild impact on traditional practices and forms. Instruction in the Canadian school remained essentially teacher centred, with a strong emphasis on obedience and conformity, and based on an educational version of philosophical Realism. Only changes that corresponded to established Canadian convictions, such as the acceptance of some activity principle in methods, were incorporated into the conduct of education, and the antagonistic reaction to such changes was disproportionate to the moderation with which reform was approached.

Traditional versus progressive education

The major change in school structure occurring during the 20th century was at the secondary level. The standard eight-year elementary program was first extended by continuation classes or schools alongside exclusive secondary schools, producing an uneven, overlapping postelementary structure. In the 1930s, an expanded school population, reaching into the secondary grades, led to decisive action on compulsory attendance and to standardization of high school provisions. Junior high schools were introduced in some provinces as a transitional level between elementary and secondary schooling, while some provinces simply developed junior and senior stages of a total secondary program. The two extremes in secondary development were probably represented by Québec and the west. In the French-speaking schools of Québec, the secondary system consisted of private classical colleges leading to a *baccalauréat* on the one side and terminal courses in special schools or institutes on the other. Only after 1956 were public high schools with a variety of courses established, and only from 1964 has there been a move to unify the entire provincial system. In the western provinces, large regional schools and composite high schools have been developed extensively, Alberta having proceeded apace in this direction. British Columbia, following the Chant Commission Report in 1960, reorganized its secondary program to include five core streams, only one of which is academic–technical.

From the attempt to broaden the curriculum in response to the requirements of an expanded school population in the second quarter of the century, regular revisions have been made in curricula and methods issued in Canadian schools. Most revision efforts have originated with developments in the United States, though they have taken a particular form in Canadian education. Present trends are common in North America and include providing flexibility in teaching assignment (for instance, team teaching), in school plant (such as "open area" schools), and in program (such as "continuous progress" principles). Both the attempt to reconcile individual educational requirements with the demands of mass systems and the current emphasis on the essential structure of subject matter lead to a search for new techniques of selecting and transmitting knowledge in schools.

Changes in the elementary program have been largely in the area of teaching methods, and this applies generally to the secondary schools also, although the secondary curriculum has been modified by some additions, by expanding the catalog of optional subjects, and by reorganizing to include new courses of study. Although the basic division remains one between the university-preparation course and a general terminal course, more alternatives are beginning to appear.

Regional pattern of Canadian universities

A new dimension in higher education was added with the establishment of provincial universities in the west (1901–08). This completed a set of regional patterns for university development that have continued to this day. Canadian universities have, within these patterns, drawn their criteria from French, British, or American models. From the 1950s, a boom in Canadian higher education has led to increasingly independent considerations on the role of universities in Canadian development, possibly the most visible result being the attention now given to the rationalization of higher education within the provincial system.

The administration of public education is the exclusive responsibility of the provinces, which have worked out schemes of local authority under provincial oversight. Although the specific structure of the departments of edu-

cation varies among the provinces, they conform to a basic structure. Each is headed by a politically appointed minister of education, who may be advised by a council. The main functions of educational supervision are now usually carried out through specific directorates for such areas as curriculum, examinations, vocational education, teacher training and certification, and adult education. Three developments, however, have strengthened local autonomy in educational administration. Throughout the second quarter of the 20th century, consolidation of rural schools and administrative units took place in the west, thus resulting in stronger educational units more competent to act independently. More recently, especially in Ontario, Québec, and New Brunswick, moves toward regional decentralization produced rather independent subprovincial units, synthesizing the advantages of local and central control. Finally, rapid urban development in Canada has led to relatively autonomous city school operations.

Federal influences in Canadian education

Canada's federal government has no constitutional authority in education and therefore maintains no general office dealing directly with educational matters. Federal activities in education are nevertheless carried out under other areas of responsibility, and certain functions of an office of education are subsumed under the secretary of state. Under its responsibility for native peoples and its jurisdiction over extraprovincial territories, the federal government, through the Department of Indian Affairs and Northern Development, finances and supervises the education of Indians and Eskimo. In the Yukon, schools are administered by the territorial government, though largely financed from Ottawa.

Through agricultural and technical assistance acts in 1913 and 1919, the federal government began to promote vocational education, and this principle was extended through emergency programs in the depression years of the 1930s and during World War II. More recently, vocational programs of wide scope have been introduced on a principle of federal support and provincial operation. The Technical and Vocational Assistance Act of 1960 was followed by a great surge in vocational education, including the construction of new schools and school additions, special institutes, and the preparation of vocational teachers. Program definitions in this area have become ever broader, and new federal policy is likely to be aimed at a broad base of support at the postsecondary level.

The federal government has maintained and supported the education of armed-forces personnel, including the Royal Military College of Canada and cadet or regular officer's training programs in schools and universities, educational aid to postwar veterans, and the operation of schools for children of overseas personnel. Research and development in higher education are promoted directly through grants and through the offices of the National Research Council and the Canada Council. Of service to educational research is the Education Division of the Dominion Bureau of Statistics, which disseminates organized statistical information on Canadian schools. Perhaps less direct but of great importance are national agencies operating in the area of mass communications media, such as the National Advisory Council on School Broadcasting and the National Film Board. Together, the activities of the federal government not only support but also strongly influence certain areas of education and complete a picture of local–provincial–federal involvement in Canadian education.

Australia. The 20th-century development of Australian education continued to be influenced by British models and to be characterized by the exercise of strong central authority in the states. Yet, because Australian national development has taken place entirely in this century, increasing attention has been given to the role of education in nation building.

Educational systems were built through the establishment of primary schools by the end of the 19th century, the extension of these through continuation programs, and the development of state secondary schools in the early part of the 20th century. The independent secondary schools that offered the bulk of secondary education before 1900 continued to be influential, either as components of the separate Roman Catholic system or as "elite" private schools of denominational or nondenominational character, but the growth of state systems carried the state high schools into numerical prominence.

The early development of educational systems before and around the turn of the century was a crude beginning, the minimal provisions being accentuated by poor teacher preparation, administrative thrift schemes, and excess in the exercise of administrative authority. Improvement of these conditions and systematic positive development can be dated from the Fink Report of 1898 in Victoria and similar reform appeals in other states between 1902 and 1909. The steady pace of progress from that time was broken by a surge of growth and innovation in Australian institutions after World War II.

Education in small, isolated communities throughout the vast Australian area has required special attention. Two particular means of reaching isolated children and adults have been used successfully. Correspondence education was begun in 1914 in Victoria, and other states followed after 1922. The procedures have been gradually refined and the levels extended. The correspondence system embraces primary, secondary, technical, and university courses and is supplemented by extensive educational broadcasting by radio. The other, more formal means of reaching rural children has been through efforts devoted to the establishment of schools in the outback.

Early efforts included provisional schools, itinerant teachers, and central schools. The small one-teacher bush schools became typical after federation in 1901 and still make up a large proportion of Australian state primary schools. Much attention has been given to methods of teaching in the one-room school, earning Australia international recognition for expertise in this area. Progress toward rural school consolidation began in Tasmania in 1936. The Tasmanian model has combined special features of school independence, pupil freedom, involvement in agricultural projects, and parental cooperation with the "area school" movement.

State and federal powers in Australia

Education is a state, rather than a federal, responsibility in Australia. Authority is concentrated in a state department of education, and control is exercised through the standard of department policy and through prerogatives distributed in a hierarchical organization of educational experts. Some decentralization has been accomplished in New South Wales and Queensland, but the administrative areas so established are coordinated with the central department rather than with other local units, thus preserving the central principle. The political head is the minister of education, and the permanent official in charge is the director or director general of education. The main divisions of the department are normally those for primary, secondary, and technical education, each directed by a senior official; additional divisions, such as special education, guidance, and in-service training, are particular to the states.

Since World War II, with the financial assets of exclusive income-taxing power, the commonwealth (federal) government has played an increasing role in educational development, particularly at the tertiary level. Through the States Grants Act in 1951, the Murray Report in 1957, the Martin Report in 1964, and the establishment of the Australian Universities Commission, the Commonwealth Scholarships Board, and the Commonwealth Advisory Committee on Advanced Education, the federal government has moved into the planning as well as the funding of postsecondary education concurrently with the states. The Commonwealth Office of Education, established in 1945, advises on financial assistance to the states and on educational matters generally, acts as a liaison agent among the states and between Australia and other countries, undertakes some educational research, and provides educational information and statistics.

About three-quarters of Australian schools are public. The remainder are made up of Roman Catholic schools (which constitute more than 80 percent of the nonpublic schools) and other private schools, many of which have

considerable influence in the leadership of Australian society. The curriculum and syllabus for each program or course in the state schools is prescribed by the Department of Education, and nonpublic schools generally follow this standard.

Primary schools are normally of six years' duration, to about age 12, though some schools retain the seventh, and even eighth year of the old pattern. Within primary schools, pupils are organized in grades and advance by annual promotion. By selection or choice, pupils in the past have proceeded to academic high schools or to an alternative form (mainly technical) of secondary education. The trend now, however, is to comprehensive schools—schools offering academic, commercial, and technical programs, either separately tracked or somehow unified with gradual differentiation from lower to the highel levels.

New Zealand. The religious and regional issues that have fettered educational development in other countries of the British Commonwealth were basically settled in New Zealand when the decisions were made in the last quarter of the 19th century to provide wholly secular primary schools and administrative centralization. The major issue in the 20th century has been the achievement of equal educational opportunity. Although New Zealand has accepted the responsibility to educate each child—without racial, social, or narrowly intellectual restriction—to the limits of his ability, the unification of the total system to this end has proved quite difficult.

New Zealand's goal of equal educational opportunity

The Education Act of 1914 consolidated the changes that had taken place since 1877. In subsequent reform periods, especially during the '30s and after World War II, barriers to pupil progress through the system were removed or modified. In 1934, the school-leaving certificate examination was established on a broader basis than the university entrance examination, and, in 1936, the proficiency examination governing secondary entrance was abolished. In 1944–45, the school-leaving age was raised to 15; a common core of early secondary studies, including English, social studies, general science, mathematics, physical education, and a craft or fine-arts subject, was established; and universities agreed to accept accredited-school courses without further examination for university entrance. These actions illustrate a gradual but steady facilitation of access through an increasingly coordinated system. This direction is carried into the third 20th-century reform period of the 1960s in the recommendations of the Currie Commission (1962) and the provisions of the Education Act of 1964.

Since 1877, the administration of education has been conducted by a central Department of Education and by elected regional boards of education. Virtually all funds come from the central source, and education is free until the age of 19 for qualified pupils. University tuition is also paid for successful students. The Department of Education is headed politically by a minister and permanently by a director general. The inspectorate is divided into primary and secondary units. In addition to these two basic divisions, the department contains offices for child welfare, technical education, higher education, Maori education, and education in the island territories.

New Zealand children generally start school at the age of five years and spend eight years in primary school. The secondary system has developed through the growth of three separate kinds of school: (1) the district high school, which represented more or less a secondary "top" on a primary school; (2) the independent, academic, one-sex secondary school proper; and (3) the technical school, which took shape between 1900 and 1908. These three types have remained, but some of their uniqueness has been lost through certain refinements. The isolated position of the fee-charging secondary schools of the 19th century was compromised by free-place legislation in 1903, and, by 1914, they were brought into the state system, though retaining a good deal of their independent status. The district high schools remained in the primary system, but their incorporation in the secondary inspection scheme and in secondary teacher classification placed them clearly within that sector of school operation. The technical high school evolved into a general high school with technical bias. Through common departmental inspection, curriculum, and examination standards, and through the effect of the movement for more general postprimary provisions after 1945, the secondary schools have increasingly approximated a single pattern.

Private and native schools

Although an extensive Roman Catholic private school system grew up after the secularization of state education, these schools are also integrated into the general system through adherence to state standards, verified by government inspection. Rural and native education have also been given careful attention in New Zealand. Pains are taken to ensure that rural schools are provided with adequate facilities and teachers. Where possible, consolidated schools have been constructed, served by an extensive school transportation system. Isolated districts are reached through the Department of Education correspondence school and through radio instruction. Maori education has been a responsibility of the Department of Education since 1879. Both Maori and European children attend the general public schools, but Maori schools, stressing oral English and Maori culture, are also present in the villages. The government has, however, been reexamining Maori education since the Currie Report in 1962, giving attention to particular areas of inadequacy in the existing provision of schooling. (R.F.L.)

REVOLUTIONARY PATTERNS OF EDUCATION

Russia: from tsarism to Communism. *Before 1917.* At the turn of the 20th century, the Russian Empire was in some respects educationally backward. According to the census of 1897, only 24 percent of the population above the age of nine were literate; by 1914 the rate had risen to approximately 40 percent. The large quota of illiteracy reflected the fact that, by this time, only about half the children between eight and 12 attended school. The elementary schools were maintained by the *zemstvo* (local government agencies), the Orthodox Church, or the state, the secondary schools mainly by the Ministry of Education.

After the Revolution of 1905 the Duma (parliament) made considerable efforts to introduce compulsory elementary schooling; it was hoped to accomplish this by 1922. At the upper stages of the educational system, progress was significant, too; nevertheless, the secondary schools (*gimnazii, realnyye uchilishcha*) were only to a small degree attended by students of the lower classes, and the higher institutions even less. Preschool education as well as adult education was left to the private initiative of the educationally minded intelligentsia, who were opposed to the authoritarian character of state education in the schools. In 1915–16 the minister of education, Count P.N. Ignatev, started serious reforms to modernize the secondary schools and to establish a system of vocational and technical education, which he regarded as most important for the industrialization of Russia. During the Provisional government (February to October 1917, old style), the universities were granted autonomy, and the non-Russian nationalities received the right of instruction in their native languages. The education system envisaged by the liberal-democratic and moderate Socialist parties was a state common school for all children based on local control and the direct participation of society.

Reform before the Revolution

1917–30. After the October Revolution of 1917, the Bolshevik Party proclaimed a radical transformation of education. Guided by the principles of Marx and influenced by the contemporary movement of progressive education in the West as well as in Russia itself, the party and its educational leaders, N.K. Krupskaya and A.V. Lunacharsky, tried to realize the following revolutionary measures as laid down in the party's program of 1919: (1) the introduction of free and compulsory general and polytechnical education up to the age of 17 within the Unified Labour School (Yedinaya Trudovaya Shkola), (2) the establishment of a system of preschool education to assist in the emancipation of women, (3) the opening of the universities and other higher institutions to the working people, (4) the expansion of professional education for persons from the age of 17, and (5) the creation of a

Revolutionary experimentalism

system of mass adult education combined with the propaganda of Communist ideas. In 1918 the Soviet government had ordered by decree the separation of church and school—that is, the abolition of religious instruction in favour of atheistic indoctrination, the coeducation of both sexes in all schools, the self-government of students in schools and universities, the abolition of marks and examinations, and the introduction of productive labour. In 1919 special workers' faculties (*rabfaki*) were created at higher institutions and universities for the development of a new intelligentsia of proletarian descent.

Lenin paid special attention to the campaign against illiteracy and to a widespread technical education; both seemed to him to be prerequisites for the establishment of an industrial Communist society in Russia. During the period of the New Economic Policy (1921–27), when there was a partial return to capitalistic methods, the revolutionary spirit somewhat diminished, and the educational policy of party and state concentrated on the practical problems of elementary schooling, the struggle against juvenile delinquency, and the schooling of adult illiterates. When the policy of five-year plans began in 1928 under the slogan of "offensive on the cultural front" and with the help of the Komsomol (the Communist youth league), the campaign against illiteracy and for compulsory elementary schooling reached its climax. At the same time the revolutionary experimental spirit regained its strength; men such as P.P. Blonsky, S.T. Shatsky, V.N. Shulgin, and A.P. Pinkevich, who can be regarded as representatives of the first period of Soviet education, attempted to establish a new Marxist educational science. It was a kind of combination of biology and sociology from a materialistic point of view, which tried to maintain the rights of the individual child within the Socialist community. This idea found its expression in the creation of a special branch of educational science, pedology (*pedologiya*), which continued to exist until 1936, when it was banned by a special party decree.

The Stalinist years, 1931–53. In connection with the policy of rapid industrialization and collectivization of farmers and with the concentration of political power in the hands of Stalin, the Soviet educational policy in the 1930s experienced remarkable changes. Starting with the decree of 1931, the structure and the contents of school education underwent in the next few years the following process of "stabilization": (1) four years was laid down as the compulsory minimum of schooling for the rural districts, and seven years for the cities; (2) the new system of general education embraced the grades one to four (*nachalnaya shkola*), the grades five to seven, which continued the elementary stage on the lower secondary level (*nepolnaya srednyaya shkola*), and the grades eight to ten, which provided a full secondary education (*polnaya srednyaya shkola*); (3) the new curriculum was to provide the students with a firm knowledge of the basic academic subjects and was to be controlled by a system of marks and examinations; (4) the decisive role of the teacher within the educational process was re-established, while the Pioneers and Komsomol organizations (for youth aged ten to 15 years and 14 to 26 years, respectively) were above all to instill a sense of discipline and an eagerness for learning; (5) manual work disappeared from the school curriculum as well as from the teacher-training institutions. These changes, carried out under A. Bubnov, the new people's commissar for education (1929–37), were accompanied by the rejection of the ideas of progressive education and the beginning of the cultivation of older Russian traditions. During World War II, furthermore, the idea of Soviet patriotism emerged fully, penetrating the theory and practice of education. The principles of the outstanding educator A.S. Makarenko, with their emphasis on collectivism, gained ground upon the former influence of Western educational thought.

Return to authoritarianism and traditionalism

The institutions of higher learning were reshaped in the 1930s, too. The number of students in institutions providing secondary specialized education, usually called *tekhnikumy*, rapidly grew from 1,000,000 in 1927–28 to 3,800,000 in 1940–41. The number of students in institutions of higher education (*vyssheye uchebnoye zavedeniye*) grew from 168,554 to 811,700 in the same period. In 1932 a central committee for technical higher education was created, and in 1936 its functions were extended to the universities and other institutions. The main characteristics of higher education that developed in this period remained unchanged for the next decades: the paramount task of higher learning was to provide specialized professional education within the framework of manpower policy and economic plans; strict control of the student's program was to be imposed by the central authorities; and the system of evening and correspondence instruction on the level of higher and secondary specialized education (*vecherneye i zaochnoye obrazovaniye*) was to parallel full-time studies.

A rather unexpected result of the rise of academic standards in higher education was the decline in the percentage of students from the working class and the peasantry. Between 1940 and 1956, students in the upper grades of secondary schools and in higher and secondary specialized establishments were required to pay tuition fees, though, as before, a great many received government stipends.

During the 1940s, "labour reserve" trade schools and factory schools for skilled and semiskilled labour were filled by drafting youth between the ages of 14 and 17. In the period 1940 to 1958, an average of 570,000 persons were annually subjected to such recruitment. The draft first affected those students who were unsuccessful academically in regular secondary schools and could not achieve even the seventh grade. For youngsters of this kind and for people who could not continue general secondary education, schools for the working youth (*shkoly rabochey molodyozhi*) and schools for rural youth (*shkoly selskoy molodyozhi*) were established in 1943–44 as part-time institutions. The main features of education policy, developed in the late 1930s, remained in force after the war: the orientation of all kinds of schooling and training to the paramount necessities of the economic system; the inculcation of Communist discipline and Soviet patriotic attitudes; and finally a rigid control of the whole educational system by party and state administration.

Post-Stalin era. Since the death of Stalin in 1953, it has been possible to observe a few changes in official policy that have affected both education and science. The 20th Party Congress in 1956 paved the way for a period of reforms inaugurated by N.S. Khrushchev. The central idea was formulated as "strengthening ties between school and life" at all levels of the educational system. The basic documents of the educational reform were the theses of the Central Committee of the Communist Party and the Soviet Government (November 1958) and the All-Union Law of December 24, 1958, followed by separate laws in each of the 15 union republics in 1959. The educational reform in the Soviet Union influenced to a high degree similar reforms in the eastern European countries.

Practicalism and polytechnization

The key issue of the reform debate in 1958 that preceded the decrees was the question of labour training in the secondary schools and of the best way to achieve it. The old idea of polytechnical education has been revived, but mainly in the sense of preparing secondary school students for specialized vocational work in industry or agriculture. Since the early 1950s there has been a growing imbalance between the output of secondary school graduates desiring higher education and the economic demands of skilled manpower at different levels. The educational reforms of 1958 pursued the aim of combining general and polytechnical education with vocational training in a way that directed the bulk of young people after the age of 15 straight into "production."

The new structure of the school system after 1958 developed as follows: (1) the basic school with compulsory education is the eight-year general and polytechnical labour school, for ages seven to 15 (*vosmiletnyaya shkola*); (2) secondary education beyond the eight-year school is provided alternatively by secondary general and polytechnical labour schools with production training (*sred-*

nyaya obshcheobrazovatelnaya trudovaya politekhnicheskaya shkola s proizvodstvennym obucheniem) or by evening or alternating-shift secondary general education schools (*vechernyaya smennaya srednyaya obshcheobrazovatelnaya shkola*), the former school for working or rural youth. Both types were three-year schools, embracing grades nine to 11. The duration of schooling until the final secondary examination was thus prolonged for one year, one-third of the curriculum being used for productive labour and vocational training.

For a transition period the admission to institutions of higher education was to be made dependent on two years of practical experience after the former ten-year secondary school had been completed. Between 1960 and 1962, however, only about 60 percent of those admitted to full-time higher education had actually had any such practice. The connection of study and productive work was to be continued during the course of higher education. Great emphasis was laid upon the further expansion of evening and correspondence education on both levels—at the level of secondary specialized education and at the level of the universities and other higher institutes. In the academic year 1967–68, 56.3 percent of all Soviet students in higher education (of the total number of 4,311,000) carried out their studies in this way. The establishment of five higher technical education institutes in factories (*zavody-vtuzy*) in 1960 was intended to train technological engineers on the actual premises of large industrial enterprises.

In the system of vocational-technical education, the reform of 1958 brought a transformation of the former labour-reserve schools into urban vocational–technical schools or rural schools of the same type (*gorodskiye i selskye professionalnyye tekhnicheskiye uchilishcha*). As a rule they require the completion of the eight-year school, but in fact there are many pupils with lower achievements; the length of training is from one to three years, depending upon the type of career.

Besides introducing polytechnical education and productive labour, the Khrushchev reforms emphasized the idea of collective education from early childhood. Preschool education for the age group up to seven years was to be rapidly developed within the newly organized unified crèches and nursery schools (*yasli i detskiye sady*); and, as a new type of education, boarding schools (*shkoly-internaty*) had been created in 1956 that embraced grades one to eight or one to 11. Some party circles wanted this kind of boarding education for the majority of all young people, but development lagged behind planning, and the idea of full boarding education has since been abandoned. Nevertheless, there were schools and groups with an "extended" school day (*shkoly i gruppy prodlennogo dnya*), and these and the boarding schools accounted for the education of some 4,500,000 children in 1969.

Qualitative reforms

The polytechnization of the Soviet school system as it took shape during the Khrushchev period turned out, in the course of its realization, to be a failure. A revision of the school reform was carried out between August 1964 and November 1966 that brought about several important results: (1) the grade 11 of the secondary school (except for the evening school) was abolished; general education returned to the ten-year program; (2) vocational training in the upper grades was retained only in a small number of well-equipped secondary schools; (3) a new curriculum and new syllabi for all subjects were elaborated. In 1966 a central Ministry of Education was established in Moscow to coordinate and guide the activities of the administrative bodies in each of the republics (see EDUCATION, SYSTEMS OF). The aim of educational policy as outlined at the 23rd Party Congress in 1966 is to achieve universal full-time education of at least ten years in the early 1970s. About 75 percent should receive it within the normal secondary school, the remainder either in specialized secondary schools or in three- or four-year establishments that combine vocational training and academic instruction (decree of April 1969). In accordance with the needs of technical modernization of the Soviet economy, the main task of education and training on the secondary level is to provide a firm knowledge in the basic subjects and to equip the pupils with motivation for further learning. Since 1958 some hundreds of secondary schools for gifted pupils in mathematics, science, or foreign languages have been developed, besides the well-known special schools for music and arts.

The educational sciences are mainly concentrated in the Academy of Pedagogical Sciences in Moscow (established in 1943). The academy maintains 12 research institutes and takes part in educational policy making. In each republic there is a special research institute of education. In the field of higher education the new statute of 1969 extended to some degree the responsibility of the individual institution and the academic staff in economic affairs and the planning of curricula. After the period of numerical expansion of higher education that made the Soviet Union rank second only to the United States, the improvement of quality in teaching and organization has top priority. The ideal of Soviet education is a learned society within the framework of Communism and in connection with the historical traditions of Russia. (O.A.)

China: from Confucianism to Communism. *The modernization movement.* The political and cultural decline of the Manchu dynasty was already evident before the 19th century, when mounting popular discontent crystallized into open revolts; and the dynasty's weakness was starkly exposed by its inability to cope with the aggressive Western powers during the 19th century. At the time when Western nations were closing in on China, the Manchus were harassed by a series of internal rebellions, the best known of which was the Taiping Rebellion (1850–64). After the military defeats administered by the Western powers, even Chinese leaders who were not in favour of overthrowing the Manchus became convinced that change and reform were necessary.

Most of the proposals for reform provided for changes in the educational system. New schools began to appear. Missionary schools led the way in the introduction of the "new learning," teaching foreign languages and knowledge about foreign countries. New schools established by the government fell under two categories: foreign-language schools to produce interpreters and translators and schools for military defense. Notable among the latter were the Foochow Navy Yard School to teach shipbuilding and navigation and a number of academies to teach naval and military sciences and tactics.

China's defeat by Japan in 1894–95 gave impetus to the reform movement. A young progressive-minded emperor, Kuang-Hsü, who was accessible to liberal reformers, decided upon a fairly comprehensive program of reform, including reorganizing the army and navy, broadening the civil-service examinations, replacing the rigid form of the old style of writing with essays on current affairs, establishing an Imperial University in the national capital and modern schools in the provinces, creating a translation bureau, and so on. The imperial edicts in the summer of 1898 spelled out a program that historians have called the Hundred Days of Reform. Unfortunately for China and for the Manchu dynasty, it met with strong opposition from the conservatives, who appealed to the empress dowager Tz'u-hsi for support. The ultraconservative Empress Dowager took prompt and peremptory action, putting the young Emperor in detention and assuming full imperial power herself to stop the reform movement. The edicts of the summer were reversed and the reforms nullified. Frustration and disappointment in the country led in 1900 to the emotional outburst of the Boxer Rebellion.

Hundred Days of Reform and the aftermath

After the Boxer settlement, which exposed China's complete helplessness before foreign powers, even the Empress Dowager saw no way out but to accept the necessity of change. Belatedly, she now ordered the very changes that she had opposed during the Hundred Days of Reform. Modern schools were to be established on all levels, teaching modern subjects such as Western history, politics, science, and technology, along with Chinese classics. The civil-service examinations were to be broadened to include Western subjects. A definite plan was ordered to send students abroad for study and recruit them for

government service upon return from abroad. But these ameliorative measures were not enough to meet the pressing demands now being presented with increasing forcefulness. Finally, an edict in 1905 abolished the examination system that had dominated Chinese education for centuries. The way was now cleared for the establishment of a modern school system.

Changing forms of the modern Chinese school system

The first modern school system was adopted in 1903. The system followed the pattern of the Japanese schools, which in turn had borrowed from Germany. Later, however, after establishment of the republic, Chinese leaders felt that the Prussian-style Japanese education could no longer satisfy the aspirations of the republican era, and they turned to American schools for a model. A new system adopted in 1911 was similar to what was then in vogue in the United States. It provided for an eight-year elementary school, a four-year secondary school, and a four-year college. Another revision was made in 1922, which again reflected American influence. Elementary education was reduced to six years, and secondary education was divided into two levels. A three-year junior middle school was in line with the junior high school movement then gaining acceptance in the United States. This six–three–three–four system continued until recent years.

Education in the republic. The first decade of the republic, up to the 1920s, was marked by high hopes and lofty aspirations that remained unfulfilled in the inclement climate of political weakness, uncertainty, and turmoil. The change from a monarchy to a republic was too radical and too sudden for a nation unschooled in political participation. The young republic was torn by political intrigue, by unscrupulous power seekers, and by internecine warfare among self-styled military chieftains known as warlords. There was no stable government, no effective leadership to make plans for the nation.

A school system was in existence, but it received scant attention or support from those responsible for government. School buildings were in disrepair, libraries and laboratory equipment were neglected, and teachers' salaries were pitifully low and usually in arrears.

New intellectual movements during the republic

It was, nevertheless, a period of intellectual ferment. The intellectual energies were channelled into a few movements of great significance. The first was the New Culture Movement, or what some Western writers have called the Chinese Renaissance. It was at once a cordial reception to new ideas from abroad and a bold attempt to reappraise China's cultural heritage in the light of modern knowledge and modern scholarship. China's intellectuals opened their minds and hearts to ideas and systems of thought from all parts of the world. Translated works of Western educators, philosophers, and literary writers were eagerly read by students and educated people. There was a mushroom growth of journals, school publications, literary magazines, and periodicals expounding new ideas. It was at this time that Marxism was introduced into China.

Another movement of great significance was the Literary Revolution. Its most important aspect was a rebellion against the classical style of writing and the advocacy of a vernacular written language. The classics, textbooks, and other respectable writings had been in the classical written language, which, though using the same written characters, was so different from the spoken language that a pupil could learn to read without understanding the meaning of the words. Now, progressive scholars rejected the heretofore respected classical writing and declared their determination to write as they spoke. The new vernacular writing, known as *pai-hua* ("plain speech"), won immediate popularity and facilitated the expression of ideas and views in a form of writing readily understandable to the many, rather than merely to the scholarly few. Breaking away from the limitations of stilted language and belaboured forms, the *pai-hua* movement was a big boon to the freedom and creativity released by the New Thought Movement and produced a new literature attuned to the realities of contemporary life.

A third movement growing out of the intellectual freedom of this period was the Chinese Student Movement, or what is known as the May Fourth Movement. The name of the movement rose from nationwide student demonstrations on May 4, 1919, in protest against the decision of the Paris Peace Conference to accede to the Japanese demand for territorial and economic advantages in China. So forceful were the student protests and such overwhelming support did they get from the public that the weak and inept government was emboldened to take a stand and refused to sign the Versailles Treaty. The students thus had a direct hand in changing the course of history at a crucial time. The May Fourth Movement was in effect the culmination of a growing political and social awakening of Chinese students as a result of the new freedom and new vistas brought within their reach by the New Thought Movement and its offshoots. Chinese youth was rebelling against the kind of education that concerned itself with books and academic knowledge detached from the challenging problems of contemporary life. For better or for worse, the old concept of the scholar was now shattered, and from now on Chinese students constituted an active force on the political and social scene.

Education under the Nationalist government. Nationalist China rose in the mid-'20s amid a resurgence of nationalism and national consciousness stimulated by post-World War I developments. It was led by the Kuomintang, the political party originally organized by Sun Yat-sen, the founder of the republic. Cognizant of the popular appeal of nationalism, the Kuomintang set up a Nationalist government pledged to achieve national unity at home and national independence from foreign control as prerequisites to a program of modernization and national reconstruction. In education, it set out to systematize and stabilize a shaky and ill-supported school system and use it as a means of national regeneration. Schools were assured of financial support, however inadequate, and placed under strict supervision and firm control by public authorities.

State control and centralization of education

State control of education by means of centralized administration was instituted. Measures were adopted to correct the abuses and chaos that had resulted from the laissez-faire educational policy of the warlords. Decrees and regulations issued by the Nationalist Ministry of Education were strictly enforced, with the aid of a centrally administered system of inspection and accreditation. Detailed regulations covered the curricula of all types of schools on all levels, minimum standards of achievement, teaching procedures, teachers' qualifications, and specifications for school buildings, libraries, laboratories, and the like. Private education was permitted, even encouraged, by means of supporting grants from public funds, but private schools were subject to government control as much as public schools and were required to follow the same prescriptions with regard to curriculum and all other details.

A uniform system of schools was in effect throughout the country. Elementary education was provided in the four-year primary school, followed by the two-year higher elementary. In areas where there were not enough funds to support longer courses, there were abbreviated schools having only one or two grades. Theoretically, the government was committed to the goal of four-year compulsory education, but financial problems prevented an early realization of this goal. Adult education was given much attention in adult schools, in mass education projects, and in different forms of "social education." The latter term encompassed a variety of educational agencies outside the schools, such as libraries, museums, public reading rooms, recreational centres, music, sports, radio broadcasting, and films. Reduction of illiteracy was a major objective.

There were three parallel types of secondary education: the academic middle school, the normal school, and the vocational school. To counteract the traditional preference for the academic type of education, the government restricted the growth of the academic middle school. At the same time, vocational schools were encouraged as an alternative type of secondary education.

Nationalist promotion of "practical studies"

A major objective of government policy was to promote "practical studies." In secondary education, "practical studies" meant the development of vocational and technical schools and more attention to science and laboratory experience in middle schools. In higher education, measures were taken to steer students away from liberal arts, law, education, and commerce to the "practical courses" of science, engineering, technology, agriculture, and medicine. Government grants for private as well as public colleges were usually designated for the science program. As a result of this policy, the years prior to World War II saw a steady increase of enrollment in the "practical courses" of study and a corresponding decline of enrollment in the arts–law–education–commerce courses. The increase of interest in science was also evident in the secondary schools.

It may be said that the thrust of educational policy in Nationalist China was to rectify the imbalance of the past, especially the nonvocational literary tradition of premodern days and the laissez-faire policy of the warlord's regime. In the attempt to counteract past tendencies, however, it was possible that the pendulum might be swinging too far to the other extreme. Some educators expressed the fear that the promotion of "practical studies" might lead to a narrow, utilitarian concept of education and a neglect of the humanities and social sciences. Others were uneasy over the danger of regimentation through centralized administration. Nevertheless, education under the Nationalist government did succeed in establishing an effective national system of education, promoting science and technical studies, and correcting the abuses and irregularities of the earlier period. Thanks to dependable financial support, state schools and universities gained in prestige and academic performance until they were recognized as among the outstanding educational institutions of the country.

Among other accomplishments of this period may be mentioned the growth of postgraduate education and research, the general acceptance of coeducation in elementary and higher education, and the teaching of the Kuo yü (National Tongue) as an effective means of unifying the spoken language and thus overcoming the difficulties of local dialects.

Education under Communism. The Communist revolution is a total revolution, demanding no less than the establishing of a new society radically different from what it calls the feudal society of traditional China and from the bourgeois society of the West. This new society calls for people with a new outlook on life, new loyalties and new motivations, and new concepts of individual and group life. The new society cannot be established without new people, and the creation of the "new Socialist man" (Maoist style) must be a task of the highest priority. It is the task of education.

Education of the new Socialist man

The People's Republic of China makes no distinction between education and propaganda or indoctrination. All three share the common task of changing man. The agencies of education, indoctrination, and propaganda are legion—newspapers, posters, and propaganda leaflets, radio broadcasting, the drama, neighbourhood gatherings for the study of current events, as well as political rallies, parades, demonstrations, and many forms of "mass campaigns" under careful direction. It is evident that the schools constitute only a small part of the educational program. During the Cultural Revolution of 1966–69, schools and universities were closed, but the Communists could well argue that those were the years when an intensive educational program was carried on to combat bourgeois and revisionist tendencies and to hasten the appearance of the "new Maoist man."

When the Communists came to power in 1949, they took up three educational tasks of major importance: (1) teaching many illiterate people to read and write; (2) producing the trained personnel needed to carry on the work of political organization, agricultural and industrial production, and economic reform; and (3) remolding the behaviour, emotions, attitudes, and outlook of the people. The literacy campaign makes use of the radio and special classes organized in factories and other production enterprises, as well as adult schools of various types. The training of farm managers, factory foremen, bookkeepers for communes and cooperatives, technicians, and cadres of all kinds is done by short-term courses, institutes, or special classes in schools or elsewhere.

The process of remolding goes on everywhere, in schools and outside. Factories, government offices, industrial organizations all require their personnel to engage in regular "study," which is the term used for political and ideological indoctrination. Besides these "study" sessions, there are numerous group meetings in which individuals are supposed to help one another in "thought reform"; in such meetings each person is expected to lay bare his thoughts and feelings to be examined by his peers under the guidance and supervision of authorized representatives of the state and Communist Party. The technique of "thought reform" by means of "criticism and self-criticism" is well planned and carefully executed and constitutes an important phase of the remolding process.

Emphasis on short-term education

Millions of cadres have been given intensive training to carry out specific programs; there are cadres for the enforcement of the agrarian law, the marriage law, the electoral law; some are trained for industry or agriculture, others for the schools, and so on. Each group is trained to perform a specific function. The training may take a few weeks or a few months. This method of short-term ad hoc training is characteristic of Communist education in general. The needs are urgent and specific, and education must meet them. There is no time for lengthened courses of study, no room for frills or trimmings, no concern for what "bourgeois" education would value as breadth or depth of understanding.

The Communists reject the concepts of "bourgeois education" and its emphasis on scholarship or academic standards. Knowledge without application is useless; Mao speaks derisively of bourgeois scholars immersed in book knowledge and aloof from politics and production. The product of education is described as the "red expert." An expert is one who has special skills and ability to contribute to production, and he must be "red" by a thorough immersion in Communist ideology. From literacy classes and short-term training programs to the schools and universities, the central aim of education is to produce zealous supporters of the proletarian revolution led by the Chinese Communist Party.

There are three types of schools: the full-time school, the work–study school, and the spare-time school. Students in the work–study school spend a part of their time in productive labour; they work on farms, in factories, or in other forms of productive enterprises. Theoretically, study is organized to help solve concrete problems of production, thus bridging the gap between knowledge and action; at no time is the student removed from actual experience on the production front. The spare-time school is designed for adults who attend after their work hours; workers do so after the day's work, and peasants attend "winter schools" in the lax agricultural season. Thus, there is no interruption of production. Spare-time schools or classes have become an important vehicle of adult education.

Prescribed courses of definite length and content belong to the era of bourgeois education, which the Cultural Revolution of 1966–69 made a renewed effort to stamp out. There is no stable system of schools. Primary schools consist of any length from one or two years to five. Secondary education has been reduced to four years, and higher education to one or two or more years, according to the nature of the course. Spare-time education claims to be able to qualify an illiterate for "secondary education" by a few years of part-time study and for "higher education" in another short period. It is obvious that the terms elementary, secondary, and higher are merely relative terms and have no specific reference to commonly understood standards in terms of curriculum content. Prerequisites for further study are so flexible that they are practically nonexistent. Colleges and universities admit workers and peasants of "advanced production experience" and cadres of high political fervour and ide-

ological orthodoxy, even though their academic qualifications are inadequate by Western standards. Those who have proved to be politically "red" and to have some degree of expertise in production are judged to be more worthy of higher education than "bookworms" erudite in book knowledge but unable to serve the revolution adequately.

Directions after the Cultural Revolution. Since the Cultural Revolution, education has moved in a different direction. Premium is put on work experience; continued study is not important and is even discouraged except for a select few. All schools and universities were closed in 1966 for reorganization and for planning to move in new directions. The schools were ordered to reopen in 1967, but compliance was slow and hesitant, and the colleges and universities did not admit regular new classes until September 1970.

Emphasis on work experience

Enrollment in higher education has been greatly curtailed. Higher education will concentrate largely on engineering and science. Professional training will be given in shortened courses. Medical colleges are proposing to shorten their courses to two or three years; "barefoot doctors," trained to treat common diseases and perform simple surgery, are produced in a few months with no academic prerequisites beyond some form of general education. This is in line with the trend of practical training in the quickest possible time, with little concern for theoretical study that does not have immediately practical value.

Two developments since 1968 deserve mention. To reduce the number of students seeking advanced study, and to implement the intensified emphasis on production and labour, Mao issued a directive to send millions of young people, students, and "intelligentsia" to the rural areas for long-term settlement and to engage in agricultural and industrial production. Even students who qualify for higher education are now required to join the production front, and their production experience is a prerequisite for admission into universities. After higher study, they are supposed to return to the production front. The intelligentsia, including teachers, are to live and work in rural areas to be "re-educated" by the workers and peasants. Political "study" or ideological indoctrination is an essential part of this re-education. Only by such experience and identification with the labouring masses can the intelligentsia hope to overcome the harmful effects of bourgeois-dominated education.

Emphasis on proletarian leadership

A second measure designed also to terminate "the domination of education by bourgeois intellectuals" is to put education under firm proletarian leadership. Backed by the army, teams of workers (in city schools) and peasants (in rural schools) were in 1968 dispatched into the schools and universities to take over the management of the institutions and to play a "leadership" role in the political education of students and teachers. Known as the Mao Tse-tung Thought Propaganda Teams, these selected workers, peasants, and soldiers, under the guidance of the Communist Party, are today the governing cadres in schools and universities, and their chief mission is to make sure that Mao's thought and directives are faithfully and literally carried out. Most members of the teams have had little schooling, and some are barely literate, but they have had experience in labour and production, and they are sufficiently versed in quotations from Chairman Mao to be able to lecture to teachers and students on the correct political line that must guide all educational effort. They constitute the new "proletarian leadership" in medical schools, hospitals, technical universities, as well as other types of higher education and all lower schools.

These are startling innovations. They serve to emphasize the rejection of what the Communists condemn as bourgeois education. It would be irrelevant to raise such questions as academic standards, freedom of thought, systematic knowledge, educational stability or continuity, or the articulation of schools of different levels. Education is designed to serve the proletarian revolution and is subject to constant change in response to the needs of the revolution. There is as yet no firmly established system, but a new pattern is emerging, and the innovations here briefly described are some aspects of this pattern.

(T.H.C.)

MODERNIZING AFRO-ASIAN PATTERNS OF EDUCATION

Japan. *Education at the beginning of the century.* Between 1894 and 1905 Japan experienced two conflicts, the Sino-Japanese and Russo-Japanese wars, which increased nationalistic feelings; and it also experienced accelerated modernization and industrialization. In accord with the government's new nationalism and efforts to modernize the country, educational reform was sought. The Japanese education system took as its model the western European educational systems, especially that of Germany. But the basic idea of education was expressed in the Imperial Rescript on Education (Kyōiku-chokugo), which stressed the need for traditional education.

Effect of Japan's nationalism and economic growth

In 1900 the length of the ordinary primary school (*jinjyō-shōgakkō*) was made four years, and it was made compulsory for all children. At the same time, the cost of compulsory education was subsidized from the national treasury. In 1907 the length of compulsory education was further extended from four to six years. As the educational system gradually improved and as modernization progressed and the standard of living increased, school enrollments soared. Whereas in 1890 only 48.93 percent of the primary-age children were in school, in 1910 the percentage had reached 98.14.

In those days, boys and girls in primary school studied under the same roof, though in separate classrooms. In secondary education, however, there were entirely separate schools for boys and girls—the *chūgakkō*, or middle school, for boys and the *jogakkō*, or girls' high school, both aiming at providing a general education. Other than these, there was the *jitsugyōgakkō*, or vocational school, which was designed to afford vocational or industrial education for both boys and girls. All these secondary schools were for students who had completed the six- or four-year course of primary education. Thus, secondary education was organized into the three parallel systems—a higher general education for boys, a higher general education for girls, and a technical and vocational education.

As for the primary and secondary curriculum, the Imperial Rescript on Education in 1890 provided the ideological basis for the system, making it clear that traditional Confucian and Shintō values were to serve as the basis of moral education. This emphasis was implemented by courses on "ethics" (*shūshin*), which served as the core of the curriculum. In 1903 a system of national textbooks was enacted, giving the educational ministry the capacity to alter, add, or delete indoctrination in accordance with political currents.

To accomplish an expansion of education there was obviously also a need for teachers, and thus a system of *shihangakkō*, or normal schools for training primary school teachers, was established in 1886 and subsequently developed under the strong control of the government. At first there was only one school in every prefecture, but after 1897 the government admitted more. Normal schools for girls were established independently. All the normal schools, in any event, were run by the prefectures, and none was private. At first only the graduates of the higher primary schools were qualified for the normal school, but in 1907 a new course was introduced for graduates of the middle schools and the girls' high schools.

For training secondary school teachers, there was after 1886 the *kōtō Shihangakkō*, or higher normal school, and the *joshi kōtō shihangakkō*, or higher normal school for women. Additionally, "temporary teachers' training institutes" were established after 1902. These were all state run. There were also state-run institutes for training vocational school teachers.

In the Tokugawa era there had been academies on the Chinese model for the study of Confucianism, but not until 1877, after the Meiji Restoration, did there appear in Japan a university, or *daigakkō*, of the European variety. In that year Tokyo University was founded, with

Japan's Imperial universities

four faculties—law, physical sciences, literature, and medicine. In the early years, research and education were dominated by foreigners: most programs were taught in the English language by English and American teachers or, in the medical faculty, in the German language by German instructors. In 1886 Tokyo University was renamed the Imperial University, or Teikoku Daigaku, by imperial order and, as a state institution, was assigned to engage exclusively in research and instruction of such sciences and technology as were considered useful to the state. Modern Western sciences formed the core of this research and instruction, though some traditional Japanese learning was revived; and the university borrowed much of the style and mode of the German universities.

In one important respect it differed from the European tradition: as an Imperial university it added to the four established faculties two others, first engineering and then agricultural science. This indicated the government's intention to make the Imperial University the highest seat of learning in all important fields and to concentrate in it all the functions relating to the training of leaders of all kinds. All these reforms heightened not only the importance of higher education but also the demand for it. Thus, in 1897 Kyōto Imperial University was established, followed by Tōhoku Imperial University in 1907, Kyūshū Imperial University in 1910, and Hokkaidō Imperial University in 1918. These universities were modelled after the imperial university in Tokyo—renamed the Tokyo Imperial University—and inherited its tradition. The higher middle schools established in 1886 were remodelled into the *kōtō-gakkō*, or higher schools, in 1894. In the 20th century the higher schools developed as preparatory schools for universities.

Higher education was advanced in another area by the College Order of 1903, which enabled certain upper level private schools to be approved as *semmongakkō*, or colleges, and to receive the same treatment as state-run universities. Until then the private colleges had not clearly been given a legal status and had been treated as rather inferior.

Efforts at reform

Education to 1940. The events of World War I and its aftermath tremendously influenced Japanese society. In the postwar days, Japan experienced the panic and social confusion that was sweeping many nations of the world. Moreover, the intensified leftist movement and the terrible Kantō earthquake of 1923 caused uncertainty and confusion among the Japanese. Nevertheless, the period was one that earned the name of the "*Taishō* democracy" era, which featured the dissemination of democratic and liberal ideas. And it was also a period that marked Japan's real advancement on the world scene and the expansion of its capitalistic economy, all conducive to the flourishing of nationalism. It was quite natural that these varied currents of the times should greatly influence education.

Educational reform was carried out through the debates and recommendations of the Extraordinary Education Council established in 1917. The reform placed an emphasis on higher education, though secondary education also experienced remarkable growth. As for primary education, the target of the reform was to improve the content and methods of education and to establish the financial foundation of compulsory education.

In 1907, as noted earlier, the period of compulsory education had been extended to six years, and the ordinary primary school course had become six years. The central problem afterward was how to revise the content of education. After World War I, the new educational movements generally called progressive in the West were introduced into Japan and came to thrive there. Many private schools advocating this "new education" were established, and the curricula of many state and public schools were also refashioned. The method of new education was gradually introduced into the state textbooks. Preschool education was also encouraged; a state-run kindergarten attached to Tokyo Girls' Normal School had been first established in 1876, and later many public and private kindergartens emerged, particularly after issuance of the Kindergarten Order in 1926.

Governmental aid for compulsory education was gradually put forward, and by 1940 this developed into a system whereby the government financed half the teachers' salaries and the prefectural governments the other half. Primary education thus further expanded. Between 1910 and 1940 the number of primary teachers and pupils almost doubled. In the latter year there were 287,000 teachers and 12,335,000 pupils.

Secondary education continued to be provided by the middle schools for boys, the girls' high schools, and the vocational schools. These schools increased remarkably both in numbers of institutions and in enrollments after World War I, reflecting the social demand. As a result, the secondary schools assumed more of a popular and less of an elitist character than they had evidenced in the Meiji Era. In 1931 two courses were provided for the middle school system; one was for those who advanced on to higher schools, and the other course was for those who went directly on to a vocation. Enrollments of all kinds leaped: whereas in 1910 the enrollments in middle schools, girls' high schools, and vocational schools had been 122,000 pupils, 56,200 pupils, and 64,700 pupils, respectively, the respective figures in 1940 were 432,000 pupils, 555,000 pupils, and 624,000 pupils.

Increases in enrollments, 1910–40

Meanwhile, the *jitsugyō hoshūgakkō*, or supplementary vocational schools, which had been built after 1893 as part-time educational institutions for working students, reached enrollments exceeding 1,000,000 by 1922. In 1935 *seinengakkō*, or youth schools, were newly established, uniting these supplementary vocational schools with the *seinen kunrenjo*, or youth-training centres, that had earlier been set up to provide military training for youth.

A drastic reform of higher education was instituted in 1918, when the University Order and the Higher School Order were issued on the recommendation of the Extraordinary Education Council. Before that, there had been only the imperial universities, which were state run. The order approved the founding of private universities and colleges. As a consequence, the old influential private colleges, or *semmongakkō*, rich in tradition, were approved as formal universities or colleges, resulting eventually in such famous universities as Keiō Gijuku and Waseda. National colleges of commerce, manufacturing, medicine, and so on were also opened. In general, universities and colleges multiplied, numbering in 1930 as many as 46 (17 state, five public, and 24 private). College-preparatory education concurrently enlarged through the establishment of public and private higher schools under the Higher School Order. The higher schools were remodelled after the German *Gymnasium* and the French lycée and offered a seven-year course.

Japan's examination system

It is also important to understand the examination system that pervaded Japanese education (not unlike the dual plans of Europe). From the beginning of the Meiji era, when the modern education system was established, elementary pupils passed on to secondary schools only if they passed a competitive, selective examination. In the 20th century, as graduates from the primary schools increased and as applicants to the secondary schools also sharply increased, the competition became ever more severe. Its ill effects were widely discussed, and the Ministry of Education eventually revised the examination system, abolishing the written achievement test. Henceforth, selection was to be based on a report of the principals of the primary schools, personal interviews, and physical tests. This new system, in turn, was abandoned for a while in favour of the old written achievement test, but in 1939 the new system was re-established on the same principle as before, with only slight modifications.

The examinations admitting secondary school graduates to higher schools or colleges were decidedly more competitive than those on the lower level; indeed, in extreme cases, the number of applicants was ten or more times the number admitted. The severity of the entrance examinations is only partly attributable to the fact that applicants outnumbered the total number of seats being offered; more important, most students were competing for the relatively few places in certain elite schools. There

had developed a deep social consciousness in Japan, which respected and gave priority to the graduates of specific elite universities—a consciousness born in the process of the rapid modernization of Japan. Although the Ministry of Education did at times try to carry out reforms, placing more emphasis on school records than on selection tests, the efforts were unavailing.

Education changes during World War II. The Manchurian Incident in 1931 escalated into the Sino-Japanese War of 1937, and national life became more and more militaristic. Education acquired an intensely nationalistic character. With the outbreak of war in the Pacific in 1941, the education system underwent emergency "reforms." Elementary schools were renamed *kokumingakkō*, or national schools, under the National School Order issued in 1941. The order proclaimed the idea of a national polity or spirit peculiar to Japan; the content and the methods of education were revised to reflect this nationalism. Moreover, the period of compulsory education was officially extended to eight years, though it actually remained six years because of the worsening war situation.

Militarism and nationalism

Secondary education was similarly made "national." In 1943 the Secondary School Order was issued in an attempt to unify all the secondary schools, but it also, because of the war, shortened secondary education to four years. Military education was especially strengthened. As the war worsened, students above the secondary schools were mobilized as temporary workers in military industries and at agricultural districts in order to increase production, and a great number of students were sent to the battlefields. As a result, classes were virtually closed at schools higher than secondary toward the end of World War II.

Education after World War II. After its surrender in 1945, Japan was placed under Allied occupation and was forcibly "demilitarized" and "democratized." Drastic reforms swept over education. They were a product not only of occupation policy but also of the reflection and criticism of the Japanese themselves regarding their prewar and their wartime education.

The Japanese lost much of their ethical foundation as a result of the defeat, the first in their history. The conventional educational principles espousing Imperial supremacy received a heavy blow, and the educational world of Japan fell into spiritual chaos and vacancy. Owing to the exhaustion of war, the basic conditions for education were fairly destroyed, as reflected in the lack of food and houses, as well as in the economic confusion. Education in Japan after the war was thus confounded for a time but arose with democratic ideas as its reviving energy.

From September to December 1945, the general headquarters of the supreme commander for the Allied powers (GHQ) issued four significant directives concerning educational administration in Japan, largely aimed at eliminating militarism and ultranationalism from the sphere of education. Many wartime teachers were banished from public office; church and state were separated in that Shintō education was removed from the curricula. These directives were issued with the aim of establishing a democratic educational system.

In 1946, American experts on educational affairs arrived in Japan and issued the "Report of the United States Education Mission to Japan," which provided the basis for the directives that GHQ issued to the Japanese Ministry of Education, the actual agency effectuating all reforms. The Fundamental Law of Education and the School Education Law, both enacted in 1947, and the Boards of Education Law of 1948 set the tone and outlines of the new education and the new democratic ideas.

The new Japanese school system

The conventional European dual system was replaced by a democratic ladder system, in which school programs were unified and simplified and the period of attendance was settled in six, three, three, and four years, respectively, for *shōgakkō*, or primary schools, *chūgakkō*, or lower secondary schools, *kōtōgakkō*, or upper secondary schools, and *daigaku*, or universities. This system was an obvious reflection of the American elementary schools, junior and senior high schools, and universities. The period of compulsory attendance was extended to nine years; coeducation was introduced; and provisions were made for education for the physically handicapped and other special education.

The reform of the content of education proceeded from a reduction of the strong state control of former days and the encouragement of teachers' initiative. State textbooks were abolished in favour of textbooks produced by private enterprise, and schools were controlled locally by elective boards of education. Ethics disappeared from the curricula and was replaced by new subjects, such as *shakaika*, or social studies, designed to prepare children for life in a democratic society. All the schools showed a rapid growth. Kindergartens developed considerably as the standard of living in Japan began rising. Compulsory nine-year education was so successful that the rate of attendance in lower secondary schools exceeded 99 percent from the outset and from 1953 on maintained 99.9 percent. Because of the democratic character of the postwar school system, the number that went on to high schools increased rapidly. The percentages of those passing from lower secondary into upper secondary schools increased from 42.5 percent in 1950 to 70.6 percent in 1965.

Changes after the end of Allied occupation

The peace treaty of 1952 not only liberated Japan from the restraints of occupation but also allowed education there to be adjusted to native conditions. Central direction of the schools was increased; controls on textbook certification were strengthened; the political indoctrination of the leftist Japan Teachers' Union was hindered; and efficiency ratings for teachers were instituted. A certain reactionary trend had thus revived. On the whole, however, the postwar educational reforms were retained and advanced, and their subsequent elaboration helped match Japan's rapid economic growth.

The educational reform after World War II also altered the character of the universities, which, in democratic fashion, were opened more widely to all the people. The former institutions—universities, colleges, and various schools for teachers' training—were reorganized into four-year universities and colleges, with a less Imperial or nationalistic tone and a more professional and practical outlook. Facilities and students increased rapidly in number. The 245 universities and colleges in 1960 had increased to 382 in 1970. The number of students rose from 630,000 to 1,400,000 in the same period. Two-year courses for a master's degree and three-year courses for a doctorate (four years for medical and dental students) were established in the new system of universities. Even junior colleges, or *tanki daigaku*, were born, as institutions giving shorter, more "practical" courses in higher education. By the 1970s the total number of students attending universities, colleges, and junior colleges had reached as high as 20 percent of the youth population. This fact eloquently suggests the transformation of universities in Japan after World War II. (A.Na.)

India. *Preindependence period.* Amid the rising nationalism of the latter part of the 19th century, Indians became more and more critical of the domination of Western learning as imposed by the British rulers and demanded instead more attention to Indian languages and culture. The Indian National Congress, several Muslim associations, and other groups raised their voices against the British system of education. Nor were British authorities altogether blind to the needs of the country. When Baron Curzon of Kedleston arrived as viceroy in 1898, his determination to improve education was immediately translated into an order for a close survey of the entire field of education. It revealed: "Four out of five villages are without a school. Three boys out of four grow up without any education and only one girl out of forty attends any kind of school." Education had advanced, but had not penetrated the country as the British had earlier expected.

Political conflicts involved in education in India

Curzon applied himself to the task of putting matters in order. He disapproved of the doctrine of state withdrawal from direct educational enterprise; rather, he considered it necessary for the government to maintain a few institutions of every type as models for private enterprise to imitate. He also abandoned the existing policy of educa-

tional laissez-faire and introduced a stricter control over private schools through a vigilant policy of inspection and control. Such a policy aroused bitter feelings among some educated Indians, since it was believed that Lord Curzon was bent on bringing the entire system of education under government control.

The main battle, however, was fought over the universities. With Eton and Balliol in mind, Baron Curzon set up the Indian Universities Commission of 1902 to bring about a better order in higher education. The commission made a number of important recommendations—namely, to limit the size of the university senates; to entrust teaching powers to universities in addition to examining; to insist on a high educational standard from affiliated colleges; to grant additional state aids to universities; to improve courses of studies; to abolish second-grade colleges; and to fix a minimum rate of fees in the affiliated colleges. The report was severely criticized, and the last two recommendations had to be dropped. Legislation in regard to the other proposals was passed despite bitter opposition in the legislature and the press.

The conflict resulted less from educational differences than from political feeling for or against centralization. In one part of the country, violent agitation had already started on the question of the partition of Bengal. In another part, the patriot Bal Gangadhar Tilak declared: "Swaraj [self-rule] is our birth-right." Thus, Baron Curzon's educational reforms were considered sinister in their intentions, and his alleged bureaucratic attitude was resented. Thus, after five years of resolute efforts, Curzon failed to inspire general confidence, and he left "education a battle field."

Swadeshi movement

The administrative policy of Baron Curzon also gave rise to the first organized movement for national education. A body known as the National Council of Education, in Calcutta, established a national college and a technical institution (the present Jadavpur University) in Calcutta and 51 national schools in Bengal. These schools sought to teach a trade in addition to ordinary subjects of the matriculation syllabus. The movement received a great impetus, because the Calcutta Congress (1906) resolved that the time had arrived for organizing a national system of education. With the slackening of the Swadeshi Movement, however, most of the national schools eventually closed. The effect of the movement was nevertheless noticeable elsewhere: Rabindranath Tagore started his famous school in West Bengal near Bolpur in 1901; the Arya Pratinidhi Sabha established *gurukulas* at Vrindāban and Hardwār; the Indian National Congress and the All-India Muslim League at their sessions in Allahābād and Nāgpur, respectively, passed resolutions in favour of free and compulsory primary education.

In 1905 Baron Curzon left India. In order to pacify the general public, his successors modified his policy to some extent, but the main program was resolutely enforced. Although Indian public opinion continued its opposition, the reforms of Baron Curzon brought order into education. Universities were reconstituted and organized, and they undertook teaching (initially, Indian universities had been simply institutions conducting examinations for degrees). Colleges were no longer left to their own devices but were regularly visited by inspectors appointed by the universities. The government also became vigilant and introduced a better system for inspecting and granting recognition to private schools; the slipshod system of elementary education was also improved. The number of colleges and secondary schools continued to increase as the demand for higher education developed.

In 1917 the government appointed the Sadler Commission to inquire into the "conditions and prospects of the University of Calcutta," an inquiry that was in reality nationwide in scope. Covering a wide field, the commission recommended the formation of a Board of Secondary and Intermediate Education with full powers to control secondary and intermediate education, the institution of intermediate colleges with two-year courses, the provision of a three-year degree course after the intermediate stage, the institution of teaching and unitary universities, the organization of postgraduate studies and honours courses, and a greater emphasis on the study of sciences, on tutorial systems, and on research work. The government of India issued a resolution in January 1920, summarizing the report of the commission. Since then, all legislation of any importance on higher education in any part of India has embodied some of the recommendations of the commission, which can be regarded as the greatest single factor in the history of university education in India.

Meanwhile, World War I had ended, and the new Indian constitution in 1921 made education a "transferred" subject (that is, transferred from British to Indian control), entrusting it almost entirely to the care of the provinces. In each province, educational policy and administration passed into the hands of a minister of education, responsible to the provincial legislature and ultimately to the people. Although European-style education was still maintained as a "reserved" subject and was not placed under the control of the Indian minister of education, this anomaly was corrected by the Government of India Act of 1935, which removed the distinction between transferred and reserved subjects and introduced a complete provincial autonomy over education.

Generally, the new constitution of 1921 was considered inadequate by the Indian National Congress. In protest, Mahatma Gandhi launched the Non-cooperation Movement, the campaign to boycott English institutions and products. Educational reconstruction was kept in the forefront: national schools were established throughout the country, and *vidyapeeths* ("national universities") were set up at selected centres. The courses of study in these institutions did not differ much from those in recognized schools, but Hindi was studied as an all-India language in place of English, and the mother tongue was used as the medium of instruction. Teaching the use of *charkha* ("spinning wheel") was intended as a symbolic preliminary step toward the complete economic independence of India from the West. These institutions, however, functioned for a short time only and disappeared with the suppression of the Non-cooperation Movement. The Congress' struggle for self-rule, however, became more vigorous, and with it spread the national movement toward education to suit national needs. The Government of India Act of 1935 further strengthened the position of the provincial ministers of education, since the Congress was in power in major provinces. The developmental program of provincial governments included the spread of primary education, the introduction of adult education, a stress on vocational education, and an emphasis on the education of girls and underprivileged people. The importance of English was reduced, and Indian languages, both as subjects of study and as mediums of instruction, began to receive greater attention.

Non-cooperation Movement

General educational trends, 1921–47

On this general background, educational developments from the inauguration of reforms in 1921 until independence in 1947 can be viewed. In the field of elementary education, the most important event was the passing of compulsory-education acts by provincial governments—acts empowering local authorities to make primary education free and compulsory in the areas under their jurisdiction. Another noteworthy feature was the introduction of Gandhi's "basic education," which was designed to rescue education from its bookish and almost purely verbal content by emphasizing the teaching of all school subjects in correlation with some manual productive craft. Considerable expansion took place in the field of secondary education. During 1946–47, there were 5,298 high schools, with an enrollment of 2,199,000 pupils. (The corresponding figures for 1916–17 had been 1,507 and 57,200.) A general demand for secondary education developed with the political awakening among the masses. Schools in rural, semi-urban, and less advanced communities were established, as well as schools for girls. Some provision was made for alternative or vocational courses when the provincial governments started technical, commercial, and agricultural high schools and gave larger grants to private schools providing nonliterary courses. But the expected results were not achieved be-

cause of a lack of funds and of trained teachers. Secondary schools still concentrated on preparing students for admission to colleges of arts and sciences.

The period is also marked by a diminishing of the prejudices against the education of women. The impetus came from the national movement launched by Mahatma Gandhi, whose appeal to Indian mothers for the country's cause encouraged a spirit of nationalism. Thousands of ladies came out of the purdah and suffered the same hardships as men for the cause of national emancipation. It was also realized that the education of the girl was the education of the mother and through her of her children. Between 1921–22 and 1946–47, the number of educational institutions for girls was nearly doubled.

In the field of university education, outstanding developments included (1) the establishment of 14 new universities, unitary as well as affiliating; (2) the democratization of the administrative bodies of older universities by a substantial increase in the number of elected members; (3) the expansion of academic activities through the opening of several new faculties, courses of studies, and research; (4) a substantial increase in the number of colleges and student enrollments; (5) the provision of military training and greater attention to physical education and recreational activities of students; and (6) the constitution of the Inter-University Board and the development of intercollegiate and interuniversity activities. With these improvements, however, the educational system of the country had become top-heavy.

Postindependence period. India and Pakistan were partitioned and given independence in 1947, and one of India's first acts was to give itself a new constitution. Although this document did not change the overall administrative policy of the country (education continued to be the prime responsibility of the state or provincial governments), the union government did assume responsibility for the coordination of educational facilities and the maintenance of appropriate standards in higher education and research and in scientific and technical education.

Five-year plans

In order to raise the standard of living, the national government developed successive five-year plans—1951–56, 1956–61, 1961–66, and 1969–1974. Under this program, the majority of educational plans are not centrally controlled but rather are formulated and implemented by state governments, with limited assistance from the government of India. In the first 20 years of independence, the enrollment in recognized educational institutions rose from 18,000,000 to 70,000,000. In 1946–47, 14,000,000 children attended primary schools, representing 35 percent of the age group six to 11. In 1968–69 this had risen to 56,000,000, or 78 percent of the same age group. A comparable increase was achieved in the age group of 11 to 14: in 1946–47 the enrollment was 2,000,000, or 9 percent of the total number; in 1968–69, it rose to 12,700,000, or 33.5 percent. In the field of secondary education, the enrollment rose from 870,000 in 1946–47 (3.8 percent of the age group 14–17) to 6,590,000 (19.4 percent). The increase in enrollment in higher education has been even more rapid. There were 19 universities in 1947. Their number in 1970 was 74. There were also in 1970 some 3,200 affiliated colleges and a number of advanced technological and research institutes conferring their own degrees. Enrollments in these various institutions of higher learning rose from about 218,600 in 1946–47 to over 2,000,000 in 1970. These enrollment figures do not represent a large part of the population of the country (over 500,000,000), but, in comparison with the enrollment in the preindependence period, the expansion was significant.

The education commissions and their recommendations

Since 1947 the government of India also appointed three important commissions for suggesting educational reforms. The University Education Commission of 1949 made valuable recommendations regarding the reorganization of the courses, techniques of evaluation, media of instruction, student services, and the recruitment of teachers. Then came the appointment of the Secondary Education Commission in 1952. The recommendations of these two commissions led to some changes in school and college courses, such as the conversion of the ten-year school into one of 11 years and the establishment of multipurpose schools with diversified courses. But the general dissatisfaction with the existing condition of education continued to grow. It could not satisfy the aspirations of the people, either the educated or the illiterate. Ultimately, an Education Commission was set up in 1964. Whereas the previous two commissions had each examined only one level of education, the new one embraced education at all levels, as it was required to advise the government on a national pattern and development of education. The commission submitted its report in June 1966 with far-reaching suggestions, dealing with seven major points—a national enrollment policy, the structure and pattern of the various stages of education, a new outlook in the school curriculum, the improvement of higher education, the vitalization of educational administration and finance, a stress on professional preparation, and the status of teachers. The commission's report led to a resolution on a national policy for education, formally issued by the government of India in July 1968.

The country has planned to spend 6 percent of its gross national income on education. Even at this higher rate of outlay, India's per capita expenditure on education is small in comparison with Western standards. A century ago the first task was the education of the higher classes; today it is that of the masses. India has failed to carry out the constitutional directive of providing universal education for children in the age group of six to 14; the girls are still lagging behind boys in educational facilities; and the number of adult illiterates is daily mounting. Money is essential for solving these difficult problems.

Of late, public attention has been increasingly concentrated on the growing disparity between the number of educated persons and the opportunities of congenial and profitable employment open to them. The remedy may lie not in extending and improving facilities for higher education but in diversifying them and in changing attitudes toward industry and labour. (S.N.M.)

South Africa. From the time of the first white settlements in South Africa, the Protestant emphasis on home Bible reading ensured that basic literacy would be achieved in the family. Throughout the development from itinerant teachers to schools and school systems, the family foundation of Christian education remained, though it was gradually extended to embrace an ethnic–linguistic "family."

Early efforts to establish education in South Africa

Despite some major 19th-century legislation on the administration of education (1874 in the Transvaal and the Orange Free State, 1865 in the Cape of Good Hope, 1873–77 in Natal) and some early efforts to establish free schools, political and linguistic problems impeded the development of public education before 1900. Natal had gone furthest in affirming government responsibility for education and setting up administrative machinery to implement this responsibility, but, by and large, provision for schooling remained voluntary and piecemeal until the beginning of the 20th century.

The Boer, or South African, War (1899–1902) suspended educational development entirely and confirmed the resolve of each white South African group to protect its own cultural prerogatives. When the Union of South Africa was created in 1910, it was a bilingual state, and, in education, English-speaking and Afrikaans-speaking schools were established for white Europeans. Furthermore, a political tightness and separateness increased among the Afrikaners after the war and strengthened their tendency to exclude nonwhites from the cultural and political life of the dominant society. The trend toward separate schools for linguistic and racial groups became a rather rigid practice in most of South Africa after union.

The preliterate tribal education of African natives in the South African colonies was replaced by church mission schools. Established from 1789, they were dedicated to converting the natives to Christianity and generally inculcating an attitude of service and subservience to white men. These schools spread from 1823 to 1842, and colonial governments began making occasional grants to

them from 1854. Some mission schools included a mixture of races, but, by and large, segregation was established by custom. There were exemplary schools—such as the Lovedale Missionary Institution—that followed rather liberal social and curricular policies, but most schools held to narrowly religious content. The African mission schools were virtually brought into the state system through government subsidies and through provincial supervision, inspection, and control of teaching, curriculum, and examination standards. Schooling, in any event, has never been made compulsory for Bantu children and only in certain areas for Coloured children (racially mixed and other nonwhites).

The union and school systems

By the time the union was formed, the new provinces had each established school systems, structured mainly for European children but including provisions for other groups. Specific arrangements varied, but basically the systems were headed by a department of education under a director and controlled through an inspectorate. Three of the provinces had school boards that localized the department administration. Provision was made for public, primary, and secondary education and for teacher training (though public secondary schooling had barely begun and continued to be provided in private schools, with or without state aid). Compulsory-attendance regulations were being effected for European children, while separate school developments were under way for other groups. The language of instruction had been established provincially, with both Afrikaans and English in use.

The South Africa Act of 1910 left the control of primary and secondary education with the provinces, while reserving higher education to the union government. The Union Department of Education, Arts, and Science became the central educational authority and expanded its responsibilities through the 20th century by accepting control of special sectors such as vocational, technical, and artistic education. Including other central departments with educational responsibilities, such as the Department of Social Welfare, Department of Agriculture, and Department of Defense, the South African government exercises wide educational powers. In 1935 an Interprovincial Consultative Committee was established to coordinate educational matters among the provinces and between the provinces and the central government. In 1962 a National Advisory Education Council was appointed. By recent legislation, the central government's general responsibility for educational policy has been affirmed, while provincial areas of administrative competence have been extended.

Primary and secondary schooling, once articulated, were organized as eight- and four-year units, respectively, and have since been reorganized on a seven–five basis. Vocational high schools, now within the general secondary provisions of the provinces, offer technical, commercial, and domestic-science courses. The general high schools have been predominantly academic institutions but have come to offer a choice of subjects that differentiate pupil programs according to ability and future career intention. Private schools, both independent and aided, are found mainly in Transvaal and Cape Province, but over 90 percent of South African white children are in state schools. With exceptions, public education is free through the tenth year and is compulsory for European children to age 16 (or until the completion of the eighth year in the Cape and Natal).

English and Afrikaans and the division of education

The language of instruction does not greatly affect the curriculum or organization of schools, but it does present a basic educational division in South Africa. Both English and Afrikaans are official languages of the republic. About 60 percent of the white population speak Afrikaans and about 40 percent English, and the two languages are used almost proportionately in schools. By law, instruction is given in the mother tongue, the other language being introduced as a second language. Coloured and Asiatic pupils are taught in the official language common to their area of residence, and under recent apartheid legislation Bantu pupils are assigned to use their own tribal language.

In 1922, when the Phelps–Stokes Commission on education in Africa offered its report, South Africa's example in the development of liberal and adaptable educational provisions for Africans, particularly in Natal and the Cape, was held up for emulation. The passing of the tribal system was noted and efforts toward interracial cooperation complimented. It was obvious at that time, however, that little of particular value to Africans was being done in the European-model schools and that noteworthy educational efforts were associated with special institutions, such as Lovedale School and University College of Fort Hare in the Cape.

Concern over Bantu education in the 1940s, led to the creation of the Eiselen Commission, whose report in 1951 accorded with the separatist views of the nationalist government coming to power in 1948 and presented a basis for apartheid legislation in education. The Bantu Education Act of 1953 resulted.

Apartheid and Bantu education

The policy of apartheid, or "separate development," has been controversial, both within South Africa and throughout the world. Fundamentally, the government position rests on three assumptions: (1) that each cultural group should be encouraged to retain its identity and develop from its present stage according to its unique characteristics; (2) that, with a population of diverse racial–social groups, the way to ensure peaceful coexistence and general progress is through legal and institutional separation; and (3) that the only agency capable of exercising overall responsibility for this development is the central government. Implementation of apartheid policy has been carried out through the 1950s and 1960s and has led to a virtually total separation of educational facilities for white, Bantu, Coloured, and Asiatic populations.

Since 1953, state funds have been withdrawn from denominational native schools, and control of Bantu education has been removed from the provinces and placed within a central Department of Bantu Education. A state secretary is in charge, with assistants for administrative and for professional matters. Regional offices are under regional directors, who supervise European inspectors and Bantu subinspectors of native schools. In the consequent organizational revisions, most attention has been given to the lower primary course (first four years). The Bantu primary course is eight years in length, with a screening examination after standard two (or lower primary) and after standard six (or upper primary). Although there has been a substantial increase in Bantu school attendance over the past decade, few nonwhite South African students enter vocational or general high schools, and those who do rarely go beyond the three-year junior certificate. In 1956 the central government assumed responsibility for Bantu teacher training on a regional basis. In 1959 the racial separation of educational programs was made complete through the Extension of University Education Act. From this date, nonwhites were barred from entrance to the established universities, and separate university colleges were set up to provide higher education to nonwhites on an ethnic–linguistic basis—(1) Zulu, (2) Xosa, (3) Sotho, Tsonga, Venda, Tswana, (4) Coloured, (5) Asiatic. Syllabuses and examinations for all nonwhite colleges are prepared by the University of South Africa in Pretoria.

There is no doubt that the separate schools in South Africa offer unequal educational provisions. Such inequalities are meaningful, however, only in terms of one of two basic educational criteria: the effectiveness of education for socioeconomic progress and the adequacy of education for human development. Both criteria are problematic in South African education today. (R.F.L.)

EDUCATION IN COLONIES AND NEWLY EMERGING NATIONS

Patterns of African education. Despite the diversity of ethnic and linguistic groups and of religious beliefs and practices throughout the different geographic and climatic zones of black Africa and despite the different kinds of social organization and levels of political, economic, and social development in precolonial African societies, there are numerous common traits among them. Thus, it is possible to speak of a traditional African education—one, namely, that took place almost entirely

within the framework of family and limited social groupings, such as peer age groups and village gatherings. Chiefly, it may be said that traditional African education had the following characteristics: (1) it was collective or communalistic and deeply democratic; (2) it was intimately tied to the practical tasks of day-to-day living and it also involved learning social relations and spiritual concerns; (3) its methodology included peer-group game playing, observation of adults, and then incremental assumption of adult activities, with more and more autonomy; (4) it was lifelong, the crowning achievement being the title of sage given to the elderly; and (5) its content was often esoteric, involving, as it did, the transmission of certain secrets of one's family or secret society—secrets such as craft techniques or the preparation of medicines or poisons. The major exception to this pattern of precolonial education was the Qur'ānic school. Islām expanded south into eastern Africa in the 9th and 10th centuries and into western Africa in the 11th century. As it was assimilated into precolonial African societies, it provided them with a script (the Arabic), and, because of the central attention given to the Qur'ān, it imposed a certain bookishness on the first formal African schools. In some areas, moreover, there developed a cultural flowering, involving an indigenous literature in Arabic transcription. Significantly, Muslim instruction in black Africa borrowed freely from the practices of traditional education—instituting "bush" schools, recruiting by age groups, arranging group gatherings for reading of the Qur'ān, and adopting practices suggestive of initiation ceremonies. The influence of the Qur'ānic school, however, was marginal even among Muslimized populations; it offered the means of learning the Qur'ānic verses essential to prayer, and it trained the marabouts, but it left most aspects of traditional education intact.

Traditional African education

Although some precolonial African societies may be credited with great accomplishments, generally to the end of the 19th century there was economic, social, and technological stagnation in Africa. This stagnation is partly attributable to the defects of traditional African education—its insufficient use of literature, its iron respect for authority, its excessive reliance on empirical knowledge and techniques, and its lack of a framework or support appropriate for further progress.

General influences and policies of the colonial powers. During the colonial period, the first direct "educational" influences from outside came from religious missionaries, first Portuguese (from the 15th century) and then French, Dutch, English, and German (from the 15th to the 19th centuries). Starting from coastal bases, they undertook to penetrate into the interior and begin campaigns to convert the black populations. The missions were the first to open schools and to develop the disciplined study of African languages, in order to translate sacred texts or to conduct religious instruction in the native tongues.

Early mission and lay schools

The partition of Africa by the colonial powers in the 19th and early 20th centuries led first to the establishment of mission schools and then to the establishment of "lay" or "public" or "official" schools. The importance of either the lay or the religious system depended on the political doctrines of the mother country, its institutions (a firmly secular state or one with a state religion), and the status of the colony and its history. Whatever the system or administrative attitude, however, the fundamental purpose of colonial instruction was the training of indigenous subaltern cadres—clerks, interpreters, teachers, nurses, medical assistants, workers, and so forth—all indispensable to colonial administration, businesses, and other undertakings. Though inspired by the system in the mother country, no colonial system compared to its prototype: local examinations and diplomas, for instance, were not equivalent to those in the mother country. From the beginnings of colonialism to the end of World War II, the larger intention was not to "educate" the subject peoples but rather to extend the language and administrative policies of the colonizer.

Such a generalization, though, is subject to a slight qualification with reference to the religious missions. Both the missions and the political administrations viewed African populations and cultures with condescension, and both wished to model the African man in accordance with their own needs and objectives. The religious missions, however, became involved in the cultures of the Africans through continual contact with them in the daily ministrations; they used African languages in instruction wherever the colonial administration permitted it. Moreover, for a long while, religious establishments were alone in offering vocational education, some secondary education, and even some higher education to Africans—frequently in the face of the fears or opposition of the colonial authorities.

German educational policy in Africa. Well before Chancellor Otto von Bismarck had granted a charter to the German Colonial Society in 1885, German missionaries, both Protestant and Catholic, were operating in various regions of West, Central, and East Africa—from 1840 in Mombasa (now in Kenya), from 1845 in Cameroon, from 1847 in Peki Blengo (Togo), and from 1876 in Mpwapwa, Buganda (Uganda), and Tanga (Tanzania). English and American missionaries also operated in German areas. Instruction was everywhere conducted in the local languages, which were objects of study by numerous missionaries and by eminent scholars.

After 1900, with the organization of colonial administrations, the central governments and local authorities added lay schools to the existing mission schools. By 1903 there were, in Tanganyika, 15 mission schools and 20 lay schools. In 1911 the missions had 30,000 primary pupils and the lay schools 3,500 pupils; some 2,500 students were receiving vocational or higher education. In Togo the missionaries were subsidized, and thus by 1912 there were 181 Catholic mission schools, 156 Protestant mission schools, and only three official lay schools. Alongside these general schools in such places as Togo and Cameroon there were a few normal schools, schools of agriculture, seminaries, and apprenticeship centres.

With the creation of the Ministry of Colonies in 1907, the coastal regions were divided into zones of influence relating to the various religious missions—Catholic, Protestant, and Methodist. In these coastal areas, education received rather favourable development (the interior Muslim areas were almost untouched by the Europeans). Because primary instruction was always conducted in the local languages—German being reserved for postprimary schools—the most lasting effect of German policy was to make education in the African languages a tradition that would be difficult to undo even after the departure of the Germans. From 1898 to 1914 in German East Africa some half dozen publications appeared in Swahili—newspapers and brochures in printings of more than 10,000 copies. German colonization in black Africa thus had some permanent results.

German emphasis on African languages

Education in British colonies and former colonies. In the British colonies, as elsewhere, religious missions were instrumental in introducing European-type education. The Society for the Propagation of the Gospel in Foreign Parts, the Moravian Mission, the Mission of Bremen, the Methodists, and Roman Catholic missionaries all established themselves on the Gold Coast (Ghana) between 1820 and 1881, opening elementary schools for boys and girls, a seminary, and eventually a secondary school (in 1909). In Nigeria, Protestant missions were opened at Badagri, Abeokuta, Lagos, and Bony from 1860 to 1899, and the Roman Catholic missions entered afterward and opened the first catechism schools, primary schools, secondary schools, and normal schools. In Uganda and Kenya, the Church Missionary Society, the Universities Mission to Central Africa, the White Fathers, and the London Missionary Society opened the first mission schools between 1840 and 1900.

The first official lay schools came later and for a long time constituted a weak minority. In 1899 in Nigeria, for instance, only 33 of the 8,154 primary schools were government run and only nine of the 136 secondary schools and 13 of the 97 normal schools. Similarly in the Gold Coast in 1914 the government was responsible for only 8 percent of the schools. In Kenya and Uganda, all schools were conducted by missions. Not until 1922 did

the British government assume some responsibility for education in Uganda by opening the first government technical school at Makerere (the future Makerere University College); its purpose was to develop craftsmen such as carpenters and mechanics, as well as civil servants, business clerks, and industrial labourers. Only in territories seized from the Germans in World War I did the British take over the administration of existing government schools (previously set up by the Germans). Generally the British preferred to leave education to missions, which were given variable financial aid, usually from local and inadequate sources.

British efforts at reform in the 1920s and 1930s

Following the publication of critical reports in 1922 and 1925, when there was growing uneasiness among the Africans, the missions, the governors, and the administrators, the necessity of a precise policy on education was imposed on the British authorities. In 1925 an Advisory Committee on Education in the Colonies, created in 1924 and presided over by W. Ormsby-Gore, published an important report. The ideas, principles, and methods formulated in this document covered the matters involved in defining a policy; namely, the encouragement and control of private educational institutions, the cooperation by the governmental authorities with these institutions, and the adaptation of education to the mentality, aptitudes, occupations, and traditions of the African peoples. Special importance was attached to religious and moral instruction, to the organization and status of education services, to subsidies to private schools, to instruction in the African languages and the publication of manuals, to the training of native teachers, to the inspection of schools and the upgrading of teachers, to professional training and technique, and to the education of young girls and women. The structure of an educational system, at the most advanced stage, was to consist of an elementary education (generally six years), diversified middle and secondary education (four to six years), technical and professional schools, specialized schools of higher education, and adult education.

In practice, subsequent British policy in black Africa was far from the recommendations of the Ormsby-Gore committee. The subsidies to mission schools were subject to regulations that varied from one colony to another and paid insufficient attention to the character of the education. The development of instruction, especially secondary, was generally curbed, and various local associations and numerous organizations therefore arose to promote the expansion of education. The colonial governments exerted real effort only on behalf of schools that trained subaltern cadres for administration and commerce (mostly schools for the children of chiefs and prominent persons and the colleges at Makerere and Achimota). Government-sponsored secondary education got started only after 1930 in the Gold Coast, only in a conditional manner in 1933 at Makerere College in East Africa, and only after 1935 in Nigeria. In Uganda no complete secondary school existed until 1945.

The Advisory Committee reports published in 1935 and 1944 raised the same questions and the same fundamental themes, indicating that the government still was playing an insufficient role in education. Development was primarily a result of the efforts of missions, of various private local or foreign institutions, and of local indigenous authorities. The efforts of certain individual administrators, such as Sir Philip Mitchell in Uganda and Sir Donald Cameron in Tanganyika, were also effective. Only the political consequences of World War II brought the appointment of commissions and officials to study more deeply the problems of education in the English colonies of black Africa. After the war, the different sectors of education were developed with the growing participation of Africans, who were gaining more autonomy. Secondary education expanded. Institutions of higher learning were improved and increased in number: university colleges were established at Accra and Ibadan in 1948, at Makerere in 1949, and at Khartoum in 1951; a College of Technology (later, University of Science and Technology) was founded in Kumasi in 1951; and the Royal Technical College of East Africa (later, University College) was founded in Nairobi in 1954. Beginning in 1950, development plans for the various colonies—Ghana (the Gold Coast), Nigeria, Sierra Leone, Kenya, Uganda, and Tanganyika—contributed to educational progress.

Post-World War II reforms

Upon achieving home rule and then independence, the new African states born of the old British colonies were inheritors of an educational system that, though better than that of the other African states, was still a cause for concern. In most states (Ghana, Kenya, and Malawi being the only exceptions) less than 40 percent of the population had a primary education. Secondary education was even less widespread, Ghana being the only country in which it exceeded 10 percent. Higher education existed in urban centres but still in an embryonic state. Other serious obstacles to the ultimate development of education for all the people included the diversity of organizations and institutions responsible for education, the necessity for students to pay fees, and the complexity of the legislation in force.

Educational improvements after independence

Every one of the various countries set upon to improve education. They offered subsidies to private schools, extended supervision over them, and regulated their tuition. They increased the number of primary and secondary schools offering free or partly free instruction and created numerous institutions of higher learning, such as the universities of Cape Coast in Ghana, of Lagos, of Ifé, and of Ahmadu Bello in Nigeria, and various branches of the University of East Africa in Kenya, Uganda, and Tanzania. The educational systems inherited from colonial rule (six to eight years of primary instruction, two to four years of higher primary instruction, and five to six years of secondary education) were subjected to "Africanization" or adaptation to local conditions. Reform, in short, was widespread and profound. Taking into account the generally high rate of population growth, however, the rate of educational growth is not spectacular. Moreover, the place made for African languages in primary education seems everywhere to have been eclipsed by English, the official language—in spite of the widespread use of African languages in magazines, newspapers, radio, television, and motion pictures.

Education in French colonies and former colonies. As elsewhere in Africa, mission schools were the first to be established in French colonies. Although public or official schools appeared in Senegal between 1847 and 1895, the first such schools in Upper Senegal, Niger, Guinea, the Ivory Coast, and Dahomey were begun only from 1896 on. In 1900 there were about 70 schools and 2,500 students throughout French West Africa.

Only after 1900, with the organization of the federated colonies of French West Africa and French Equatorial Africa, was there a French colonial policy on education. By decree in 1903, education in French West Africa was organized into a system of primary schools, upper primary schools, professional schools, and a normal school. Two further reorganizations followed decrees in 1912 and 1918, and important schools were established—the St. Louis Normal School in 1907 (transferred to Gorée in 1913), the School for Student Marine Mechanics of Dakar in 1912, and the School of Medicine of Dakar in 1916. The educational organization that remained in force in French West Africa from 1924 until 1947 included a system consisting of primary instruction for six years (regional urban schools), of intermediate–higher primary education given in upper primary schools and in professional schools (generally one for each colony), and at the top the federal schools—two normal schools, a school of medicine and pharmacy, a veterinary school, a school for marine mechanics, and a technical school. The two schools for secondary education, both in Senegal (the Faidherbe State Secondary School of St. Louis and Van Vollenhoven State Secondary School, at Dakar), were reserved for Europeans and those rare Africans having French status.

Total enrollment in French West African schools rose from 15,500 in 1914 to 94,400 in 1945. The number of students in the higher primary schools grew in the same period only from 400 to 800 or 900. (The area's total population in 1945 was almost 16,000,000.)

Educational policy was stated frankly in the official statements of governors general:

> Above all else, education proposes to expand the influence of the French language, in order to establish the [French] nationality or culture in Africa (*Bulletin de l'Enseignement en AOF*, No. 45, 1921); Colonial duty and political necessity impose a double task on our education work: on the one hand it is a matter of training an indigenous staff destined to become our assistants throughout the domains, and to assure the ascension of a carefully chosen elite, and on the other hand it is a matter of educating the masses, to bring them nearer to us and to change their way of life. (From *Bulletin de l'Enseignement en AOF*, No. 74, 1931.)

After World War II, all inhabitants of the newly established "French Union" became citizens in common who were represented in the French Parliament. This political policy carried over into education, which became even more assimilationist: the old higher primary schools, for instance, became classical and modern secondary schools on the French model. An Investment Fund for Economic and Social Development provided financial and developmental aid to education—to the extent that primary enrollments rose to 156,000 in 1950 and to 356,800 in 1957, and higher primary enrollments rose to 5,800 in 1950 and to 14,100 in 1957. Technical and professional education also expanded, from 2,200 students in 1951 to 6,900 in 1957. Scholarships, awarded by the central government, the colonies, and local groups, enabled an increasing number of African youth to pursue higher education in France. In Senegal in 1950 the first French West African university was established, the Institute for Higher Studies, later called the University of Dakar, followed by those of Abidjan and Brazzaville.

In 1957 and 1958, when the colonies achieved autonomy and then a kind of commonwealth status within the new French Community established by the Gaullist constitution, education began a more intensive development, at least quantitatively. More primary schools were opened; larger staffs and more schools were created on the secondary level; teacher training was accentuated; and more scholarship students went to France. Within three years, after the French African countries had achieved full independence, this upgrading of education accelerated. Curricular reforms, however, have been slow. Although such countries as Guinea, Mali, and the Congo (Brazzaville) introduced such reforms as the Africanization of history and geography, generally the traditional French system persists, and courses are taught in French. The so-called ruralization of primary education—that is, the spread of education out beyond the towns—proceeds under the aegis of the governments and French educational officials.

The rise in the number of primary students has been spectacular: between 1955 and 1967, for instance, the number of primary graduates increased in Guinea from under 34,000 to over 164,000, in Senegal from about 62,000 to almost 219,000, and in Chad from about 21,000 to almost 140,000. Admittedly, such progress depended on recourse to unqualified teaching personnel, thus putting in doubt the qualifications of all the graduates. Also, in the former French areas, the number of students attaining a higher education has remained among the lowest in Africa.

Education in Belgian colonies and former colonies. As elsewhere and perhaps more than elsewhere, the Catholic and Protestant missions played the prime role in the development of education in the Belgian Congo (today Zaire) and in Ruanda-Urundi (the present states of Rwanda and Burundi). In the period before 1908, when the Belgian king Leopold II treated the Congo as virtually his private preserve, the missions had assumed an unofficial responsibility for education. After 1908, when the Belgian parliament assumed control of the Congo, the Catholic mission schools were given a privileged official status, with government subsidies, while the Protestant schools, though financially unassisted, were also officially authorized to operate. Throughout the colonial period the overwhelming majority of schools were missionary, and until 1948 the systems were limited to two-year primary schools, three-year middle schools, and a sprinkling of technical schools for training indigenous cadres for administration, health, teaching, and mechanical labour. In 1948 the Belgian government issued a new plan entitled "Organization of Free Subsidized Instruction for the Indigenous with the Assistance of Christian Missionary Societies," which promised more diversification in primary education (both vocational and secondary–preparatory), the institution of a six-year secondary-education program, and the future creation of institutes of higher learning. In practice, the fullest emphasis was subsequently given to the so-called general two-year primary education, which was oriented toward rural crafts, technical training, and a "democratic" distribution of education. (Education was of course still designed in the interests of colonialism.) It is in this context that the great number of persons with a primary education in Zaire must be appreciated: 909,500 students in 1950 (compared with 370,000 in 1935 or with 2,066,800 in 1965).

Catholic missionary schools in Belgian areas

In 1962 the government of the newly independent Congo proceeded with a reform of the old educational system: the first primary degree was standardized and its length extended from two or four years to six years; and pupils were to take a common primary course prior to their orientation toward general, normal, or technical secondary education or toward professional education. Numerous specialized schools of higher education were created: the National Institute of Public Works and Construction, the School of Civil Engineers, the National Institute of Mines, the Higher Pedagogical Institute, the Higher Institute of Architecture, and the National School of Administration. To the already existing Lovanium University of Kinshasa (founded in 1954) and the Official University of the Congo (founded in 1955 in Lubumbashi) there was added the Free University of the Congo (founded in 1963 in Kinshasa). The result is that the Congo is among the African countries enjoying the best school and university infrastructure, despite the long historical delay. Although less pronounced, an analogous evolution has characterized Burundi and Rwanda.

Education in other colonies or former colonies. Except for Equatorial Guinea, which became independent of Spain in 1968, the African territories administered by Spain and Portugal are the only ones still retaining the status of colonies; in all of them there are nationalist movements seeking independence and carrying on armed struggles. Quiescent or suppressed movements also exist in South Africa, South West Africa, and Rhodesia.

Generally, the ruling policy in these countries is to perpetuate colonial control within the framework of racial discrimination, or apartheid. In the Spanish and Portuguese colonies, education is designed to assimilate a thin stratum of mulattos for the propagation of the Spanish and Portuguese languages and the Catholic religion; hence the importance of mission schools. The educational system imitates that of the mother countries and includes four years of primary instruction and seven of secondary. The per capita rate of registration for primary and secondary education, however, is among the lowest in Africa—in Angola, for example, only 2 percent in 1951 and only 16 percent in 1964. Higher education is nonexistent, and only those who are "assimilated" receive it—in Spain or Portugal itself. Any improvement of the system results from the political pressures of independent African nations and of nationalist movements.

Education in Spanish and Portuguese areas

Ethiopia was less influenced by Italy, which ruled it briefly (from 1935 to 1942), than by Great Britain, which traditionally held sway in East Africa. Its educational system therefore has remained very close to that in the English former colonies. Although Amharic has become the language for primary education, English is used in secondary and higher education. School registration remain among the lowest in Africa: 8 percent in 1965.

Problems and tasks of education in contemporary Africa. The independent African states face numerous problems in implementing an educational policy that will encourage economic and social development. Pedagogical problems and economic and political problems all intermix. The difficulties confronting most governments,

however, are basically political: they are usually oriented toward maintaining political, economic, and cultural ties with the old colonial power, and they therefore maintain the old educational programs and policies—expanded, perhaps, but usually revised, if at all, very conservatively. There is a pronounced distrust of anything that deviates from the traditional path of colonial education or the colonial powers' educational system.

Thus, numerical increases in school enrollments, though occasionally spectacular, fail to correspond to the legitimate aspirations of the people or even, more modestly, to the initial objectives fixed by the governments themselves. The Conference of Nairobi in July 1968 viewed as rather alarming the lack of progress in education and literacy in the context of growing populations. There has also been concern about the financial difficulties of the different states, the unsuitability of current educational systems to local needs, the waste and duplications in primary and secondary education, and the insufficient liaison between educational policy makers and the planners of economic and social development. In short, an educational crisis has developed and ripened in black Africa. (A.M.)

Patterns of education in the Arab world. The countries comprising the modern Arab world form a vast territory stretching from the Atlantic Ocean eastward to the Persian Gulf, with over 106,000,000 Arabic-speaking people, predominantly of the Muslim faith. Arab education was shaped in the postmedieval centuries largely by two significant developments. One was the rise and fall of the Ottoman Empire, which began with the fall of Constantinople and the end of the Byzantine Empire in 1453, reached its peak of territorial growth and power under Süleyman the Magnificent in 1520–66, and then gradually declined until it was all but liquidated in the Treaty of Sèvres in 1920. What remained was reconstructed by the Treaty of Lausanne (1923) as the modern Turkish republic under Mustafa Kemal Atatürk. The other influential development was the far-reaching territorial and educational control of Arab states by European powers, chiefly France and Great Britain, gradually replacing the steadily declining Ottoman Empire. During this long period, education in the Arab world was governed to a large extent by Ottoman educational policies, French influence in areas of French control, British influence in areas of British control, and Italian influence in Libya, and finally, largely after World War I, the development of national systems of education in modern times in independent Arab states.

Historical background of modern Arab education

In efforts to break away from Ottoman and European political, economic, and educational domination, the Arab nations have tried to create among themselves unity of purpose, educational improvements, economic cooperation, cultural interchange, and military defense. Among the important means used to this end have been such movements as Pan-Islāmism, which began in the late 1800s; Pan-Arabism, which was initiated during World War II; and the Arab League, which was founded in 1945 to promote closer cultural, economic, political, and military ties among the pact members. Member nations in the Arab League include Saudi Arabia, Lebanon, Syria, Jordan, Egypt, Yemen, Libya, The Sudan, Morocco, and Tunisia.

Since their independence, almost all Arab states have made sweeping advances in the development of national systems of education. Education in the modern Arab nations draws upon not only recent European and American ideas and practices but also upon, certain traditional concepts and practices rooted in medieval and postmedieval Muslim education and surviving in scattered centres of Arab society to the present day.

Educational background and modern development in some Arab states. The impact of this historical background upon Arab education can best be illustrated by historical sketches of a few Arab states: Algeria, Egypt, Iraq, Lebanon, Morocco, and Syria.

Algerian education

Algeria was under the Ottomans at least nominally until 1848 and under French colonial rule from about 1830–48 until 1962. Only Qur'ānic schools existed before 1830, teaching religion and some reading, writing, and grammar, rarely beyond the primary level. These schools underwent a period of decline following the advent of French colonial rule and were mostly replaced with French schools for French children but open also to Muslim pupils. The legal provisions of France's free, compulsory public-school system were extended to Algeria in 1833, leading to the establishment of a complete French public educational system by 1909 and a provision for compulsory education for ages six to 14 by 1936. The University of Algiers, which combined existing schools of medicine, law, science, and letters, was established in 1879 and reorganized in 1909.

Until it was abolished in 1949 there existed a dual school system, European and Muslim, but all elementary schools established by the French government concentrated on teaching the French language. Moreover, private schools received financial support from the government only if they offered instruction in French. Without financial support, the large majority of Muslims were thus left without any education of their own. In 1951 the French government did begin a few experimental secondary schools (*lycées d'enseignement franco-Musulman*) that offered a seven-year program to French and Muslim pupils completing a five-year elementary school and, more significantly, combined elements of both Arabic and French cultures in their curricula. The aim, nevertheless, was to prepare students for the examinations of the French baccalaureate.

As early as 1931 a native group, the Association of the Learned Men, dissatisfied with the traditional Qur'ānic schools, established modern Arabic primary schools, which added such subjects as North African geography and arithmetic to the traditional religious subjects; 193 such schools were operating in 1955. A few madrasahs offered advanced training in Arabic, but they were closed in 1957, except for three, which were converted to normal schools for training teachers in Arabic culture. The best was the Institut Benbadis in Constantine, with a four-year Arabic secondary program. This was closed along with many other Muslim schools in the 1962 revolution, initiating the present educational system.

Although a dual system of elementary and secondary education continued immediately after independence, steps were taken between 1962 and 1964 to transform Algerian education into a national system under the Ministry of National Education. The aim was to introduce compulsory teaching of Arabic at all levels, to train Arabic-speaking teachers, and to Algerianize the curricula of elementary, secondary, and higher education. Classical and modern high schools were needed, as well as technical colleges, teacher-training colleges, vocational schools, and universities. Education, in short, was to be democratized to make it available to all Algerian students at all levels. Professional training was needed for teachers and for education counselors, and programs of adult education were planned. It was necessary to construct schools and supply equipment sufficient to these needs, to prepare thousands of Arabic texts, and to arrange educational relations and exchanges with other countries.

Egyptian education

Of all the Arab countries in the Middle East, the Arab Republic of Egypt has the longest educational history. Its higher education (religious) dates back over 1,000 years, as represented by the al-Azhar University. Whereas other Arab nations broke away from the educational program of the Ottoman Empire, Egypt has retained in part the Ottoman system that dates back for centuries. With such a long history, Egypt has developed a comparatively complex educational system, continually modified by laws initiating new types of educational organization and programs based on new ideas and needs. Because of this historical background, the system today is at once modern and ancient, with some practices dating back to medieval times and others based on modern Western policies, programs, and practices.

The Egyptian Ministry of Education, founded in 1836–37, has continued (except for a short interval under Sa'īd Pasha) to exist to the present day. It was influenced first

by French methods of education and later, after British occupation, by British influence until 1922; during this period, both French and English languages and cultures continued to influence, if not shape, Egyptian education. When Egypt received its independence in 1922 and the Ministry of Education went to Egyptian hands in 1923, it began dramatic progress under central control.

As early as the 1820s, Egypt had been sending technical missions to Italy and France for new ideas on the development of a modern school system and a modern press. But the plans for a modern education were not realized until after independence and, particularly, not until after the 1952 revolution under Gamal Abdel Nasser. A ten-year educational plan was launched in 1952 to build 400 schools annually. The present educational enterprise thus includes a steadily increasing enrollment, not only in primary and secondary schools but also in preparatory and technical schools, including industrial, commercial, and agricultural education. In higher education, there are the Universities of Cairo, Alexandria, al-Azhar, 'Ain Shams, Asyūt, and others, and there are also higher institutes and vocational-training centres.

Iraqi and Lebanese education

In the golden age of Islāmic education from 800 to 1100, Mesopotamia (modern Iraq) was one of the great centres of learning in eastern Islām. It fell under the Ottoman rule from the middle of the 16th century to World War I. During this long period the Turkish language was the medium of instruction in almost all schools, and, because of the linguistic barriers, the large majority of Iraqi children were deprived of any kind of education. Instruction in religion and classical learning dominated education until 1850: elementary schools emphasized reading, writing, and memorization of selections from the Qur'ān and other religious literature; the little higher education that existed dealt largely with Muslim theology and the study of Arabic classics. Iraq did share some of the Ottoman educational reforms that established a secular system of public education throughout the Ottoman Empire in the mid-19th century, but the new school system did not replace the traditional *kuttābs*, or religious schools.

Since World War II, steady modernization of education has taken place at all levels in Iraq. Efforts have been exerted toward establishing kindergartens; primary education is both free and compulsory by law but awaits new schools and facilities for fuller implementation; and secondary education has been growing at a slow pace.

Prior to 1926, Lebanon was an integral part of Syria, which after the fall of the Ottoman Empire in 1918 was mandated to France in 1922. In 1943 Lebanon gained independent sovereignty. By that time, however, it had already had an organized system of education dating back about 200 years. With a Christian majority, the country has historically attracted the establishment of many sectarian schools, such as 'Ayn Waraqah in 1789, which boasted of teaching four languages—Arabic, Syrian, Italian, and Latin—in addition to all courses usually taught in the elementary and secondary schools of Europe, including theology, philosophy, and law. The school became a model for the developing national schools in Lebanon, concurrently with the establishment of many other sectarian schools. The American missionary schools were first established during the first quarter of the 19th century; and British–Syrian mission schools were set up during the third quarter of the 20th century. All these schools contrasted with the Latin types represented by French and Italian schools and with the schools representing Greeks, Danes, Germans, and Russians.

Because only one Ottoman public school existed prior to World War I, the French easily took over and initiated their own public school system, including an upper elementary level. Nevertheless, at the time of independence (1943), private schools in Lebanon assumed by far the largest share of formal education (953 private compared with 252 public schools). Independent Lebanon soon decreased the ratio by establishing a complete public school system. Thus by 1964 there were 1,273 public elementary and secondary schools and 1,456 private schools at both levels. In 1966 enrollments in primary schools numbered 360,000; in secondary schools 82,000; and in higher education 23,400. The system also includes vocational and technical schools and teacher education. Lebanon enjoys today the highest rate of literacy, 87 percent, of all the nations of the Middle East.

Moroccan and Syrian education

Before independence in 1956, Morocco was under Spanish and French protectorates and was an international zone for 44 years. The French government maintained an exclusive educational system for French students, open to only a few Moroccans. Except for the traditional Qur'ānic reading and memorization schools, the establishment of modern schools in Morocco awaited independence, after which began a continuous program of educational expansion in number and types of schools as well as in a variety of curricula to meet diversified needs. As far as possible, subjects in primary schools are being taught in the Arabic language. Primary education is preparatory for secondary schools, which lead to higher education in al-Qurawīyīn University, the oldest university in Morocco, and the modern Muhammad V University at Rabat, with faculties of law, humanities, science, sociology, medicine, engineering, and theology.

Syria achieved independence in 1941 after 19 years of French mandate. In 1958 Syria joined Egypt in forming the United Arab Republic but dissolved the union, along with Yemen, in 1961 and has since been known as the Syrian Arab Republic. In the last two decades the republic has made great educational advances, having, next to Lebanon, the highest literacy rate in the Arab world.

The educational system of the Syrian Republic provides, under the Ministry of Education, elementary, general-preparatory (junior-high-school), general-secondary (senior-high), technical, and higher education, including college and university. There are also separate teacher-training institutes to meet the needs of elementary and secondary schools; lower technical schools, including industrial, commercial, agricultural, and feminine arts; and higher technical institutes, including commercial, agricultural, fine arts, and industrial training. University education, as that in Aleppo, includes faculties of medicine, dentistry, nursing, pharmacy, law, arts, sciences, commerce, engineering, education, and Islāmic law.

Arab education: prospects and problems. Independence caught the Arab nations unprepared for the demands and advantages of the postwar technological world. To make the adjustment swiftly, they are undergoing a revolution through education. In 1965 out of a population of 105,000,000 they have enrolled 12,500,000 students in their schools at all educational levels—9,700,000 in primary schools, ages six to 12; 1,840,000 in secondary schools. Similar advances have been made in the 26 universities and in a number of colleges and institutes; their enrollment totalled 248,000 in 1963–64.

Although elementary and secondary education is tuition-free in all Arab states and higher education in some, the unprecedented increases in population—at the rate of 3 percent a year, one of the highest rates in the world (from 92,000,000 in 1960 to 105,000,000 in 1965)—and a lack of funds and facilities are still making the advantages of education inaccessible to many Arab students. Almost one-half of the Arab population is still illiterate. Remarkable advances have been made, however, in the status and education of girls in most Arab states, at all levels, including higher education.

In the area of knowledge and skills there have been notable advances in scientific research, social sciences, and the arts, with relatively insufficient development in scientific and technological studies, at both the secondary and the university levels. Although every segment of the educational enterprise is under continuous revision, the greatest need in Arab education today is sweeping advances in applied sciences preparing trained individuals to manage a developing industrial society. There is a danger that many university graduates, who specialize in large numbers in humanities, arts, and social sciences, may find themselves unemployable in a society with a pressing need of technologists.

Education in Israel. Geographically in the midst of the Arab world but ideologically and politically in conflict with it is Israel, a relatively new homeland for a

people who have traditionally placed great importance on education and scholarship.

Although the new Israeli government in 1949 expediently reaffirmed the status of the various educational systems inherited from British rule, the State Education Law of 1953 reformed them by setting up a national school system of two parts. There is the Hinukh Mamlakhti, or State Education System, which offers general and vocational education; and the Hinukh Mamlakhti Dati, or Religious State Education System, having a vocational, religious, and Zionist background. Each system has its own teachers and administrators, but both are supervised by the Ministry of Education and Culture. In addition, there is a third and private system, the Aguyat Yisrael, of religious but non-Zionist cast, which receives government subsidies so long as its basic curriculum meets state requirements. For Arab minorities there are separate schools conducted by local Arab authorities.

According to the 1953 law, the aim of Israeli education for Jews is

> to base elementary education on the values of Jewish culture and the achievements of science, on love of the homeland and loyalty to the State and the Jewish people, on practice in agricultural work and handicraft, on *ḥalutzic* [pioneer] training, and on striving for a society built on freedom, equality, tolerance, mutual assistance and love of mankind. (From *Israel Government Yearbook, 1968–1969.*)

Israeli emphasis on Jewishness

The emphasis on Jewishness is continued on the secondary level, at which study of Hebrew language and literature, the Bible, and the Talmud are stressed along with more conventional sciences and humanities.

Institutions of higher education in Israel are generally self-governing but are advised by a national Council of Higher Education. The chief institutions are the Hebrew University of Jerusalem, Tel Aviv University, Technion (or the Israel Institute of Technology) in Haifa, and the Bar-Ilan University in Ramat Gan. (M.K.N.)

Patterns of education in Latin America. The term Latin America is a facile concept hiding complex cultural diversity. This abstraction covers a conglomerate of areas, distinguished by differences not only in the Indian and Negro population base but also in the superimposed nonindigenous patterns—Spanish, Portuguese, French, Dutch, and Anglo-Saxon. In this brief survey, generalizations will be limited to the major Spanish and Portuguese patterns, which account for 95.7 percent of the population and 98.3 percent of the area.

The heritage of independence. At the beginning of the 19th century, the Spanish colonies enjoyed a prosperity that led to optimism, thoughts of independence, and republican rule. In the prolonged struggle for independence, they were all but ruined, and the change from absolute monarchy to popular democracy was far from easy. The revolutionaries had a blind faith that a separation of the branches of government and a written constitution would miraculously change overnight a reality three centuries in the making. Thus novel customs clashed with those of the past; governmental practice did not follow political theory; and the legal equality of the citizens hardly corresponded to economic and educational realities. An abyss separated the educated, rich Spaniards born in America from the rest of the population, and the majority of the population lacked the education to benefit from its new rights. Although a small group of poor criollos and mestizos who had served in minor administrative posts during colonial times sought to bring about changes, they themselves and the rich, in positions of advantage, split into liberal groups and conservative groups. The struggles among them would lead to almost a century of anarchy. In this respect at least, Brazil was more fortunate, for it obtained its independence peacefully, passing from colony to empire and then to a republic.

Catholicism and education in Latin America

When educational ideologies began to be defined, the Roman Catholic Church became a source of controversy. The conservatives wanted to follow established educational patterns and considered Catholicism a part of the national character. The liberals thought it a major obstacle to progress. To uproot the church's privileges, governments in many countries made education both secular in character and a state monopoly. In others, as in Colombia, by way of a concordat with the Holy See, religious education became the official one.

Everyone was aware of the need to educate in order to create a new national consciousness, and the task was begun by importing the Lancaster method of "mutual" instruction (so named from its developer, the English educator Joseph Lancaster), which in monitorial fashion employed brighter or more proficient children to teach other children under the direction of an adult master or teacher. Its obvious advantage was that it could accomplish an expansion of education rather quickly and cheaply. Beginning in 1818 it was introduced in Argentina; 1821 in Chile and Colombia; 1822 in Peru and Mexico; and 1823 in Brazil. In 1824 Simón Bolívar personally invited Lancaster to visit his Greater Colombia. Until well into the second half of the 19th century, it was to be the most widely used system.

Almost all the heroes of independence tried to establish schools and other educational institutions. José de San Martín founded the National Library and the Normal Lancasteriana, a teacher-training school, in Lima; Bolívar established elementary schools in convents and monasteries and founded the Ginecco (1825), known afterward as the Normal Lancasterian School for Women. Bernardino Rivadavia, the first president of Argentina, also stimulated educational development, including the establishment of the University of Buenos Aires. In midcentury, Benito Juárez in Mexico also championed education as the only bulwark against chaos and tyranny.

By the 1870s the liberals had won the day almost everywhere throughout Latin America. Education was declared to be compulsory and free, the lack of teachers and teacher colleges notwithstanding. A program to remedy this situation was launched. Chile paid for educator Domingo Faustino Sarmiento's travels to the United States and Europe and enabled him to found, on his return in 1842, the Normal School for Teachers. This was the first non-Lancasterian teachers college and was to be followed in 1850 by the Central Normal School in Lima and in 1853 by the Normal School for Women in Santiago. Countries with more acute educational problems, such as Ecuador, simply imported the Brothers and Sisters of the Sacred Heart and put them in charge of organizing their educational system. During the 1870s and 1880s, foreign teachers began to be imported, while at the same time students were sent abroad. Sarmiento had already called in North American teachers to open his normal schools in the 1860s, and Chile invited Germans for its Pedagogical Institute (1889). Germans and Swiss came to Mexico and Colombia; a number of distinguished Mexican educators were trained by Germans in the Model School in Orizaba. With the foreign professors came new pedagogical ideas—especially those of Friedrich Froebel and Johann Friedrich Herbart—and also new ideologies, foremost among them positivism, which flourished in Argentina, Brazil, Chile, and Mexico.

Administration. With independence, the task of overseeing public instruction fell to the state and local authorities. Fiscal poverty and a lack of trained personnel soon proved them unequal to the task. Furthermore, since most existing schools were confessional and private, the need for intervention by the central authorities to enforce unity became obvious. In 1827 the Venezuelan government established a Subdirectory of Public Instruction, which in 1838 became a directory. Mexico established a General Directory of Primary Instruction in 1833. Soon, some countries decided to assume responsibility for centralization through a ministry for public instruction—Chile and Peru in 1837, Guatemala in 1876, and Venezuela in 1881. Other governments abstained from accepting total responsibility. In Mexico, no ministry was created till 1905 and then only with jurisdiction over the Federal District and territories; even that became a victim of the revolution of 1910. In 1922 a Mexican ministry was re-established, now in charge of the whole republic and taking up the functions that the states could not fulfill. In Argentina the Lainez Law, decreed in

Efforts to centralize educational administration

1905, authorized the National Council of Education to maintain, if need be, schools in the provinces.

Today, in all countries the control over education is in the hands of a ministry of public education or a similar government unit. Its functions include planning, building, and administering schools, authorizing curricula and textbooks for public elementary and secondary schools, and supervising private ones. In some countries, the states sustain their own educational systems, which the federal government then supplements, but, because of the disparity between city and countryside, these federal governments often have had to shoulder almost the total burden of rural elementary education.

Primary education and literacy. At the time of independence, elementary education consisted of teaching reading and writing, the religious and civil catechisms, and rudiments of arithmetic and geometry. By the second half of the century, it became differentiated between "elementary primary" and "superior primary" education, and the curriculum was enlarged to include the teaching of national language, history, geography, rudimentary natural sciences, hygiene, civics, drawing, physical education, and crafts for boys and needlework for girls. The elementary primary school was increased to five or six years, and the superior primary was to become the secondary school of the 20th century. These educational levels absorbed the greatest part of the governmental efforts and became a means to do away with illiteracy and also to create a concept of citizenship.

Primary instruction was improved by special programs and teacher training, and both benefitted from educational influences coming from abroad but also from improvements resulting from the study of national problems. Today, primary school teachers are trained in teachers colleges having the status of secondary schools.

Problems of rural and native education

Thanks to solid foundations laid during the 19th century, public education in Argentina and Chile reached a high level of competence. In other countries, because of such factors as a more heterogeneous population, a higher level of demographic growth, and greater geographical barriers, the results of great efforts have been less than impressive, Although all countries have declared primary instruction to be free and compulsory, the situation in reality is rather complex. Whereas, in towns, many children have gone from kindergarten to secondary schools since the beginning of the century, in the rural areas, even today, many schools have only one teacher to handle students of all levels. Furthermore, because many Indian citizens do not understand Spanish, special instruction is required. In the 20th century, governments have established special institutions for Indians, such as the "departments of native instruction," "native institutes," "special normal schools," and "cultural missions." The first cultural mission was created by the Mexican secretary of education, José Vasconcelos, in 1923. The idea was to send an elementary school teacher, an expert in trades and crafts, a nurse, and a physical-education teacher to underdeveloped communities, in which, during a limited period, the population would be provided with some general education. The United Nations Educational, Scientific and Cultural Organization (UNESCO) has helped in the training of teachers for these special areas through two regional centres of fundamental education for Latin America (CREFAL), one in Mexico and the other in Venezuela. Many countries have tried to master the dropout problem by offering at least one free meal a day to those who continue their schooling.

Uruguay, Argentina, and Chile have been able to multiply their schools and thus to provide facilities for their entire population of school age. In other countries, the efforts may be gauged by comparing statistics. In Peru, only 29,900 children went to school in 1845; but there were 59,000 in 1890; 1,075,000 in 1955; and 2,054,000 in 1965. In Brazil, there were 115,000 pupils in 1869; 300,000 in 1889; 4,878,000 in 1955; and 9,923,000 in 1965. In Mexico, there were 349,000 in 1874; 800,000 in 1895; 3,349,000 in 1955; and 7,813,000 in 1969. Unfortunately, the high population-growth rate (2.9 percent) makes it difficult to keep up with the ever-increasing needs.

Fight against illiteracy

Illiteracy has been fought by various means in accordance with the political and socioeconomic situation. Until the middle of the 19th century, illiteracy in Latin America was in excess of 90 percent. Of Brazil's population, only 1.5 percent were literate in 1823. Around the beginning of the 20th century, illiteracy had decreased to 39 percent in Argentina (1908), 50.4 percent in Uruguay (1908), and 68.2 percent in Chile (1895); in other countries it fluctuated between 80 and 98 percent. By 1960 illiteracy was down to 8.6 percent in Argentina, 9.7 percent in Uruguay, 16.4 percent in Chile, 37.8 percent in Mexico, 38.9 percent in Peru, and 39.4 percent in Brazil. Nations with the greatest illiteracy were Guatemala, with 62 percent; Bolivia, with 68.9 percent; and Haiti, with 89.3 percent.

Secondary education. During the 19th century, many countries established new secondary schools on the basis of colonial institutions. Thus, in 1821 Argentina converted its College of San Carlos into its College of Moral Sciences. Mexico attempted a total reform in 1833 but would not complete it until 1867 with the founding of the National Preparatory School, which involved reforming the whole system on the basis of Positivist philosophy. In Brazil in 1837 the Pedro II College was founded as a model secondary school, but secondary instruction did not prosper till the return of the Jesuits in 1845 and was to be supplemented later by *gimnasios*—that is, *Gymnasien* on the German model. Peru and Venezuela established national colleges, and Chile and Argentina created *liceos* (modelled on the French lycées) and, later, national colleges. (The term college in all cases here is used in the continental European sense to refer to secondary institutions, not institutions of higher education.)

Secondary emphasis on university preparation

In all countries (except perhaps Chile), secondary instruction has been considered a preparation for the university. All attempts to make it more formative and practical have failed, in spite of the fact that the government has taken charge. The secondary-preparatory course lasts from five to six years, with a degree of bachelor (*bachillerato*) usually given upon its completion. Its teachers come from the humanities departments of the universities and the superior normal schools (which have existed since 1869 in Argentina, since 1889 in Chile, and in the 20th century in the other countries).

Polytechnical education—industrial, commercial, and agricultural—had been a concern of liberal governments since the end of the 19th century but has developed only recently. Traditional prejudices against practical instruction were overcome only after industrialization began. It has been emphasized only in Argentina, Venezuela, Chile, and Mexico.

Higher education. Imbued with a revolutionary spirit in which education was a vital element, Latin Americans founded ten universities between 1821 and 1833, among them the University of Buenos Aires (1821). Bolívar himself established two in Peru—Trujillo (1824) and Arequipa (1828). With independence, practically all theological faculties had disappeared, and their position of pre-eminence was taken over by faculties of law.

Four universities were founded in the 1840s, Chile's among them, and ten more in the second half of the 19th century. In Mexico the new institutions called themselves institutes of arts and sciences, because the University of Mexico (founded in 1551) was associated with colonialism and had become a favourite target of the liberals. The University of Mexico was suppressed in 1865, after several earlier attempts, not to be reopened until 1910, the year of the revolution. Argentine liberals solved their problem by passing the Avellaneda Law (1885), which allowed only national universities, prohibiting private universities (until the reform of 1955).

In Brazil, the republican constitution of 1891 decreed that all secondary and higher education would be a federal or state function. Several higher schools were established, but the country's first university, at Rio de Janeiro, was not founded until 1920.

Almost all higher education in Latin America came to be secular and state operated. The fact that Latin American governments, themselves unstable, generally took

Emphasis on state universities

charge of higher education, however, explains in part its uncertain existence.

Some colonial religious institutions nevertheless survived. During part of the 19th century, for instance, the University of the Republic in Montevideo maintained its ties with the church. In 1855 the University of San Carlos in Guatemala, through a concordat with Rome, reverted to pontifical status. But, with the exception of the Universidad Católica in Chile, founded in 1888, and the Pontificia Universidad Católica del Peru, founded in 1917, all religious universities are recent creations. Indeed, today the majority of private institutions are religious or confessional, with the significant exception of some recently established technological institutes (Monterrey, in Mexico; Buenos Aires). The need for technical education was also recognized by the Mexican government when it founded, in 1936, the National Polytechnical Institute as its second national institution of higher learning, with several branches in the country (regional technological institutes) to serve the particular needs of each region.

Until the 20th century, universities were mainly professional schools. Often, they also supervised primary and secondary education (Uruguay 1833–37; Chile 1842–47; Mexico 1917–21). Today, they also conduct research and try to encourage regional developments. Unofficially, they have sometimes played a role in political life. Since the reform movement for student representation at the University of Córdoba, in Argentina, in 1918, they have become involved in political controversies. The Mexican government tried to extricate the National University from political strife by giving it autonomy in 1929. Student demonstrations by the late 1960s, however, proved this measure to lack effectiveness.

Higher education has proved to be the best means of furthering social mobility. In spite of this, institutions of higher learning in Latin America have suffered from several handicaps. Foremost is the lack of sufficient funds, which usually results in poor research facilities. Second, both students and professors are generally engaged only half time. This increases the dropout rate and decreases performance. Thus, most highly qualified professionals are trained abroad. At the same time, both the political situation and economic pressures have induced an exodus of the most highly educated Latin Americans to the United States.

In 1960 there existed 148 institutions of higher learning in Latin America. Four countries had the highest number of university students, in absolute terms: Argentina, 195,000; Brazil, 95,600; Mexico, 95,500; and Colombia, 45,000. But, on the basis of the number of students per 10,000 population, the four highest places were taken by Argentina, 93; Uruguay, 48; Cuba, 38; and Chile, 35. Colombia had only 32, Mexico 27, and Brazil 14.

(J.V.de K.)

Historical background of modern Southeast Asian education

Patterns of education in Southeast Asia. The history of education in the countries in Southeast Asia (namely, Bangladesh, Burma, Cambodia, Sri Lanka [former Ceylon], Indonesia, Laos, Malaysia, Pakistan, the Philippines, Singapore, and North and South Vietnam) bears the imprint of the distinctive cultural heritage of each of these countries—a product of its geographical, ethnic, linguistic, religious, economic, and political background. The course of educational development in these countries, however, seems to fall into a general pattern with three notable phases.

The first phase represents the earliest period, steeped in legend and history. Education during this period centred on the social and cultural values of the religions that prevailed in some of the greatest civilizations of the time—namely, the Indic, the Chinese, the Buddhist, and the Islāmic. The second phase marked the process of westernization of education through contact with Western countries. The aim was primarily to educate an elite group on Western lines in the furtherance of the interest of the ruling power. The third phase came in the wake of the resurgence of the people in this region after World War II. The most striking feature of this period was the attempts at planned development—to expand education rapidly, to improve its quality, and above all to correct the old imbalance by shifting the emphasis from the traditional liberal, general education to scientific and technical education. The gap between resources and needs is the chief constraint on faster progress.

All countries in this region are committed to the objective of universal primary education of at least seven years' duration. But about half the children of school-going age in the region are still out of school, and the dropout rate was as high as 53 percent in 1960. The majority of the adult population is still illiterate (though Sri Lanka, Indonesia, Malaysia, the Philippines, and Thailand are exceptions in this regard). Education of girls lags far behind that of boys. The proportion of girls enrolled as compared with boys, though rising gradually, was only 38 percent in 1967 (an exception is the Philippines).

By and large, the rate of growth in higher education has been faster than in any other levels. A development of far-reaching social consequence is the movement to extend the use of national languages in all fields of higher education. Although English remains the most important foreign language in the region, in such countries as Indonesia and Thailand, the national languages are in use as mediums of education at all levels, with English as the second language. Other trends include increasing attention to research and the development of services for academic counselling and career guidance.

Education in all these countries is organized in three levels—primary, secondary, and higher. The duration of primary education varies from five to seven years (in a few cases divided into two stages), the commencing age being five to seven years. Secondary education with two to three stages also has a duration of five to seven years, ending about age 17 to 19, the commencing age for higher education. The responsibility for the administration of education is shared between the national and the provincial or local governments in varying measure in the region. In Pakistan, for example, it rests predominantly with the provincial governments, but it is largely with the national governments in Indonesia, Thailand, and the Philippines. In all countries, the national governments are responsible for overall planning and coordination of education.

During the 20th century the indigenous system of education continued in all of these countries in some form, along with the new system evolved on Western lines. Western models with modifications to meet the needs of the changing society became the mainstream in each country. In spite of the local changes, it retained its resemblance to the educational system of the country on which it was patterned. By and large, the British pattern is found in Bangladesh, Burma, Sri Lanka, Malaysia, and Pakistan; the Dutch in Indonesia; the American in the Philippines; and the French in Cambodia, Laos, and North and South Vietnam.

Bangladesh. Comprising what was formerly the eastern wing of Pakistan, Bangladesh emerged in December 1971 as an independent sovereign state with a population of over 75,000,000, the eighth-largest in the world. It has a fine educational and cultural tradition. According to an early-19th-century survey, 100,000 schools were already in existence in Bengal when the beginnings of Western education for an elite group were laid in 1835 under British rule. An extremely valued educational asset of the country is its rich national language, Bengali.

Notable among the earliest educational institutions established on Western lines were Dacca College (1841), Rajshahi College (1872), and the University of Dacca (1921). In 1969–70, there were six universities and 286 colleges, with an enrollment of over 270,000; 5,200 secondary schools, with an enrollment of 1,400,000; and over 29,000 primary schools, with an enrollment of 6,300,000. Post-secondary technical and vocational institutions of various types numbered 57, with an enrollment of 13,000, indicating a wide leeway to be made up in this critical area of education. Major reforms are under way to orient the educational system to a new social order inspired by the ideals of "nationalism, democracy, socialism, and secularism" on which the new nation is founded.

Burma. The indigenous system of education in Burma

Buddhist and Socialist influence in Burma

consisted mainly of Buddhist monastic schools imparting education at both primary and higher levels. They were based on (1) the moral code of Buddhism, (2) the divine authority of the kings, (3) the institution of *myothugyi* (township headmen), and (4) widespread male literacy. The Western system of education was established after the British occupation of the country in 1886. The new system recognized women's right to formal education in public schools, and women began to play an increasingly important role as teachers. The Government College at Rangoon and the Judson College established in the 19th century were incorporated as the University of Rangoon under the University Act of 1920.

Ever since independence in 1948, Burma has passed through significant political changes. A unified system of state schools with Burmese as the language of instruction was introduced in 1950. The major objective was indicated in the proclamation on the "Burmese way of socialism" made by the Revolutionary Council: "An educational system equated with livelihood and based on socialistic moral values will be brought about. Science will be given precedence in education." As a result, the curriculum has been redesigned to achieve a balance among academic subjects, civic education, and practical arts. The teacher-education program is oriented to the needs of rural schools and to local economic growth and social development. As a result of various measures, enrollments increased between 1946–47 and 1962–63 nearly fourfold in the primary, 14-fold in the secondary, and sevenfold in the higher stage of education.

Buddhist, Muslim, and Western influences in Ceylon

Sri Lanka. Following the advent of Buddhism in the 3rd century BC, the monastic system of education centring on the teaching of Buddha became the principal system in Ceylon (now Sri Lanka). With the beginning of the Muslim settlements in the 7th century AD, schools for Muslim children were established. Western education was introduced by missionary workers arriving in the wake of the Portuguese, Dutch, and British occupations of the island between 1543 and 1802. The first schools established by the British were English schools at central stations. It was, however, the reforms introduced during the administration of Sir Hercules Robinson in the 1860s that provided the system instrumental in building the modern Sri Lanka.

Various legislative measures extended elementary education and eventually raised the compulsory school age to 14 years in 1917. Further reforms introduced during 1931–46 paved the path of educational progress. With the literacy rate rising from 56.3 percent of males and 21.2 percent of females in 1921 to 70.1 percent of males and 43.8 percent of females in 1946, Ceylon established an early lead in literacy in the region. Until 1947 education beyond the primary level was practically confined to the middle class. The movement to put the entire educational system within a national orientation, which gained strength with the introduction of universal franchise in 1931 and of parliamentary government in 1947, found official acceptance after independence in 1948.

A milestone in the growth of higher education was the establishment of the University of Ceylon in 1942. A separate university was formed from the Colombo campus of the University of Ceylon in 1967. Within the indigenous classical system of education, two universities, Vidyodaya and Vidyalankara, were established in 1959 for the promotion of Sinhala and Buddhist culture.

Malay, Chinese, and English heritage in Malaysia

Malaysia and Singapore. The origin of the educational system of Malaysia can be traced to the Qur'ānic schools, which after the advent of the British rule became secularized and developed into government schools with Malay as the medium of instruction. In 1916 these schools were remodelled according to the recommendations in the Winstedt Report. There were also three other streams of schools using either Chinese, Tamil, or English as the language of instruction. The first major development in higher education was the establishment of the Malay College at Kuala Kangsar in 1905 "to prepare the highborn Malays" for administrative service. Among other institutions of higher education are the University of Malaya established at Singapore in 1949 and a division of this university established at Kuala Lumpur in 1959; this division became the National University of the Federation of Malaya in 1961; the division at Singapore is called the University of Malaya. The second university in Singapore, Nanyang University, sponsored by a group of wealthy Chinese businessmen in 1953 and incorporated in 1958, is predominantly Chinese in character.

The Ordinance of 1957 as amended by the Education Act of 1960 provides the legal framework of the present structure of education in Malaysia. It is based on the concept of a single national school system, a unified teaching service, and a federal inspectorate, with local authorities for primary and vocational education. Under the first Malaysia Plan, 1966–70, there were striking gains: enrollment in primary schools exceeded 90 percent of the age group; enrollment in secondary schools rose by 70 percent and, in the university stage, by 200 percent.

Muslim and British influences in Pakistan

Pakistan. The educational system that grew up in the subcontinent during 2,000 years before the advent of the British was in turn Brahminic, Buddhist, and Islāmic in character. The Muslim educational system became the mainstream of education from the 11th to the 18th century. Although British rule was established in 1757 over parts of the subcontinent, no significant step in education was taken by the new government until 1835, when a few schools teaching in English were set up in the principal towns. The first universities, purely affiliating in character and modelled on London University, were established in 1857. The period between World War I and the end of World War II marked a steady but unplanned growth in education and intensification of the struggle for independence. On August 14, 1947, Pakistan emerged as a national sovereign state.

For the new state the initial years proved to be a period essentially of consolidation and exploration. The constitution of Pakistan adopted in 1956 recognized the obligation of the state to provide education among other basic necessities of life, and education was an important element of Pakistan's three five-year plans covering the period 1955–70. As recommended by the Commission on National Education in 1960, the educational system was redesigned in both structure and content. Under the new education policy adopted in March 1970, further reforms were launched to make the educational system more responsive to the goals of national development. Notable among them were measures for a faster growth of scientific, technical, and vocational education and greater autonomy for the institutions of higher education as a fillip to research and creative work. An index to the progress of education during the postindependence period was the rapid rise in enrollment at all levels between 1947–48 and 1969–70, threefold in primary, fivefold in secondary, and nearly 15-fold in higher education.

Muslim and Dutch influences in Indonesia

Indonesia. The educational system in Indonesia was profoundly influenced by the Hindu and Buddhist religions from AD 100 to 1500 and by Islām from 1400 to 1600. The first schools on Western lines were established during the Portuguese and Spanish period (1500–1600). In 1848 the Dutch government for the first time officially assumed responsibility for the provision of education for the native population. The educational policy issued through the Royal Decree of 1892 divided the native primary schools into first- and second-class schools, the former for the Indonesian upper classes and the latter for the masses. The impact of the growth of liberal ideas in Europe could be seen with the advent of the 20th century in the provision for teaching Western science and culture to the native children.

After final independence of the country from the Dutch in 1949, the educational system was substantially remodelled. Article 3 of the provisional constitution of the Republic of Indonesia recognized the right of every Indonesian citizen to education according to his free choice. The Ministry of Education and Culture had carried out a ten-year program of development in education during 1950–60 and issued seven major objectives (*sapta usaha tama*) for educational development in schools and universities, followed in October 1960 by "five principles of Indonesian education" (*pantja wardhena*). The first

Five-Year Plan (1969–70 to 1973–74), adopted in 1968, aimed at linking education to "the needs and possibilities of economic and social development." Notable gains were made between 1950–51 and 1962–63, as indicated by the substantial increase in enrollments—twofold in primary, more than eightfold in secondary, and over tenfold in higher education.

Philippines. The pre-Spanish Philippines possessed a system of writing similar to Arabic, and it was not uncommon for adults to know how to read and write. Inculcation of reverence for the god Bathala, obedience to authority, loyalty to the family or clan, and respect for truth and righteousness were the chief aims of education.

Spanish and American influences in the Philippines

After the Spanish conquest, apart from parochial schools run by missionaries, the first educational institutions to be established on Western lines were in higher education. The Santo Tomás College, established in 1611 and raised to the status of a university in 1644–45, served for centuries as a centre of intellectual strength to the Filipino people. Educational growth, however, was slow, mainly because of lack of government support.

With the advent of the American rule, the stress laid on universal primary education in the policy announced by U.S. Pres. William McKinley on April 7, 1900, led to a rapid growth in primary education. A number of institutions of higher education were also established between 1907 and 1941, including the University of the Philippines (1908). Private institutions of higher education, however, far outnumbered the state institutions, thus indicating a trend that remains a characteristic feature of the system of higher education in the Philippines.

The new Republic of the Philippines emerging after World War II launched a national plan for reconstruction and development of education followed by the five-year socioeconomic program in 1963. The Four-Year National Plan (1971–74) and the recommendations of the Presidential Commission to Survey Philippine Education (1970) envisaged far-reaching reforms to make education "responsive to the challenge of modernization and the goals of national development." During 1950–65 the country made impressive progress in education. The number of technical and vocational schools doubled from 112 to 224 during 1959–69, and 102,070 students were enrolled under the Bureau of Vocational Education programs in 1969–70.

Thailand. The traditional system of education in Thailand was inspired by the Thai philosophy of life

The Thai philosophy

based on (1) dedication to Theravāda Buddhism, with its emphasis on moral excellence, generosity, and moderation, (2) veneration for the king, and (3) loyalty to the family. The beginning of the present system of education can be traced to 1887, when King Chulalongkorn set up a department of education with foreign advisers, mostly English educationists. Gradually, temple schools were established. The process of westernization of education was strengthened with the establishment of a Medical School in 1888, a Law School in 1897, and a Royal Pages' School in 1902 for the general education of "the sons of the nobility." It was converted into the Civil Service College in 1910.

The abolition of the absolute monarchy after the 1932 revolution led to the broadening of higher education with cadres of high-level specialists in various fields. By 1953 the following state institutions of higher education had come into existence: Chulalongkorn University, Thammasat University (of Moral and Political Sciences), the University of Medical Sciences, Kasetsart University for Agriculture and Veterinary Sciences, Silpakorn University for Fine Arts, and the College of Education.

Thai is the medium of education at all levels, but English is taught as a second language from grade five. The educational system is organized by the state, and all registered institutions are under its supervision. Compulsory primary education for four years—introduced by law in 1921 and since raised to seven years—has progressed well, with 5,700,000 children (84 percent of the age group) enrolled in 1969.

Cambodia. For nearly four centuries before the advent of the French in 1863, the educational system in Cambodia grew up around Theravāda Buddhism, which became the established religion toward the end of 1430 under the Thai influence. In 1887 Cambodia became a part of the French Indochina Union and did not achieve complete independence until 1954. Pagoda schools, imparting education at the primary level, were remodelled and integrated into the primary school system administered by the Ministry of Education.

Indochinese education under colonialism and war

Modern higher education is a fairly recent development, dating from 1946, when a school for health officers was established. A number of other institutions of higher education came into being between 1949 and 1958. The Royal Khmer University was inaugurated in 1960, incorporating some of the existing institutions of higher education. A striking feature of later plans had been the increase of expenditures on both higher education and vocational education (in relation to primary education), though plans were frustrated by civil war.

Laos. The pagoda school was the main unit of the traditional educational system in Laos. Efforts toward modernization came in the wake of the country's becoming a French protectorate in 1893 and finally after its inclusion in 1904 within the French Indochina Union. The medium of education was changed to French when the French Education Service was created.

After independence, education developed rapidly, as indicated by the rise in enrollment between 1956 and 1963 from 75,000 to 133,000 in the primary and from 1,500 to 4,000 in the secondary stage. This improvement was made in spite of the loss of schools in the territory controlled by the antigovernment Pathet Lao in the civil war. The rural areas have been served mainly by pagoda schools, village primary schools, and Rural Community Centres (one for every five villages in 1963–64). Primary schools with a full six-year course are for the most part in towns and townships. Plans were made to treble enrollments by 1980.

Vietnam. Long Chinese domination over the emperors of Vietnam resulted in strong Confucian and Taoist influences on the Vietnamese educational system, though it centred on Buddhism. The establishment of French rule, commencing with the occupation of Saigon in 1859, led to the gradual growth of a pattern of education similar to that of the rest of the former Indochina Union. Vietnamese attempts to develop education were thwarted by the continued fighting from World War II onward and, after the partition of the country in 1954, by fighting between the South Vietnamese government forces and the Vietcong. The plans for national development have been tentative in view of the uncertainty created by the war. A notable reform in the educational systems in both North and South Vietnam, however, was the diversification of education and provision of technical and vocational education in secondary and higher levels.

(M.S.H.)

BIBLIOGRAPHY. A very large number of general histories of education have appeared since World War II, and publication of such works seems always to be increasing. A few are regularly reviewed in the *History of Education Quarterly* (1961– ; formerly the *History of Education Journal*, 1949–60), which, however, concentrates mostly on items of interest to Americans. All the general histories seem concerned with the educational history of either one country or the West. Even books in English entitled *A History of Education* tend to centre on Western education; non-Western developments, if treated at all, are usually treated in the initial chapters on ancient civilization or in the very last chapters in terms of the Westernizing of such countries as Japan and India; non-Western events between these extremes are usually left unnoticed. Medieval Muslim education is frequently treated because of its impact upon Western developments.

Given these limitations, the best general histories of education are WILLIAM BOYD and EDMUND J. KING, *The History of Western Education*, 8th ed. (1966); R. FREEMAN BUTTS, *A Cultural History of Western Education*, 2nd ed. (1955); S.E. FROST, *Historical and Philosophical Foundations of Western Education* (1966); HARRY G. GOOD and JAMES D. TELLER, *A History of Western Education*, 3rd ed. (1969); ADOLPH E. MEYER, *An Educational History of the Western World* (1965); JAMES MULHERN, *A History of Education: A Social Interpretation*, 2nd ed. (1959); MEHDI NAKOSTEEN,

The History and Philosophy of Education (1965); I.N. THUT, *The Story of Education: Philosophical and Historical Foundations* (1957); ROBERT ULICH, *The Education of Nations: A Comparison in Historical Perspective*, rev. ed. (1967), and *The History of Educational Thought*, rev. ed. (1968); and JOHN E. WISE, *The History of Education* (1964).

Despite its age, the five-volume *Cyclopedia of Education*, ed. by the American educator PAUL MONROE (1911–13, reprinted 1968), remains one of the most comprehensive and best sources of information on many individual historical topics. Its excellence was early recognized; when the English educator FOSTER WATSON developed his four-volume *Encyclopaedia and Dictionary of Education* (1921–22), he acknowledged the influence of Monroe's opus; Foster's work, incidentally, though largely written by Britishers, had many foreign contributors, including John Dewey and Benedetto Croce. There have since appeared many other national encyclopaedias of historical interest in education. The Japanese educational encyclopaedia *Kyokkugaku Jiten*, 6 vol. (1957–59), gives especially great attention to both Japanese and foreign educational history; regrettably there is no English version.

For historical surveys of individual countries, the following are a few of the useful works: W.H.G. ARMYTAGE, *Four Hundred Years of English Education*, 2nd ed. (1970); S.J. CURTIS, *History of Education in Great Britain*, 5th ed. (1963); CHARLES FOURRIER, *L'Enseignement français*, vol. 1, *De l'antiquité à la révolution* (1964) and vol. 2, *De 1789 à 1945* (1965); NICHOLAS HANS, *History of Russian Educational Policy, 1701–1917* (1931, reprinted 1964); WILLIAM H.E. JOHNSON, *Russia's Educational Heritage* (1950, reprinted 1969); ADOLPH E. MEYER, *An Educational History of the American People*, 2nd ed. (1967); R. FREEMAN BUTTS and LAWRENCE A. CREMIN, *A History of Education in American Culture* (1953); CHARLES E. PHILLIPS, *The Development of Education in Canada* (1957); F. HENRY JOHNSON, *A Brief History of Canadian Education* (1968); TOKIOMI KAIGO, *Japanese Education: Its Past and Present*, 2nd ed. (1968); T.N. SIQUEIRA, *The Education of India*, 4th ed. (1952); and PING WEN KUO, *The Chinese System of Public Education* (1915).

Education in primitive and early civilized cultures: Monographs dealing solely with education in primitive or primordial civilizations are not common; information is to be found chiefly in works treating larger subjects. The best for primitive education are YEHUDI A. COHEN, *The Transition from Childhood to Adolescence* (1964); MARGARET MEAD, *Continuities in Cultural Evolution* (1964); and GEORGE D. SPINDLER (ed.), *Education and Culture: Anthropological Approaches* (1963). An elaborate illustration of primitive education is provided in JOHN W.M. WHITING, *Becoming a Kwoma* (1941). Relatively rich descriptions and interpretations of Egyptian and Mesopotamian education are offered in JAMES MULHERN, *A History of Education: A Social Interpretation*, 2nd ed. (1959); and THOMAS WOODY, *Life and Education in Early Societies* (1949, reprinted 1970). TOM B. JONES, *Ancient Civilization*, rev. ed. (1964), sheds light upon the development of several ancient civilizations, including those of the Near East and Egypt. HOWARD S. GALT, *A History of Chinese Educational Institutions*, vol. 1 (1951); and KWANG-CHIH CHANG, *The Archaeology of Ancient China*, rev. ed. (1971), illuminate the historical background for North China. The important sources for New World civilizations are ABRAHAM ARIAS-LARRETA, *Literaturas aborígenes: Azteca, incaica, maya–quiche* (1951; Eng. trans., *Pre-Columbian Literatures: Aztec, Incan, Maya–Quiché*, 1964); MIGUEL LEON-PORTILLA, *Las literaturas precolombinas de México* (1964; Eng. trans., *Pre-Columbian Literatures of Mexico*, 1969); SYLVANUS GRISWOLD MORLEY, *The Ancient Maya*, 3rd ed. (1956); THEODORE A. WILLARD, *The Lost Empires of the Itzaes and Mayas* (1933); J. ERIC S. THOMPSON, *The Rise and Fall of Maya Civilization* (1954); LOUIS BAUDIN, *La Vie quotidienne au temps des derniers Incas* (1955; Eng. trans., *Daily Life in Peru Under the Last Incas*, 1961); and A. HYATT VERRILL, *Old Civilizations of the New World*, new ed. (1938).

Education in classical cultures: Only a limited number of works deal solely with ancient Asian education. A short, broad survey of events from the dawn of Indian civilization is contained in T.N. SIQUEIRA (*op. cit.*). More specifically for ancient India, D.G. APTE, *Our Educational Heritage: The Schools of Background of Hindu Education* (1961), is a brief, elementary account; and R.K. MOOKERJI, *Ancient Indian Education, or; Brahmanical and Buddhist*, 4th ed. (1969), is a somewhat more sophisticated account for advanced students who are already acquainted with the broad outlines. FRANK E. KEAY, *Indian Education in Ancient and Later Times*, new ed. (1938), surveys education from the Vedic period (on up to the advent of British rule). A detailed description of ancient Chinese education from the earliest times to the 10th century AD is given in HOWARD S. GALT (*op. cit.*); the further development of this scholarly work was cut short by the author's death. PING WEN KUO (*op. cit.*), is a succinct but well-organized account of Chinese education from ancient times (on up to the beginning of the 20th century). Some excellent accounts of education among the ancient Hebrews and Jews are NATHAN DRAZIN, *History of Jewish Education from 515 B.C.E. to 220 C.E.* (1940); ELIEZER EBNER, *Elementary Education in Ancient Israel During the Tannaitic Period (10–20 C.E.)* (1956); and NATHAN MORRIS, *The Jewish School: An Introduction to the History of Jewish Education* (1937). For the ancient West, HENRI I. MARROU, *A History of Education in Antiquity* (Eng. trans. 1956), provides a general survey of Greek and Roman education, but to be preferred to this inadequate English translation of the 1st French edition is the 6th enlarged French edition, *Histoire de l'éducation dans l'antiquité* (1965). Other important general surveys of the West include EDGAR B. CASTLE, *Ancient Education and Today* (1961); JOHN F. DOBSON, *Ancient Education and Its Meaning to Us* (1932); FRANK P. GRAVES, *A History of Education Before the Middle Ages* (1909); S.S. LAURIE, *Historical Survey of Pre-Christian Education*, 2nd ed. rev. (1900, reprinted 1970); and WILLIAM A. SMITH, *Ancient Education* (1955). On special but important topics there are JOHN P. MAHAFFY, *The Progress of Hellenism in Alexander's Empire* (1905); MARTIN P. NILSSON, *Die hellenistische Schule* (1955); OSCAR W. REINMUTH, *The Ephebic Inscriptions of the Fourth Century B.C.* (1971); and MARTIN L. CLARKE, *Higher Education in the Ancient World* (1971). Specifically on Roman schools there are AUBREY O. GWYNN, *Roman Education from Cicero to Quintilian* (1926, reprinted 1966); and THEODORE J. HAARHOFF, *Schools of Gaul: A Study of Pagan and Christian Education in the Last Century of the Western Empire* (1920).

Education in Persian, Byzantine, early Russian, and Islāmic civilizations: One of the best detailed studies of ancient Persian culture and education is MANECKJI N. DHALLA, *Zoroastrian Civilization* (1922). JAMES DARMESTETER (trans.), *The Zend-Avesta*, 2 vol., 2nd ed. of vol. 1 (1895), is excellent for the Zoroastrian foundation of ancient Persian civilization, while A.V.W. JACKSON, *Zoroaster, the Prophet of Ancient Iran* (1899), is a briefer, though still scholarly, survey of Zoroastrianism. One of the best summaries of Byzantine education is GEORGINA BUCKLER, "Byzantine Education," in NORMAN H. BAYNES and HENRY ST. L.B. MOSS (eds.), *Byzantium: An Introduction to East Roman Civilization* (1948). Brief treatments of Byzantine education are also contained in GERVASE MATHEW, *Byzantine Aesthetics* (1963); DEAN A. MILLER, *The Byzantine Tradition* (1966); CHARLES DIEHL, *Byzance, grandeur et décadence* (1919; Eng. trans., *Byzantium: Greatness and Decline*, 1957); and ROBERT BYRON, *The Byzantine Achievement: An Historical Perspective, A.D. 330–1453* (1929, reprinted 1964). A good brief account of early Russian education may be found in THOMAS DARLINGTON, *Education in Russia* (1909). Kievan culture and education are treated in BORIS D. GREKOV, *The Culture of Kiev Rūs* (1947; orig. pub. in Russian, 1944); early Ukrainian education is sketched in MICHAEL S. HRUSHEVSKY, *A History of the Ukraine* (1941; orig. pub. in Russian, 1918). MEHDI NAKOSTEEN, *The History of Islamic Origins of Western Education, A.D. 800–1350* (1964), is an extensive introductory, documented survey of medieval Muslim education and its impact upon Western education. BAYARD DODGE, *Muslim Education in Medieval Times* (1962), is an excellent brief survey, as are B.S. KHUDA, "The Educational System of the Muslims in the Middle Ages," *Islamic Culture*, 1:442–472 (1927); and AHMAD SHALABI, *History of Muslim Education* (1954). Two other significant works are DE LACY E. O'LEARY, *How Greek Science Passed to the Arabs* (1949) and *Arabic Thought and Its Place in History*, rev. ed. (1939, reprinted 1963).

The European Middle Ages: Some of the best surveys of medieval European education are contained in the general histories of education listed at the beginning of this bibliography. On the earlier periods, some useful works are THEODORE J. HAARHOFF, *The Schools of Gaul: A Study of Pagan and Christian Education in the Last Century of the Western Empire*, 2nd ed. (1958); HUGH GRAHAM, *The Early Irish Monastic Schools* (1923); W.G. HANSON, *The Early Monastic Schools of Ireland, Their Missionaries, Saints and Scholars* (1927); and JAMES B. MULLINGER, *The Schools of Charles the Great and the Restoration of Education in the Ninth Century* (1877). For later periods there are two significant works in French: EMILE LESNE, *Les Écoles de la fin du VIII^e siècle à la fin du XII^e siècle* (1940); and PHILIPPE DELHAYE, "L'Organisation scholaire au XII^e siècle," *Traditio*, 5:211–268 (1947). On significant special topics there are SIR JOHN EDWIN SANDYS, *A History of Classical Scholarship from the Sixth Century B.C. to the End of the Middle Ages*, 3rd ed. (1921); REGINALD L. POOLE, *Illustrations of the History of*

Medieval Thought and Learning, 2nd ed. rev. (1920); JAMES W. THOMPSON, *The Medieval Library* (1939); JAMES A. CORBETT (ed.), *The De Instructione Puerorum of William of Tournai* (1955); A.L. GABRIEL, *The Educational Ideas of Vincent of Beauvais* (1956); and CLARA PATRICIA MCMAHON, *Education in Fifteenth-Century England* (1947, reprinted 1968). There are many studies of medieval universities, a few of the most important being HASTINGS RASHDALL, *The Universities of Europe in the Middle Ages*, new ed., 3 vol. (1936); ROBERT S. RAIT, *Life in the Medieval University* (1912, reprinted 1969); CHARLES H. HASKINS, *The Rise of Universities* (1923; paperback ed., 1957); and A.L. GABRIEL, *Garlandia: Studies in the History of the Medieval University* (1969).

Education in Asian civilizations, c. 700 to the eve of Western influence: India's medieval, or Muslim, period is well covered in three works: S.M. JAFFAR, *Education in Muslim India* (1936), is a vivid account based on contemporary chronicles and other documents; N.N. LAW, *Promotion of Learning in India During Muhammadan Rule* (1916), surveys the efforts of Muslim rulers to promote learning in India and contains very useful illustrations; MANSOORUDDIN QURAISHI, *A Survey of Muslim Education in the Middle Ages* (1956), in focussing on the advent of the Muslims in India and on their dissemination of Islāmic teaching, relates how this Muslim education acquired an Indian character. For education in China, EDWARD A. KRACKE, *Civil Service in Early Sung China, 960–1067* (1953), may be added to the books of HOWARD S. GALT (*op. cit.*), for the period up to the 10th century; and PING WEN KUO (*op. cit.*). For general surveys of the history of Japanese education, the chief works in English are TOKIOMI KAIGO, *Japanese Education: Its Past and Present*, 2nd ed. (1968); and the MINISTRY OF EDUCATION (JAPAN), *An Outline History of Japanese Education, Literature and Arts* (1876). RONALD P. DORE, *Education in Tokugawa, Japan* (1965), provides good coverage of the period immediately prior to the modern.

European Renaissance and Reformation: Three works in English by WILLIAM HARRISON WOODWARD are useful introductions to Renaissance education: *Studies in Education During the Age of the Renaissance, 1400–1600* (1906, reprinted 1965), *Vittorino da Feltre and Other Humanist Educators* (1897, reprinted 1964), and *Desiderius Erasmus Concerning the Aim and Method of Education* (1904, reprinted 1964). HANS BARON, *The Crisis of the Early Italian Renaissance*, 2 vol. (1955; rev. ed., 1 vol., 1966); STANFORD E. LEHMBERG, *Sir Thomas Elyot: Tudor Humanist* (1960); and FOSTER WATSON, *Vives on Education: A Translation of the De tradendis disciplinis* (1913), are other important treatments of the Renaissance. In Italian, two excellent works by EUGENIO GARIN are *L'Educazione in Europa, 1400–1600* (1957), which covers the Humanistic movement, the Renaissance, and the Reformation and Counter-Reformation in Europe generally, and *L'Educazione Umanistica in Italia* (1949), an earlier work covering the effects of the Humanistic movement on education in Italy, where the movement started. JOHN LAWSON, *Mediaeval Education and the Reformation* (1967), contains an analysis of the transformation of the schools and universities as a result of the Reformation, especially in terms of the new curricula. Other important works on the Reformation and Counter-Reformation are FREDERICK EBY (ed.), *Early Protestant Educators: The Educational Writings of Martin Luther, John Calvin, and Other Leaders of Protestant Thought* (1931, reprinted 1971); F.V.N. PAINTER, *Luther on Education* (1889, reprinted 1965); ALLAN P. FARRELL, *The Jesuit Code of Liberal Education: Development and Scope of the Ratio Studiorum* (1938); EDWARD A. FITZPATRICK (ed.), *St. Ignatius and the Ratio Studiorum* (1933); and HOWARD C. BARNARD, *The Port-Royalists on Education* (1918).

European education in the 17th and 18th centuries: The general histories of education cited at the beginning of this bibliography offer good accounts of the 17th, 18th, and 19th centuries, when educational theory flourished as it never had before and when educational institutions proliferated. For the 17th century, an additional useful work is JOHN WILLIAM ADAMSON, *Pioneers of Modern Education, 1600–1700* (1905). Some major theories are also covered in MATTHEW SPINKA, *John Amos Comenius, That Incomparable Moravian* (1943); WILL S. MONROE, *Comenius and the Beginnings of Educational Reform* (1900, reprinted 1971); and JOHN WILLIAM ADAMSON (ed.), *The Educational Writings of John Locke* (1912). HOWARD C. BARNARD, *The French Tradition in Education, Ramus to Mme. Necker de Saussure* (1922, reprinted 1970), covers France down to the 18th century. For the 18th century itself, useful introductions are NICHOLAS A. HANS, *New Trends in Education in the Eighteenth Century* (1951); DAVID SALMON, *The Education of the Poor in the Eighteenth Century* (1908); FRANCOIS DE LA FONTAINERIE (ed. and trans.), *French Liberalism and Education in the Eighteenth Century* (1932); and HERBERT MCLACHLAN, *English Education Under the Test Acts; Being the History of the Nonconformist Academies, 1662–1820* (1931). For information on the major theorists there are AUGUSTE PINLOCHE, *La Réforme de l'éducation en Allemagne au dix-huitième siècle: Basedow et le philanthropinisme* (1889); WILLIAM BOYD, *The Educational Theory of Jean Jacques Rousseau* (1911, reprinted 1963), and *The Minor Educational Writings of Jean Jacques Rousseau* (1911, reprinted 1962); and RICHARD L. ARCHER (ed.), *Rousseau on Education* (1912). The best studies available on colonial Latin America are JOSE LUIS BECERRA LOPEZ, *La organización de los estudios en la Nueva España* (1963); and JOHN T. LANNING, *Academic Culture in the Spanish Colonies* (1940, reprinted 1971).

Western education in the 19th century: As already noted, the general works listed at the beginning of this bibliography treat the 19th century in excellent fashion. There are useful specialized works, however. The best works on the theorists are KATE SILBER, *Pestalozzi: The Man and His Work*, 2nd ed. (1965); WILLIAM H. KILPATRICK, *Froebel's Kindergarten Principles Critically Examined* (1916); CHRISTIAN UFER, *Vorschule der Pädagogik Herbarts*, 5th ed. (1894; Eng. trans., *Introduction to the Pedagogy of Herbart*, 1901); FRANK H. HAYWARD, *The Meaning of Education As Interpreted by Herbart* (1907); and E.I.F. WILLIAMS, *Horace Mann, Educational Statesman* (1937). In this century of German pre-eminence in education, it is worthwhile to consult WILHELM H. LEXIS, *A General View of the History and Organisation of Public Education in the German Empire* (1904); WILLY MOOG, *Geschichte der Pädagogik*, vol. 3, *Die Pädagogik der Neuzeit vom 18. Jahrhundert bis zur Gegenwart* (1933); and FRIEDRICH PAULSEN, *Das deutsche Bildungswesen in seiner geschichtlichen Entwicklung* (1906; Eng. trans., *German Education, Past and Present*, 1908), this last being a classic analysis. For England there are JOHN W. ADAMSON, *English Education, 1789–1902* (1930); and R.L. ARCHER, *Secondary Education in the Nineteenth Century* (1921). For 19th- and early-20th-century Russia, there is PATRICK L. ALSTON, *Education and the State in Tsarist Russia* (1969). Two interesting specialized works are A.G. AUSTIN, *Australian Education, 1788–1900* (1961); and ARTHUR G. BUTCHERS, *Young New Zealand* (1929). For India during the era of British rule, there are SYED NURULLAH and J.P. NAIK, *A History of Education in India During the British Period*, 2nd ed. rev. (1951); S.N. MUKERJI, *History of Education in India: Modern Period*, 5th ed. (1966); and B.D. SRIVASTAVA, *The Development of Modern Indian Education*, 2nd rev. ed. (1964). The works of TOKIOMI KAIGO and the MINISTRY OF EDUCATION (JAPAN), already cited for Japanese education, are useful here.

Education in the 20th century: All the works listed at the beginning of this bibliography contain surveys of 20th-century practices and theories. ADOLPH E. MEYER, *The Development of Education in the 20th Century*, 2nd ed. (1949), treats the development of education in a generalized story form and is one of the better books on the history of education in the 20th century. Good descriptions of the various contemporary philosophies of education may be found in JOHN PAUL STRAIN (comp.), *Modern Philosophies of Education* (1971); THEODORE B.H. BRAMELD, *Patterns of Educational Philosophy* (1971); and JOHN S. BRUBACHER, *Modern Philosophies of Education*, 4th ed. (1969). Although otherwise wholly oriented toward American institutions and practices, LEE C. DEIGHTON (ed.), *The Encyclopedia of Education*, 10 vol. (1971), does contain individual articles on the nations of the world and, in these articles, offers a measure of recent history. More detailed surveys, often with a sociological slant, are being offered in the series of small volumes being issued by the Commonwealth and International Library—each, though separately authored, with the title *Society, Schools, and Progress in*, followed by the name of the country or region (*e.g.*, *Society, Schools, and Progress in Australia; Society, Schools, and Progress in Eastern Europe*); over a dozen volumes have appeared so far on systems, for example, in Australia, Canada, China, eastern Europe, England, France, India, Israel, Nigeria, Peru, Scandinavia, Tanzania, the U.S., and the West Indies. An excellent cross-national comparison is to be found also in JOHN F. CRAMER and GEORGE S. BROWNE, *Contemporary Education: A Comparative Study of National Systems*, 2nd ed. (1965).

Works on individual countries and most regions are legion; only a sample can be cited here. For Europe there are EDMUND J. KING, *Other Schools and Ours*, 3rd ed. (1967), and *Education and Development in Western Europe* (1969); URSULA K. SPRINGER, *Recent Curriculum Developments in France, West Germany, and Italy: A Study of Trends at the Middle Level of Education* (1969); RAYMOND POIGNANT, *L'Enseignement dans les pays du Marché commun, étude comparative sur l'organisation et l'état de développement de l'enseignement dans les pays du Marché commun, aux États-Unis, en Grande-Bretagne et en U.R.S.S.* (1965; Eng. trans., *Education and Development in Western Europe, the United States, and the*

U.S.S.R.: A Comparative Study, 1969); WILLIAM O. LESTER SMITH, *Education in Great Britain*, 5th ed. (1967); JEAN CAPELLE, *L'École de demain reste à faire* (1966; Eng. trans., *Tomorrow's Education: The French Experience*, 1967); JOACHIM H. KNOLL, *The German Educational System* (1967); and HERMAN RUGE, *Educational Systems in Scandinavia* (1962). For the United States there are LAWRENCE A. CREMIN, *The Transformation of the School: Progressivism in American Education, 1876–1957* (1961); WILLIAM E. DRAKE, *The American School in Transition* (1955); NEWTON EDWARDS and HERMAN G. RICHEY, *The School in the American Social Order*, 2nd ed. (1963); and S. ALEXANDER RIPPA, *Education in a Free Society*, 2nd ed. (1971). For elder members of the British Commonwealth, there are, for Canada, FRANK MacKINNON, *The Politics of Education* (1960); and FREEMAN K. STEWART, *Interprovincial Co-operation in Education* (1957); for Australia, UNESCO, *Compulsory Education in Australia* (1962); and for New Zealand, ARTHUR G. BUTCHERS, *Education in New Zealand* (1930); and WILLIAM B. SUTCH, *The Quest for Security in New Zealand, 1840 to 1966* (1966). For education under Communism, two general treatments are WASYL SHIMONIAK, *Communist Education: Its History, Philosophy, and Politics* (1970); and EDMUND J. KING (ed.), *Communist Education* (1963). For the Soviet Union, two important analyses in German have been written by OSKAR ANWEILER: *Geschichte der Schule und Pädagogik in Russland, vom Ende des Zarenreiches bis zum Beginn der Stalin-Ära* (1964), which covers the 20th century up to the early 1930s, and *Die Sowjetpadägogik in der Welt von heute* (1968), which focusses largely on the 1950s and 1960s. Other important works, in English, are JAAN PENNAR, IVAN I. BAKALO, and GEORGE Z.F. BEREDAY, *Modernization and Diversity in Soviet Education, with Special Reference to Nationality Groups* (1971); GEORGE Z.F. BEREDAY, WILLIAM W. BRICKMAN, and GERALD H. READ (eds.), *The Changing Soviet School* (1960); GEORGE Z.F. BEREDAY and JAAN PENNAR (eds.), *The Politics of Soviet Education* (1960); and NIGEL GRANT, *Soviet Education* (1964). For education in 20th-century China prior to Communist rule, the most useful works are KNIGHT BIGGERSTAFF, *The Earliest Modern Government Schools in China* (1961); WOLFGANG FRANKE, *The Reform and Abolition of the Traditional Chinese Examination System* (1960); and THEODORE E. HSIAO, *The History of Modern Education in China* (1932). Post-1949 education in mainland China is well covered in R.F. PRICE, *Education in Communist China* (1970); and THEODORE H.E. CHEN, *The Thought Reform of the Chinese Intellectuals* (1960). Three works on Japanese education, which largely relate it to socio-economic development, are the MINISTRY OF EDUCATION (*op. cit.*), the JAPANESE NATIONAL COMMISSION FOR UNESCO, *The Role of Education in the Social and Economic Development of Japan* (1966); and HERBERT PASSIN, *Society and Education in Japan* (1965). For India two works by S.N. MUKERJI are especially useful: *Education in India, Today and Tomorrow*, 6th ed. (1969), which focusses on current problems and policies, and *History of Education in India: Modern Period*, 5th ed. (1966), which treats mostly of the British period. To these may be added B.D. SRIVASTAVA, *The Development of Modern Indian Education*, rev. ed. (1963). Racial and linguistic divisions in South African education are treated in S. PIENAAR and ANTHONY SAMPSON, *South Africa: Two Views of Separate Development* (1960). Nation building in Africa has elicited several works on education and development in Africa, past and present; the most significant recent works are ABDOU MOUMOUNI, *L'Éducation en Afrique* (1964; Eng. trans., 1968), which largely concentrates on former French colonies; JOHN CAMERON, *The Development of Education in East Africa* (1970), which is concerned with former British colonies; K.A. BUSIA, *Purposeful Education for Africa* (1964); JOHN W. HANSON, *Imagination and Hallucination in African Education* (1965); and L. GRAY COWAN, JAMES O'CONNELL, and DAVID G. SCANLON (eds.), *Education and Nation-Building in Africa* (1965). The best surveys of the Arab countries are RODERIC D. MATTHEWS and MATTA AKRAWI, *Education in Arab Countries of the Near East* (1949); ABDEL AZIZ EL-SAYED, "Education in the Arab World," *Mid East*, vol. 7 (1967); and FAHIM I. QUBAIN, *Education and Science in the Arab World* (1966). Historical surveys of education in Latin America are relatively poor, except for those dealing with certain countries, such as Mexico. The most useful works are ROBERT J. HAVIGHURST, *La sociedad y la educación en América Latina* (1962); HAROLD R.W. BENJAMIN, *Higher Education in the American Republics* (1965); JOSEFINA VAZQUEZ DE KNAUTH, *Nacionalismo y educación en México* (1970); and ROBERT J. HAVIGHURST and JOAO ROBERTO MOREIRA, *Society and Education in Brazil* (1965). Surveys of education in Southeast Asia are even scarcer, the best studies being MUHAMMAD SHAMSUL HUQ, *Education and Development Strategy in South and Southeast Asia* (1965), which analyzes planned educational development in Southeast Asia with special reference to Pakistan, India, Indonesia, and the Philippines; and UNESCO and the INTERNATIONAL ASSOCIATION OF UNIVERSITIES, *Higher Education and Development in Southeast Asia*, 3 vol. (1967), which contains useful background materials in the form of country profiles.

Education, Philosophy of

Philosophy of education is a field of inquiry, speculation, and application in which philosophical methods are applied to the study of a problem, topic, or issue in education. Characteristic of these methods is the attempt to think as accurately, clearly, coherently, and systematically as possible. Analytical philosophers would say that philosophy of education should end with the attempt to clarify and justify educational statements and arguments. In practice, however, the field includes much more than that. It considers as relevant material much that has been written on education by influential philosophers of the past. There is, therefore, much overlap with the field of history of education, and ultimately no clear demarcation can be drawn between the two fields, which nourish and illuminate each other in their interconnectedness (see EDUCATION, HISTORY OF). Many philosophers of education also go beyond analysis in being concerned with establishing a commitment to value judgments and substantive positions. They take pains to attempt to clarify and justify those judgments and positions on the grounds that clear and substantiated judgments have greater probability of being sound.

Different conceptions of the philosophy of education

The philosophy of education has a special concern with the *applications* of knowledge and theories. Thus, many philosophers of education are especially interested in the relationship between theory and practice. Moreover, they are often concerned with the ways in which philosophy relates to other fields of study in the attempt to shed light on educational problems and issues. This gives them a wide-angled approach to education, which some philosophers have called "educational theory" to distinguish it from a more narrowly analytical form of philosophy of education.

With regard to the term "education," there is a similar spectrum of views ranging from the narrow view of education as that which goes on in schools and universities to a definition of education as all those experiences that affect the growth and development of a person throughout life. The former view has tended to give way, with increased recognition of the crucially important part that informal experiences and relationships play in determining what and how an individual learns.

COMMON ASSUMPTIONS ABOUT HUMAN LEARNING

Symbolization. First, it is to be noted that human learning is largely dominated by symbolic processes. Much of the learning that man acquires during his lifetime is gained through his growing ability to understand and manipulate symbols—verbal, mathematical, artistic, musical, and so on. This symbolic emphasis gives human learning much of its power. It brings power because symbols enable man to deal abstractly and at a high level of generality with words, data, and ideas. Thus, he can learn to think rapidly, to link varied items in fruitful ways, and to create shortcuts to new knowledge.

Dangers of abstraction

The symbolic emphasis also brings dangers. It can trap man into circling around at a high level of generality without ever feeling the need to tie his abstractions to concrete applications. Schooling thus can become an apparently self-justifying cycle, which in fact fails to pay off in terms of improving men's lives. Moreover, since learners assimilate and master these symbolic processes at different rates, it is tempting for educators to identify all education with symbol manipulation, to classify those who take longer to master the process as inferior, and to reserve the privileges of higher schooling for those who master this particular process rapidly. This is to mistake a part for the whole and leads to enormous losses of human power and development.

Human motivation. Motivations to learning are more variable in man than in subhuman species. In other animals learning patterns develop largely through instinctual

motivations. This means that lower animals need less training than humans before they are independent and self-reliant. Human beings are the most dependent of creatures when young, and they take longest to educate for independent activity. It is precisely this long period of dependence and education that constitutes man's superiority over lower animals, for it is in this period that the growing human being absorbs much of his culture and develops the skills and knowledge that enable him to build on the work of his predecessors. His lower degree of reliance on innate mechanisms also means that man is more flexible in his responses and is capable of adjusting those responses in the light of previous experiences. Thus, he is not condemned to repeat previous patterns of living and thinking but can create change, both in himself and in his culture.

Educators have often acted as if man were a lower animal, for schooling has often resembled training, with emphasis upon imitation, obedience, repetition, drill, and control, more than it has resembled education, with emphasis upon initiative, creativity, choice, decision making, and freedom. As a result, schooling has been more successful in improving man's skill in mechanical and repetitive activities at the lower end of his capacity scale than in developing higher qualities of intellect, feeling, and will that constitute his peculiar humanness.

Human capacity to learn. So far there have been discovered no limits to man's capacity to learn. From earliest times, however, men in positions of power or influence have suggested that the learning capacity of certain individuals or groups is severely limited and that they should not be expected to profit greatly, if at all, from education. These "ineducable" individuals have usually been members of minority or disadvantaged groups. But, repeatedly, when their cultural disadvantages have been removed, these groups have shown that their previous failure to learn has been due not to incapacity but to lack of fully realized opportunity.

Categorizing learners

These findings have led educators to be much more modest and less hasty in their labelling and classifying procedures. It has been realized that labels affixed to children tend to become self-fulfilling prophecies, that those who are expected to learn usually do so, and those who are expected to fail to learn also usually do so. Hence, when educators resort to classifying children at all, they increasingly tend to use their labels as temporary rather than permanent, as saying something only about a quality of the child rather than about his person, and as something to be abandoned as soon as the child's performance proves the label wrong.

Similarly, no one has been able to confirm any certain limits to the speed with which man can learn. Schools and universities have usually been organized as if to suggest that all students learn at about the same rather plodding and regular speed. But, whenever the actual rates at which different people learn have been tested, nothing has been found to justify such an organization. Not only do individuals learn at vastly different speeds and in different ways, but man seems capable of astonishing feats of rapid learning when the attendant circumstances are favourable. It seems that, in customary educational settings, one habitually uses only a tiny fraction of one's learning capacities.

Complexity of human learning. Human learning is complex rather than simple. Learners are apt to learn more than one thing at a time. Sometimes this process is conscious, as when one simultaneously or rapidly assimilates many specific items of a whole. More often, the process is entirely or partly unconscious, as when the student learns some "content" consciously but at the same time absorbs unwittingly a great deal more from interrelationships, tones of voice, and so on.

Correlative learning

Educators are therefore becoming increasingly concerned with these concomitant learnings. They are aware that the long-term significance of the arithmetical skill that the student consciously learns may be nugatory compared with the importance of what he learns about himself as a learner, about his capacities and limits, about his relationship with his teacher, about power and authority, about his relationships with his fellow students, about equality, collaboration, competition, and friendship. As educators become more knowledgeable about the importance of learning climates, they are impelled to abandon simplified techniques of teaching in favour of a more complex approach that views learning in the context of a matrix of relationships and forces that act upon the student, the teacher, the school, and the community.

Intellectual–emotional–physical–volitional integrity of learning. Human learning concerns the whole person. The intellect is not the only agent of learning. This activity is shared by the body, the emotions, and the will. Moreover, the process cannot be limited to any one of these domains without affecting the others. Educators are most conscious of intellectual learnings, which tend to play the largest part in their plans and intentions. But there is increasing evidence that makes clear the folly of attempting to confine education to the training of intellects. If the teacher does so, he is destined to fare badly, for the child who is emotionally frozen or whose stomach is empty or who is determined to thwart the teacher will not perform intellectually as the teacher intends.

Educators are also becoming more aware of the other side of this coin—that is, that the learner's powers are vastly enhanced when not only is his intellect stimulated but also his feelings are respected, his body is nurtured, and his will to learn is strengthened. Effective education, therefore, is found when the learner is regarded as a person to be respected, nurtured, strengthened, and stimulated, rather than as an intellect to be trained.

DIFFFERING CONCEPTIONS OF EDUCATION

Model theories of the educated person. In different historical eras and in different cultures a variety of conceptions has been put forward concerning the nature of mind, of truth, of knowledge, of moral goodness, of aesthetic beauty, and of the purposes of life. These views carry implications for the purpose and nature of the educational process, the ends toward which one should educate, the means by which one can best achieve those ends, the degree of consistency between the ends and the means, and the model, if any, of the ideally educated person that one has in mind when engaging in the educational process. These relate to some of the perennial questions that concern the philosopher of education. The following represent some of the most influential models that have been presented of the educated person.

Platonic view. The first great attempt to create a philosophically coherent model of the educated person was that of Plato. Reacting against the turbulence and chaos of his own times, Plato envisaged in *The Republic* and the *Laws* a permanent, stable, hierarchical society in which those most adept at education would rule, those moderately adept would become warriors and carry out the orders of the rulers or guardians, and those least adept at education would fill the lowliest worker functions in society. The education that would thus order and segregate the different groups in society would be based upon the ability to understand the perennial, ideal Forms that, Plato maintained, underlie the transient, concrete manifestations of those Forms in everyday life. Since, supposedly, not all people could surmount the difficulties of such an intellectually demanding program of study, it would be justifiable to select, classify, segregate, and reject students at the various stages of the educational process.

Intellectually segregated education

The Platonic scheme has been enormously influential, especially in the Western world, during the last 25 centuries. Most educational programs in the West have consistently reflected this hierarchical pattern. Schooling has usually been a process of selection and rejection, with great effort and ingenuity expended on testing, measuring, classifying, and segregating in accordance with the best available knowledge. Similarly, Plato's concept of mind has exerted lasting influence on Western education. Patterns of schooling have reflected the greater prestige accorded to the study of ideas and abstractions and the lower prestige given to practical studies and manual

work. There has persisted a dichotomy between the so-called liberal arts, which have been considered suitable educational fare for potential leaders of society, and so-called vocational studies, which have been considered more suitable for potential followers. This pattern has lasted to the present day, not only in Europe and North America, but also in those parts of Asia, Africa, and Latin America greatly influenced by European culture. Thus, one finds in many former colonial societies that students often strongly prefer to follow "liberal" studies, even though the choice may condemn them to genteel unemployment in a society whose economy cries out for people with technological, practical, and manual skills.

Thomist view. In the 13th century, St. Thomas Aquinas made a monumental attempt to reconcile the two great streams of the Western tradition. In his teaching at the University of Paris and in his writings—particularly the *Summa theologiae* and the *Summa contra gentiles*—Aquinas tried to synthesize reason and faith, philosophy and theology, university and monastery, activity and contemplation. In his writings, however, faith and theology ultimately took precedence over reason and philosophy because the former were presumed to give access to truths that were not available through rational inquiry. Hence, Aquinas started with assumptions based on divine revelation and went on to a philosophical explication of man and nature. The model of the educated man that emerged from this process was the Scholastic, a man whose rational intelligence had been vigorously disciplined for the pursuit of moral excellence and whose highest happiness was found in contemplation of the Christian God.

This Scholastic model has greatly affected the development of Western education, especially in fostering the notion of intellectual discipline. Aquinas' theological–philosophical doctrine has been a powerful intellectual force throughout the West and has constituted the official basis of Catholic theology since 1879. Although Aquinas made an important place in his hierarchy of values for the practical uses of reason, later Thomists have often been more exclusively intellectual in their educational emphasis. For Aquinas, the primary agent of education was the learner, and his model was, thus, a person capable of self-education. Intellectually autonomous, he should be able to conduct his own process of research and discovery. The Roman Catholic Church, however, has usually put the learner firmly under the authoritative superordination of the teacher.

Lockean view. John Locke, the 17th-century English philosopher, has been credited with formulating the classical liberal defense of individual freedom against the authorities of state and church. Opposed both to what he deemed the stagnation of unreflective tradition and the perils of enthusiastic radicalism, Locke saw science, reason, and experience as the best safeguards against these dangers. Responding to the rise of the new bourgeoisie and the new science, he became the principal spokesman for the increasingly powerful middle class, who were predominantly skeptical and practical in their intellectual temper. The model of the gentleman had traditionally been the English ideal of the educated person. Locke's achievement was to take this ideal and modify it to a form acceptable to the new bourgeoisie. Originally an aristocratic model, the gentleman ideal became infused, under Locke's influence, with democratic, Puritan, and practical characteristics.

Education of the practical gentleman

Locke's notion of the mind at birth as a *tabula rasa*, a "blank tablet" devoid of innate ideas, gave enormous importance to the role of experience and sense perception in the educational process. In *Some Thoughts Concerning Education* (1693) and the *Conduct of the Understanding* (1706), Locke outlined the heavily experiential education that would be appropriate for a gentleman. His four cardinal aims of education, in order of importance, were virtue, wisdom, breeding, and learning. Some critics have insisted that the order has remained important to the present day in English education, that learning must always be lightly worn and never ostentatious. Not only was Locke's thinking influential in shaping the subsequent development of English educational thought, but his Puritan individualism also had a considerable effect on the growth of American educational ideals. Locke's failure to recognize the possibilities of the uses of institutional power and legislation for interventions that would enhance rather than restrict freedom was an omission that has often bedevilled American educational thinking.

Naturalistic view. Most movements in philosophy and education have seemed to engender their own opposition. Hence, the rationalism and scientific objectivity of the Enlightenment found a reaction in the subjectivity and emotional spontaneity of romantic naturalism, the principal spokesman of which was Jean-Jacques Rousseau. In particular, Rousseau reacted against the excessive formalism and rationalism of 18th-century France. Out of this reaction there came a model of the educated person as the natural man, a figure presented in contrast to what Rousseau saw as the pathetic products of contemporary civilization. Against civilized values like rationalism, conscious reflection, control, complexity, and objectivity, Rousseau offered his own values of romanticism, intuitive spontaneity, freedom, simplicity, and subjectivity.

In *Émile* (1762), one of the most influential books on education ever written, Rousseau argued that one should protect the child from the corruptions of civilization and carefully nurture his natural, spontaneous impulses, which were always healthy. It was important, he maintained, to avoid premature intellectualization of emotion so that the child's intellectual powers could develop without distortion. Feeling should precede thinking, and the child should be controlled only by things, not by adults' wills. In these ideas lay some of the germs of progressive education, which spread throughout the world during the 19th and 20th centuries. Constantly in tension with the classical demands of reason, discipline, authority, and scholarship, the romantic naturalism of the progressive-education movement has continued to remind educators that their ultimate concern should be the growth of the unique, ultimately unfathomable child.

Subjectivity and spontaneity in learning

Marxist view. Probably the most influential writer of recent times was Karl Marx, whose writings, mostly in collaboration with Friedrich Engels, altered the course of history and continue to affect the lives of millions of people in all parts of the world. A central concern of Marx and Engels was to cure the alienation and dehumanization of man caused by what they saw as the exploitative forces of capitalism. In Marx's writings, the material dimension of history appeared as primary. Economic production was deemed the basis of life, and the prevailing ideas (religious, educational, and political) of a society were seen as being determined by its economic structure. The dominant ideas of an epoch or society were considered to be the ideas of its ruling class—that is, the class that controlled the means of material production. In order to find the hidden interest behind an idea, Marx argued, one had to examine its social function rather than its intellectual content. Marx saw the need for a proletarian revolution in order to bring about Communism. Under Communism, he argued, the opposition between the individual and the group would disappear; each man's interests would be seen to be identical with the interests of all, and alienation would be banished.

In Marx's view, what was needed for man's growth toward maturity was genuine community; that is, the voluntary drawing together of autonomous and socially responsible persons. The model of the educated person that Marx put forward was not the irresponsible individualist nor the coerced collectivist but the accountable communal man, who attained his freedom not by fleeing from social relationship but *through* social relationships. Individual freedom required social authority.

Marxist emphasis on social education

Marx's and Engels' writings, particularly *Das Kapital* (1867–94), *The German Ideology* (London ed., 1938) and *The Communist Manifesto* (1848), have reshaped the world. Millions have been inspired by their vision of unalienated work and education. Unfortunately, like many influential ideas, they have been abused and distorted.

Several states have tried to bring about a Communist utopia through collectivist coercion and social manipulation in a distortion of Marx's arguments. There is a need for state intervention to remove gross inequalities and to expand opportunities, in Marx's view, but ultimately human regeneration is a task for each individual.

Pragmatist view. Out of the America that was created by immigration, urbanization, and industrialization in the second half of the 19th century came the philosophy of Pragmatism. Associated with such thinkers as C.S. Peirce, William James, and John Dewey, Pragmatism, as the dominant American philosophy, exerted a strong influence on the shape of education in the United States, and affected educational ideas and practices in Europe and Japan. In the hands of Dewey, Pragmatism evolved into a philosophy that saw man as formed through interaction with his natural and social environment. The educated person was always viewed by Dewey in a social context. Neither the individual nor society had any meaning without the other. Dewey created a model of the educated person as the reflective man, one who was critical of the authority of custom and tradition as the determinant of belief and action and who preferred the method of science, of "organized intelligence" as the best way to solve his problems.

Progressive emphasis on blending individual and social needs

In *My Pedagogic Creed* (1897), *How We Think* (1910), *Democracy and Education* (1916), *Human Nature and Conduct* (1922), and a stream of other philosophical and pedagogic writings, Dewey formulated a viewpoint that constituted a rigorous intellectual core of the progressive-education movement, although he criticized many of the manifestations of progressive pedagogy in practice. For example, he viewed the child's interests as vitally important and to be neither repressed nor humoured. To repress them was to commit the fault of much traditional education by ignoring the child's unique bent. But to humour them was to commit the fault of some progressive education by failing to discover the underlying power below the passing whim. Subject matter, Dewey argued, should consist of activities that enabled the child to reflect upon his social experiences. When subject matter preceded or was unrelated to the child's experiences, it was largely meaningless. It gained meaning through being made the medium for continued reflection upon, and reconstruction of, experience.

Behaviourist view. In the 20th century, the use of science has been extended to the study of virtually all aspects of human affairs. The possibilities of scientific control of men and events have brought profound changes in philosophy and education. By studying only the behavioral aspects of man, science has been able to predict and control in ways that have powerful and sometimes frightening implications. The behaviourist view has been most notably represented by the American psychologist B.F. Skinner. In his writings, including the utopian novel, *Walden Two* (1948), *Science and Human Behavior* (1953), and *Beyond Freedom and Dignity* (1971), he has firmly rejected the conventional model of man as a free agent who acts in accordance with the decisions of an inner self that is neither fully explicable nor fully controllable by scientific means. Instead, Skinner envisages the use of scientific knowledge about the control of human behaviour to create a planned man, one who will be conditioned to behave in the way best calculated to achieve society's goals. Behavioral engineering will have removed all of his antisocial tendencies, and he will want only what is good for himself and his society.

Programmed learning and behavioral development

Behaviourism in general and Skinner in particular have exerted considerable influence on Western, and especially American, education through developments that range from the invention of programmed instruction (or the so-called "teaching machine") to the widespread emphasis on behavioral objectives in educational programs. The philosophy of Behaviourism compels one to examine carefully the issue of control in education. Skinner wants to use scientific control to bring about a society in which it will be easy to be good and to bring about an educational process through which it will be easy to be excellent. There is no alternative to control, in his view. It is simply a matter of *who* is to control. One does not grant the child "freedom" merely by leaving him alone. To refuse to use scientific control to shape human behaviour is, for the Behaviourist, a failure in responsibility.

Existentialist view. The application of science to ever more aspects of the study of man has been particularly marked in the West; but, in reaction to this trend, there has developed some criticism of the scientific way of viewing man as an object to be categorized, studied objectively, or subsumed under a generalization. Prominent in this reaction have been Existentialist philosophers, among whom the Jewish philosopher, Martin Buber, played an outstanding role in deepening contemporary understanding of man and education. In such books as *I and Thou* (1923) and *Between Man and Man* (1947), Buber's model of the educated person appeared as one whose life was shaped by existential decision making. Such a person did not determine choices in advance of existential situations. He used principles and traditions only as checks or reminders, not as infallible guides. His values were created in the concrete here and now and were manifested as he related to other men. Thus, each man was seen as a unique person rather than as a member of a category.

Dialogue between learning and life

A vital concept for Buber was responsibility, viewed in terms of one's response to another. Thus, the dialogue became a central focus in his educational philosophy. The educated person was one who could listen as well as talk. And since genuine dialogue depended upon authenticity, upon *being* rather than *seeming*, one needed the courage to be oneself in relationships. Buber also urged the recognition of a continuity between learning and life, rather than the encouragement of knowledge for its own sake. He insisted that learning be related to consequent action. Thus, for Buber, the educated person was not one who merely had had his cognitive faculties trained but one whose inmost spirit had been infused by what he had learned.

Areas of disagreement among educational theorists. *Conceptions of what constitutes knowledge.* Although there is wide agreement that the assimilation of knowledge by the learner is a principal goal of education, this agreement is more apparent than real, because there is much disagreement about what constitutes knowledge. Clearly, if one person identifies knowledge with information, his pedagogic approach is likely to differ from that of one who identifies knowledge with a process of thinking. Some philosophers have attempted to clarify a distinction between "knowing that" (something is true or false) and "knowing how" (to do or learn to discover something). The former emphasis is common in educational programs that focus on the product—that is, on the information that the learner has amassed and can reproduce at the end of his program. The latter emphasis is common in programs that focus on the process—that is, on the skills and attitudes that the learner has adopted that enable him to learn more.

Apprenticeship, teaching, and dialogue

Conceptions of how knowledge should be communicated. Throughout history, the commonest way for knowledge to be communicated has been through what can be loosely called apprenticeship. The one who does not know (the apprentice) watches the one who does (the master); he imitates the master, probably fails, watches again, tries again, and so on until he knows what the master knows. This is primarily the way young people learn role identity, for example—what it means to be a husband or wife, a father or mother, a male or female. It is not an effective method for stimulating the growth of new knowledge. A second method, one that has been time-honoured in institutions of formal schooling, is telling. The one who knows (the teacher) tells the one who does not (the student). The student listens, tries to remember, and is usually required to reproduce his memorizations at some stage so that the teacher can judge whether or not the knowledge has been assimilated. This method is a much less effective way of communicating knowledge than would be guessed from its prevalence as an educational practice. Other methods try to

remove the difficulties of human relationships from the act of communication by removing one party from the scene. For example, the apprentice can watch the master on film or videotape. Or the teacher can write his lectures in a book, which the student can then read. Another method is that of the dialogue. Unlike the previous methods, which imply superiority–inferiority relationships, this method implies equality between the members of the dialogue. All may at different moments be teachers or learners. The method assumes that all have some knowledge to give and that all need to learn.

Conceptions of the purpose of education. One of the commonest areas of disagreement among educational theorists concerns what education should be for. Two broad divisions of opinion can be identified. They are, first, that education should serve the needs of the individual and second, that it should serve the needs of society. In industrialized nations it is common for both of these goals to be held but for different classes of the population. For the elite, the needs of the individual tend to prevail, and thus upper-class schooling often tolerates diversity and encourages idiosyncrasy. For the masses, the needs of society tend to dominate, and schooling usually serves to prepare children to become obedient, well-drilled, uncomplaining workers in industry and agriculture. In addition to this dichotomy, there are many national and cultural variations. In France and Germany, for example, there is a strong tendency to see the primary purpose of education as some form of intellectual development, with the French *lycée* and the German *Gymnasium* serving as characteristic institutions. In England, education has usually been seen as primarily serving the purpose of character building. The English public school is sometimes taken as an epitome of this emphasis. In Roman Catholic and Communist countries, moral or religious training is usually the primary purpose of education. Hence, every subject is imbued with religious (or, if it is Marxism–Leninism, quasi-religious) content, and there is a constant attempt to draw moral lessons from educational material. In the United States, where large-scale immigration once brought the fear of social disintegration, preparation for citizenship and development of national consciousness have tended to be emphasized. Even today, nationalistic rituals such as displaying flags and reciting the "Pledge of Allegiance" in classrooms distinguish American schools from those in most other countries.

Needs of the individual and of society

Conceptions of uniform versus differential grouping. One of the controversies that most consistently divides educational theorists is whether students should be clustered together with similar students for instruction or whether instructional groups should be made deliberately diverse. The criteria that are employed for differentiating students are manifold. They include chronological age, mental age, IQ, skin colour, sex, social class, geographical location, parental income, performance ability on various tests, and so on. Some of these criteria, such as skin colour, are used "officially" in some countries, such as South Africa, and "unofficially" in others, such as the United States. Other criteria, such as parental income, are used overtly in some schools (expensive, private schools) and covertly in others (public schools located in expensive suburbs). Some criteria, such as mental age, IQ, and performance ability, often have the effect of separating rich from poor or elites from cultural minorities because of the cultural content of the tests used. The customary arguments used in favour of homogeneous grouping tend to emphasize speed and efficiency in instruction and learning. Arguments in favour of heterogeneous grouping tend to emphasize social outcomes, such as an enhanced understanding of others and tolerance of diversity.

Means of segregating students

Conceptions of control and discipline. When education is viewed as a transaction between superior (the teacher) and inferior (the student), questions of control invariably arise; *i.e.*, can students be relied upon to do what is best if they are not controlled by teachers? Answers to this question refer back to assumptions about the nature of man as learner. Those theorists who tend to see man as basically untrustworthy will tend to find justification for a regimen of strict control over students, in order to prevent their natural impulses from harming themselves and others. Theorists who view man as basically trustworthy will tend to justify more lenient, loose, and permissive systems of control. Most theorists would agree that an important goal of discipline would be for the student to develop self-discipline. But disagreement arises over whether one best learns self-discipline through practice in obeying others or through practice in commanding oneself.

Conceptions of the need for competition and collaboration. In most cultures of the world, children from an early age seek to compete against and to surpass their fellows. Whether this competitive spirit is innate or whether it is acquired through the culture is at present unverifiable. The question that divides education theorists, however, is whether these existing competitive strivings should be encouraged or reduced. Some argue that, since competition is endemic in the culture and since schooling should assimilate the child to the culture, the school should encourage individual competition and help children to develop the strengths necessary to compete successfully. On the other hand, others argue that competitiveness in the culture is pernicious and that schooling should not serve to adjust children to harmful aspects of the culture but should foster cultural progress. Usually, the school is in conflict over these two attitudes and gives mixed messages to students. As a result, the students often suffer the double strain of trying to compete successfully but to do so without appearing to compete.

Conflict of aims in school

Conceptions of the boundaries of formal education. Considerable disagreement exists over what should legitimately be studied in formal institutions of learning. Usually some topics or subjects are considered taboo. These taboos may come about through interpretations of the law, as in the case of the exclusion of religious teaching from public schools in the United States. Or they may come about as a result of pressure from interest groups, as in the attempts to exclude sex education, teaching about Communism, and sensitivity training from American schools. The major considerations are that young children are not well equipped to resist heavy bias in teaching and that they are usually compelled by law to attend school and thus constitute a captive audience. It is sometimes argued, therefore, that they should be protected against religious and political propaganda and against material or experiences that require greater maturity to be handled creatively. On the other side, the danger in such arguments is that they can be used to keep all controversy out of schools and to render them places characterized by dull uniformity of thought.

BIBLIOGRAPHY. For an excellent example of the analytical approach to philosophy of education, see ISRAEL SCHEFFLER (ed.), *Philosophy and Education*, 2nd ed. (1966). A representative British example is P.H. HIRST and R.S. PETERS, *The Logic of Education* (1970). Several scholars have attempted to group philosophies of education into schools of thought. The best collections are JOHN PAUL STRAIN (comp.), *Modern Philosophies of Education* (1971); and THEODORE B.H. BRAMELD, *Patterns of Educational Philosophy: Divergence and Convergence in Culturological Perspective* (1971). For examples of attempts to focus on a single problem or issue and to combine clarity and commitment, see PAUL NASH, *Authority and Freedom in Education: An Introduction to the Philosophy of Education* (1966); KENNETH D. BENNE, *Education in the Quest for Identity and Community* (1962); and ROBERT ULICH, *The Human Career: A Philosophy of Self-Transcendence* (1955). A good collection of Existentialist statements on education may be found in MAXINE GREENE (comp.), *Existential Encounters for Teachers* (1967). For selections from the classics on conceptions of the educated person, see PAUL NASH (ed.), *Models of Man: Explorations in the Western Educational Tradition* (1968); and ROBERT ULICH (ed.), *Three Thousand Years of Educational Wisdom*, 2nd ed. enl. (1954). Original essays on all of these major figures may be found in PAUL NASH, ANDREAS M. KAZAMIAS, and HENRY J. PERKINSON (eds.), *The Educated Man: Studies in the History of Educational Thought* (1965).

(Pa.N.)

Education, Social Aspects of

Education is a form of human interaction. It is a social act that is made possible through a network of human relationships. It is this network, and the individual relations and roles of people within it, that really determines the nature of education in a particular society.

The social aspect of education is illustrated by considering the dependence of individuals upon one another for learning. The nonhuman animal, bird, or insect can live out its life guided largely by biological inheritance, a genetic "program" for the behaviour of the organism. Inborn patterns teach it to raise its young, find food, and protect its territory. What it needs to know is transmitted largely through the genes.

Most of what human beings need to know, on the other hand, is not genetically programmed. Very early in life a child must start to learn a vast, intricate, and often confusing way of life. That way of life—the culture—cannot be biologically inherited; it must be learned anew by each person.

Schools, as formal institutions for learning, impose a particular set of conditions on education. As a result, school learning is quite different from that which takes place in the family, in peer groups, or in the community. "Education," in its widest sense, may be thought of as the socialization process through which a person learns his way of life. It is a continuing process from birth to death.

This article discusses four social dimensions of education: important social functions performed by the kind of education that takes place in schools; the social factors within the community that influence school education; the school's own bureaucratic organization, which influences school practices; and the social system of the school and its influence on teaching and learning.

THE SOCIAL FUNCTIONS OF EDUCATION

The transmission of culture. Schools are, first of all, agents of cultural transmission. Without some means of inheriting, as well as extending, the culture each new generation would have to start afresh. Culture depends upon continuity, and schools help to provide it by transmitting accumulated knowledge and skills and values and beliefs.

In primitive societies informal education was adequate for transmitting knowledge and skills. A child required no special preparation to learn the things he needed to know. He learned from his father and other adults the arts of hunting and food gathering. In somewhat later societies he learned to cultivate, plant, and harvest. Tasks were unspecialized and were easily transmitted to the child. The learning process was spontaneous, with parent and child largely unaware that it was happening.

The transmission of skills. Endowed with a capacity to learn, even very primitive man started to control his environment for his own purposes. He developed fire to keep himself warm, clothing to shield himself from the wind and cold, baskets and kettles to carry more than his hands could, and finally the wheel, to which he transferred burdens from his back. Some men were better at certain skills than others; they became artisans, making copper pots, weaving baskets, making wheels. Young men came to watch and to learn. They became apprentices, helping and learning at the same time. The first formal schools probably sprang up around such artisans, arising from clearly definable social and economic needs that were in turn a result of man's increasingly complex way of living.

These new agencies on the social scene separated children from their families for periods of time and put them in schools. Schools became institutions for learning; teaching became a profession, and learning a recognized pursuit of the young. Schools, in a cultural sense, are institutions created by society to transmit important knowledge and skills to the young.

The transmission of values and beliefs. Schools also transmit values and beliefs. Such qualities as honesty, thrift, and industriousness are, of course, encouraged by teachers in schools the world over. In some countries, Great Britain being an example, a certain amount of instruction in the established faith is required by law, whereas in many others, the United States and France being examples, instruction in the state schools is resolutely secular.

Education and indoctrination

But even in those countries in which education is secular, political and ideological beliefs may be inculcated to varying degrees. Such is the case in all Communist countries, in which Marxism–Leninism is taught in the manner of an official faith. In non-Communist states, governments may resort to transmitting propaganda through state controlled schools in an attempt to secure the future commitment of youth to the regime. An extreme example of this occurred in Ghana during Kwame Nkrumah's rule.

In Western countries political and ideological values are transmitted not by means of a program of overt indoctrination but through the fostering of patriotic ideals. The extent to which attempts are made to influence the political and social ideas of youth varies considerably from country to country and even from school to school. Most school curricula, however, provide for instruction in the history and practice of parliamentary democracy, and students are encouraged to participate in open decision making, as, for example, in debating societies and in the election of student officials.

In the United States the political content or slant of education has at certain periods and in certain local educational jurisdictions gone beyond this. Local education boards, made up of opinion leaders in the community, have in some cases allowed political attitudes to influence such matters as the selection and dismissal of teachers and the political content of school textbooks. The latter have sometimes reflected the hostile attitudes of local community and opinion leaders to foreign countries such as the Soviet Union and to domestic United States policies regarded as Socialistic.

Preparation for working life. Another social function that schools perform is to allocate youth to particular work roles. Modern society has an infinite variety of tasks to be performed, which, in terms of the education required to perform them, vary widely. Some may be performed with an elementary school education, and some require sophisticated higher education. The number of tasks that can still be performed by those with only an elementary education is diminishing rapidly in modern technological society; but occupations necessitating education at higher levels are increasing rapidly.

The process of channelling students toward a particular type of career begins in the elementary schools, in which grading is based typically on the level of capacity for achievement. The level of achievement in elementary school is usually an important factor in determining the type of education a child will receive at the secondary level. In the secondary school the basis for selection shifts to the *type* of achievement and the kind of job to which the education is likely to lead the student as an adult. The differentiation is therefore on the type rather than the level of capacity. The period since World War II has witnessed an increasing vocational emphasis in secondary education. This is a consequence of the desire of governments to secure an economic return from burgeoning education budgets, of the urgent need to prepare youth (a steadily expanding age group in the total work force) for gainful employment, and of the demands of industry for increasing numbers of skilled workers.

Vocational emphasis of secondary education

Such practical requirements have led to a typical tripartite division of secondary education: academic, technical, and general. In some secondary education systems —that of the United States for example—all three types of curriculum may be offered within the same school. In other systems—such as those of Great Britain, France, Germany, and the Soviet Union—separate schools are maintained for each type of ability and aptitude. In any case, the school thus puts its imprint on youth, certifying the tasks that they are presumably capable of performing in society.

Schools allocate youth in yet other ways. A school certificate is now almost a requirement for any kind of em-

ployment in all industrialized countries. A university degree carries its own status. And beyond the general values of certificates and degrees there is the hierarchy of values attached to the status of the institution from which they were taken. An "academic" secondary school, a "polytechnical" school, a "vocational school," and a "community college" all carry perceived images in society that are related to the likely performance of their graduates. At a still more elite level certain prestigious institutions, such as Oxford and Cambridge, the French *écoles nationales*, and the American "Ivy League" colleges imprint upon the individual an image of quality and high expectation.

The caretaking of youth. A further social function of education is the caretaking of youth. This function is seldom listed among the stated objectives of education. Some sense of the importance of this function in an industrialized society may be gained by looking at the relatively easy way in which youth enter an unindustrialized one. In a society that is composed of small farmers, herdsmen, artisans, shopkeepers, and servants, children tend to step quite directly into their father's or mother's occupations. Indeed, it is frequently necessary for economic reasons that they do so. In industrialized societies, on the other hand, few youths have an intimate knowledge of their father's work. Except in rural areas, few have the opportunity even to see their fathers at work, let alone gain an understanding of that work. The chances are that the father works away from home in an environment unknown to the child. The parents' work is also likely to be such that it is almost impossible for children to be apprenticed in it. Often the work is specialized, technical, and beyond the reach of children.

State nurseries

In some countries schools are expressly assigned a caretaking function, the purpose being to free parents to enter the labour force. The state nurseries of the Soviet Union and the kibbutzim of Israel are examples. In the kibbutzim, nursery-age children are cared for while parents are in the fields or factories. Thus parents are relieved of some of the day-to-day concerns involved in raising children, and the mother is free to take her place in the labour force.

Other schools serve a caretaking function simply as a social convenience, a place for socially minded mothers to "park" their children while they engage in other pursuits. This purpose is especially apparent in those affluent societies in which leisure time is abundant and in which sufficient money is available to permit mothers to engage in noneconomic pursuits.

Promotion of peer-group relations. The promotion of peer-group relations is a further important social function performed by education. All individuals have a basic need to relate to one another. The fear of being alone, of being cut off from meaningful human contact, of being a "nobody," runs deep within society. Social needs are among the most persistent and demanding.

Adolescent peer groups serve very real functions in society. They provide a way in which children can learn to become independent of family authority. In modern society maturity is equated with independence, with the ability to formulate one's own judgments, and with the capacity to take independent action and live by the consequences of that action. Peer groups provide children with experience of egalitarian relationships not possible in the family. Through peer groups the child is exposed to the values and experiences of dozens of other families, many of which may be greatly different from his own. Through these contacts the child's horizons are broadened, his perceptions widened.

In order for peer groups to serve these important functions the child must get outside of the family and interact freely with children of his own age. The school is ideal for this purpose. Its corridors and classrooms, clubs and activities provide a natural and convenient setting for the young to socialize. In addition, schools provide an environment in which boy–girl relationships and understanding may develop. This mixing of the sexes in youth performs a valuable function in Western society, in which the selection of mates is based largely on personal choice. Adolescence is a time for testing relationships and forming standards of selection.

ENVIRONMENTAL FACTORS AFFECTING SCHOOLS

There are various ways in which social and cultural factors in the community affect the school and its practices. It is possible, for purposes of discussion, to group schools in various ways according to the dominant cultural characteristics that they share with their communities. It should be realized, of course, that such categories are arbitrary and that all schools will not fit neatly into the types selected for convenience of study. One useful typology is that which places schools in broad demographic categories, which may then be compared in respect to such matters as political and social beliefs, economic wealth, social class, and ethnic composition. In this manner schools may be typed as village, town, suburban, and depressed-neighbourhood.

Demographic classification of schools

Village schools. Village schools play a particularly important role in the educational systems of underdeveloped countries. They are found in the bush and on the veld of Africa and throughout Asia and are spread across the length and breadth of South America. They symbolize the interest of people everywhere in providing formal education for their children.

Small village schools continue to exist in Europe and North America, but they belong to a vanishing era. This is not the case in other parts of the world, where the spread of education is symbolized by an open plot of ground in some remote area where parents have gathered after a hard day in the fields to build a school for their children, often on land they have purchased.

Village schools have an organic unity with the community. They impose no new ideals, open no strange vistas, and are rarely concerned about the future and preparing for it. The student's future is known, determined within the expectations of the village and the family. He is to take no sharp departures from the family's occupation, learn no disturbing new ideologies, perform no acts that lack time-worn precedents. Thus the student is held close to his past, fixing in him its value, motivating him to follow its precepts. His schooling sanctions what is already regarded to be of value. It looks to the past rather than the future, rating stability over change.

The town school. The town school has a cultural style with distinctive characteristics of its own. The student body of the town school is more mixed than that of the village school. In it are both "town" and "country" children who represent more than one social and economic class. Social class differences are usually not great, however, and children do not always divide themselves rigidly along social class lines.

The cultural and political style of the town is reflected in the elected school board. On it are the "solid" members of the community who represent the upper and middle classes. There is little disagreement among the social classes of the town regarding the functions of the school. Although there is not in the town the organic link with the school that one finds in the village, there tends to be general agreement about what the school ought to do.

The smallness of the town contributes to the school's cultural style. The small number of students provides an important deterrent to social class bias, for school activities must include almost everybody. If the school were larger it would likely be more selective. Like the community, the school requires the participation of all of its members. Children with a variety of backgrounds are in this way accepted in the school and begin to adjust early to the realities of human differences.

The town school thus socializes the young in terms of town values. It selects and sorts students to some degree but is relaxed about the process. It does not pressure students to go in directions in which they do not seem inclined to go. Because its broad scale of values permits a kind of natural selection to take place, the town school is not expected to force the child upward. In this regard the town school's cultural style is in sharp contrast to that of the suburban school, in which the dominant value is achievement for the future.

The suburban school. Social stratification takes place among and between suburbs. Some suburbs spring up in areas favourable to high status residential development —around golf courses, lakes, and along scenic rivers, for example. Other suburbs develop in less favourable locations. The city dweller who seeks a home in the suburbs soon discovers that the amount of money he can pay will largely determine the desirability of the neighbourhood in which he can afford to live and also the type of school his children will attend. Because of economic and social stratification in middle class neighbourhoods, the social composition of student bodies in the schools tends also to be stratified along social class lines.

Mobility is an important feature of upper middle class suburban culture. People who are upwardly mobile are likely to be transferred by their employers, to take new positions in other cities, or simply to move further out to other suburbs. Change rather than stability is the dominant feature. The future rather than the past or the present is the dominant direction.

Competitive ethos of the suburban school

The emphasis in suburban schools is typically on achievement in university entrance examinations. Counsellors help students secure college admission, and academic subjects are stressed.

The culture style of the suburban school is reflected in its teachers. A large number of the teachers are often strangers in the community, seeing parents in only a professional capacity. They rarely own property within the suburb and are not otherwise involved in its problems. The student is profoundly influenced by the professional role of the teacher in the suburban school. He sees the teacher only as "teacher," rather than as also a community leader, or for that matter, simply as a "human being." Thus the teacher's partial view of the child is matched by the child's partial view of the teacher. This relationship, however, is probably congenial to the main business of the suburban school—academic preparation. The teacher's narrowly defined professional role sharpens the focus on the academic business at hand. The broad values of the village and town teacher tend to be screened out by the drive for good grades and other symbols of academic achievement. The cultural style of the suburban school thus mirrors the values of the mobile elite of society. In this school the selection and sorting process tends to become feverish: the school transmits values connected with achievement, and socialization revolves around the acquisition of academic and social status.

The depressed-neighbourhood school. The cultural pattern of the depressed-neighbourhood school contrasts sharply with that of village, town, and upper middle class suburb. Located usually in a crowded quarter of the city, it has an asphalt pavement for a playground and high brick walls for landscape.

Most of the students are from lower working class or poor families. It is in depressed-neighbourhood schools that one is most likely to encounter extremely hostile relations between student and teacher and high rates of juvenile delinquency. Students generally leave school as soon as they are legally able to do so, and a high proportion of the teacher's time and energy is devoted to maintaining discipline and attempting to ensure at least functional literacy.

The uncongenial nature of the depressed-neighbourhood school serving the slums is not conducive to the fulfillment of any of the school's educational and socializing functions. Student attitudes, often reinforced by negative, alienated, and antisocial norms in the surrounding community, result frequently in hostile responses to attempts to introduce middle class values and aspirations.

Even in those cases in which students do aspire to break away from the depressed neighbourhood, its schools, and the concomitant lack of opportunities, upward mobility is particularly difficult. Students with a deficient educational background are likely to experience difficulty in finding employment when they enter the job market. A partial solution to this problem in several European countries and in the United States has been to provide more opportunity for further education beyond the minimum school-leaving age. Institutions providing this type of education, on a full- or part-time basis, have minimal entrance requirements and concentrate on remedying deficiencies in basic education and on providing elementary industrial and technical skills.

ORGANIZATIONAL AND SOCIAL FACTORS INFLUENCING SCHOOL PRACTICES

School organization. The bureaucratic view of school organization holds that teachers are employees in an organization in which superior–subordinate relationships should be clearly drawn and standards of business efficiency applied. The bureaucratic model places power at the top, supported underneath by line and staff relationships. Bureaucratic organizations such as this usually exhibit three characteristics. They are *standardized;* the stress is on uniformity among personnel and their interchangeability. Standardization calls for uniform rules and work procedures. Bureaucratic organizations are *centralized* in decision making and, thus, diminish the responsibility of employees. Centralization demands that the primary responsibility of the employees be to the organization. A bureaucratic organization is *specialized;* it stresses efficiency and technique.

Characteristics of bureaucratic organization

A professional organization, in contrast, stresses quite opposite conditions: the uniqueness of personnel and their skills; diffuse and flexible rules; the responsibility of professionals, rather than of administrators, for making policy decisions; primary loyalty to clients and colleagues rather than to the organization; and aiding students rather than applying efficiency and techniques. In a school that is organized along professional lines, teachers, as trained specialists, have more control over both the conditions of their work and over standards, and they are usually judged by peers rather than by outsiders.

Persons holding the bureaucratic and professional views take quite different positions on how conflicts should be resolved. Those taking the bureaucratic view argue that conflict should be resolved through meliorative processes of discussion and persuasion. Professional teacher groups, on the other hand, tend to hold the view that conflict is to be resolved by adversary processes, such as collective bargaining, in which parties bargain as equals. The melioristic approach is based on the assumption that educational conflict is amenable to resolution through good faith and serious communication. The adversary approach to conflict resolution assumes that conflict exists and will continue to exist in schools in which groups have different self-interests.

Formal and informal relationships. Each school has two types of social system, formal and informal. The formal social system is based on official positions, such as "teacher," "administrator," and "student," on the status relationships among them, and on the interactions of the people who occupy them. The status relationships of the formal social system may be observed in the administrative organization chart of the school. Such a chart diagrammatically describes the school's social hierarchy, with administrators assigned to boxes at the top, with teachers and students near the bottom, and supervisory personnel and "middle management" between. Lines connecting the boxes describe status relationships. A person's position in this hierarchy is usually determined by his authority and responsibility.

Whereas the formal social structure is diagrammatically neat, the informal social system is a scramble of relationships that tie people together within and across many boxes and lines. It represents the ability of people to humanize cold bureaucratic structures. The informal social system also represents a way for people to defend themselves against other persons of formal power and to develop power of their own. For the informal social system does have power. Teachers who do not like a program that they feel the administration has imposed on them can quietly agree among themselves to subvert it. Students who do not like a school policy can take measures to force a change. In these cases the struggle becomes one between the "establishment" and the informal social structure.

The informal social system of students has considerable

influence on academic achievement. Research suggests that the students' informal social system may apply sanctions to support athletics and social activities and discourage scholastic achievement. In the informal system the status of the brilliant student may thus not be valued, even though it is valued in the formal system.

Teacher–administration relationships. Perhaps the most crucial teacher–administrator relationship is that between the teacher and the school administrator. Of particular relevance is the question of whether the leadership efforts of the administrator have, or can have, a significant effect on instruction.

Importance of administrative leadership

Some research indicates that the school administrator's leadership does indeed have a positive impact on teacher morale and performance. One study of the leadership of elementary school principals in the United States found a positive relationship between leadership and teacher morale and performance. The research indicated, for instance, that the principal's leadership score was related to the classroom performance of the teachers. The higher the leadership score of the principal the more likely were the teachers to "do everything possible to motivate students," "plan classes so that different types of students can benefit from them," "provide opportunities to go beyond the minimum demands of assigned work," "try new teaching methods in their classroom," and "take a strong interest in the social and emotional problems of their students" and the less likely were teachers to "do textbook teaching only." These findings suggest that principals who provide a high degree of professional leadership are more likely to have teachers who are more productive and enjoy higher morale.

Parent–teacher relationships. Most parent–teacher interaction focusses on the child's school performance. In this respect the teacher is the judge of the performance, not the parent. In this most crucial matter, it is the teacher who is the dominant figure in the interaction. This focus of the relationship becomes even more crucial when it is remembered that most parents attach great importance to school achievement as a means to finding employment or gaining university admission. Middle class parents are convinced that their child's present school performance will determine his future life chances.

The achievement of the child compared to other children is of key importance in teacher–parent relations. To achieve scholastically the child must compete successfully against others, yet parents often disavow interest in how well their child is doing competitively. Teachers also attempt to play down the competitive emphasis and accentuate individual talents and abilities. The teacher attempts to reconcile conflicting values and expectations. The community values success as expressed in school performance. Yet children, being different, do not and cannot all achieve scholastic success in the same measure. The grading system, by definition, grades. The teacher's attempt to stress other values represents a sensible reconciliation in an area in which the school is a battleground of conflicting values.

Peer-group relations. One of the social functions of schools is to provide for peer-group activity under the supervision of adults. Peer groups serve important functions in society, and the school provides an ideal place for these functions to be performed.

Particularly noticeable are the changing positions of peer groups, with respect to adults, as children move from one age group to another. These changing positions may be described as a circle on which peers and parents begin at the same point. A small child tends to accept parental values uncritically. As he grows, however, the child gradually moves away from parental influence and toward peer influence, until as an adolescent he stands almost on the opposite side of the circle from that of his parents. Adolescence creates within youth an earnest effort to become independent, to reject the symbols of dependence, and to free themselves of authority and control. The peer group becomes supreme during this period, and it frequently wields more control than the family or any other group. By the time the youth reach the secondary school, adult values seem irrelevant or at least remote. What adults teach in the secondary school seems to have no vital relation to the peers' immediate personal and social concerns, which are intense and all absorbing.

Yet it is during this period that maximum pressure is exerted by the school and society for academic achievement. For those students who have university ambitions this is the critical period of decision. The record that determines future opportunities for advancement is being compiled. Peers are caught in the cross fire of pressures. From their fellows comes the urgent call of allegiance to the group and to peer solidarity and for a strong show of independence. From the adult world come pressures to achieve in matters not seemingly relevant at the present but supposedly necessary for the future. The student must reconcile these apparently irreconcilable appeals.

From this point on the circle a youth starts his journey back toward identification with adult values. If he enters a university, the need for the group begins to wane. He starts to feel independent, able now to focus more on his future. The need to prepare for a vocation and for marriage and a home comes more sharply into view. As he proceeds through the undergraduate course and possibly through graduate school, or out into marriage and a job, he thinks and behaves more and more like an adult.

The peer group and individual aspirations

In the peer group the students' informal social structure can be seen, a structure distinct from the formal program of classes, courses, and teaching. The formal and informal systems exist side by side and apparently converge at one critical point: the adolescent's level of educational aspiration, which is positively related to his peer-group membership. If the child's peer group values education, his desire for an education is reinforced. The opposite is also true.

Teacher–student relationships. In school, teachers perform a variety of social roles related to students, and these roles help to define the nature of teacher–student relationships. The teacher is, for example, a mediator of learning. Standing between students and the subject matter to be learned, the teacher seeks to bring the two together. This role is the one for which teachers are professionally trained and the one in which they tend to feel the most comfortable with students. As an examiner the teacher evaluates the work and progress of students in terms of stated or implied standards and objectives. The teacher is also a disciplinarian who is expected to promote conduct that is acceptable within the norms of the school. Sometimes the teacher is a confidante to students, sharing their trust and perhaps performing the functions of a parent counsellor. The teacher is expected to be a surrogate of middle class morality, representing the community in the classroom in upholding moral values and norms. Traditionally, the teacher was expected to be a parent substitute who acted with "reasonable care" in respect to the student's safety and general welfare. This role continues with respect to small children but is disappearing with respect to adolescents and young adults on university campuses. Finally, the teacher is an academic adviser who assists students in planning programs and in solving problems associated with learning.

The teacher's relationship with the student is sometimes described as a "client" relationship, like that of a doctor with his patient or a lawyer with his client. The analogy is only partly accurate. The relationship is indeed a professional one in that the roles of teacher and student are defined by law and convention. The teacher, however, is not granted the same freedom that a doctor enjoys with his patient or a lawyer with his client. Further, the teacher cannot select his client, as can the doctor or lawyer. Nor does the student client usually have the freedom to choose his teacher. For the teacher, the parent of the student is the indirect client, monitoring, as it were, the professional work of the teacher with the child.

BIBLIOGRAPHY. J.G. ANDERSON, *Bureaucracy in Education* (1968), an examination of alternative patterns of administration and their consequences for the individual and the organization; C.S. BREMBECK, *The Social Foundations of Education* (1971), a discussion of environmental influences in teaching and learning, both inside and outside the school; W.B. BROOKOVER and D. GOTTLIEB, *A Sociology of Education*, 2nd ed.

(1964), a standard textbook presenting a social interactionist theory of school learning; EMILE DURKHEIM, *Education and Sociology* (Eng. trans. 1956), an analysis of the major role that socio-cultural differences should play in the establishment of educational programs; R.J. HAVIGHURST and A.J. GOUVEIA, *Brazilian Secondary Education and Socio-Economic Development* (1969), a study of the Brazilian education system demonstrating the successful operation of models of manpower planning; M.N. RUTKEVICH (comp.), *The Career Plans of Youth* (Eng. trans. 1969), a translation of original Russian literature describing the official policy of the Soviet government toward manpower needs and education; L.M. SHARP, *Education and Employment: The Early Careers of College Graduates* (1970), a report on large-scale studies of recipients of bachelor's and master's degrees that deals empirically with the effects of formal education.

(Co.S.B.)

Education, Systems of

The term system of education refers to a society's total pattern of formal institutions, agencies, and organizations that transmit knowledge and the cultural heritage and that influence the social and intellectual growth of the individual. Although it would be possible to analyze a metropolitan, state, or provincial system of education, this article deals with national education systems.

A point to be emphasized when viewing systems comparatively is that no educational system is fixed or static but is always in a process of change, even if the change is not always clearly identifiable or perceptible. Since education is concerned with human beings—and they are changing—it is reasonable to infer that man's institutions and activities are also subject to modification. As one authority on comparative education, I.L. Kandel, put it in *The New Era in Education* (1955), no national educational system "has reached a stage of equilibrium; all are in a state of becoming." Also helpful in understanding such systems is the awareness that any system is affected by social, cultural, spiritual, economic, and political tendencies and forces, some of which originated centuries ago. Moreover, the educational ideas, practices, and structure derive frequently, in whole or in part, from countries other than the one being considered. It is clear, therefore, that the study of educational systems is a complex procedure combining description with analysis. Each individual system or issue must be understood fully and in terms of its various historical and contemporary contexts.

The article is divided into the following sections:

I. General characteristics of educational systems
  Varieties of administrative structure and control
  Factors in the formation of a system
    Ethnic and linguistic factors
    Political factors
    Religious and other factors
II. Types of educational systems and their characteristics
  Centralized systems
    France
    Italy
    Scandinavian countries
    Latin American countries
    Other centralized systems
  Decentralized systems
    United States
    Federal Republic of Germany
    India
    Canada
    Australia
  Joint national and local systems
    England
    Japan
  Systems controlled by political parties
    U.S.S.R.
    People's Republic of China
    Spain
  Sectarian systems
    Ireland
    Canada: Quebec and Newfoundland
    Pakistan
    United States: parochial education
      Roman Catholic parochial schools
      Protestant parochial schools
      Jewish parochial schools
III. Trends and prospects

I. General characteristics of educational systems

VARIETIES OF ADMINISTRATIVE STRUCTURE AND CONTROL

Types of systems. For purposes of classification it is useful and convenient to consider educational systems with reference to administrative structures and controls. In a large majority of countries, there is a centralized governmental agency, usually called the ministry of education or the ministry of education and culture, which organizes, administers, finances, and controls the formal and cultural aspects of education in all regions and localities. The laws, curricula, personnel, and materials and methods of instruction are in large measure determined in the central office, with variations in specific situations as occasion demands. However, even in a classic example of centralization—France—there is *in practice or action* no perfect centralization.

Centralization and decentralization

In some countries, such as Switzerland, education is carried on under the jurisdiction of the constituent states —the 22 cantons and 3 half-cantons. Though this situation is described as decentralized, there may be centralizing tendencies of a nonformal educational nature or common elements in the educational programs of the various units.

As a general rule, centralization often is taken to imply uniformity and conformity, if not indeed a monolithic or totalitarian situation, whereas decentralization is often equated with democracy and diversity. In point of actual fact, the line of demarcation is seldom that sharp. It is possible to have local linguistic differences in education, as in the Soviet Union, but centralized ideological control. On the other hand, in a country like Denmark, which has a centralized educational system, there exists considerable freedom of thought and opinion. One must be careful, therefore, to look under the surface of the classification in each particular instance.

Still another type of educational system is one in which administration and control are shared by national and local authorities. Perhaps the best example of this system is in England, where the educational laws originate in the Parliament but the actual administration is in the hands of the local governments. The Department of Education and Science (formerly the Ministry of Education), on the national level, issues reports, disburses funds, and offers guidelines, but it does not control the schools.

Nonpublic or private schools and other educational facilities may be controlled but not financed by the central authority in some cases. In others they may receive full or partial subsidies with varying degrees of autonomy. When the schools are related to a church or religious movement, they are frequently under the authority of a centralized agency, even though they may not be controlled by the government. In the worldwide perspective, there is a variety of patterns of educational systems.

FACTORS IN THE FORMATION OF A SYSTEM

Ethnic and linguistic factors. All systems of education are rooted in the past and are compounds of a variety of constituent factors influencing educational institutions and their functions. Racial and ethnic attitudes, for example, have played an important role in the formation of educational policy and in the administration of education in several countries. Historically, the attitude of the dominant white majority toward the blacks in the United States has resulted in discriminatory practices and segregated school facilities, a situation that has been undergoing change in recent decades. In England the recent influx of Indians, Pakistanis, and West Indians has provoked issues of curriculum, language, and equality of opportunity. The South African government's official policy of segregating various racial groups—white, black, coloured, and Asian—has brought about a quadripartite racial system of education. The caste system of India has continued to exert an influence on education, as newspaper reports periodically testify, in spite of the fact that discrimination against the untouchables is no longer legal. Also pertinent is the widespread Japanese effort to exclude the Eta (historically, an "untouchable" caste of uncertain origin) from any relationship, educational or

Racial considerations

otherwise, with the majority of the population, so that the Eta continue to have the lowest educational, social, and economic standard in the country.

Language is another significant factor affecting education in many countries. In some, such as South Africa, Belgium, and the Province of Quebec in Canada, the problem of bilingualism must be taken into account in administration, physical plant, curriculum planning, teacher preparation and selection, textbooks, and the like. Switzerland has been faced by the need of teaching in three, actually four, official languages. In the Soviet Union each constituent republic has a language of its own, and an Uzbek national living in Tadzhikistan has to learn the Tadzhik, Russian, and Uzbek languages if he is to participate and profit fully from his residence in the republic. The educational system of Israel, apart from maintaining schools in the Hebrew and Arabic languages, must also take into consideration the problem of teaching Hebrew to the new immigrants who arrive with multifarious linguistic backgrounds. Countries like Peru and Ecuador, with a large Indian population speaking many tongues, are endeavouring to solve the problem of teaching the national language. North and South Chinese can communicate with each other only via the written script or through such a neutral medium as pidgin English. After more than two decades of independence, India still is involved in the issue of national versus local languages, with English and Hindi dominating at the present time. Finally, recent social and economic trends resulting from migration have brought about changes in societies in which a single language has long been dominant. In certain parts of the United States, as in New York City, the influx of Puerto Ricans has necessitated the addition of Spanish-speaking teachers and other changes in the urban school system. The arrival of immigrants from Asia has brought about pressures in some of the cities of England with regard to teaching Tamil, Hindi, and other Asiatic languages in the schools.

Political ideology

Political factors. The political development of a country historically is a potent influence on the educational system. A totalitarian government will naturally organize a totalitarian educational system. In Nazi Germany, the objectives of Aryan or Nordic supremacy were applied in such a way as to deprive Jews and other non-Aryans of access to higher education, as well as to persecute them in countless ways. The authorities of Fascist Italy imposed their political credo—"Believe, obey, fight!" and "Mussolini is always right!"—upon all pupils and students and monopolized all media of public expression. The rules of the U.S.S.R. prohibit the teaching of religion, non-Marxian economics, and other doctrines deemed to be inconsistent with Communist ideology. Nations officially committed to the principles of democratic government and life tend to reveal a more pluralistic attitude toward other viewpoints and peoples. Though there may be discrepancies and contradictions in educational systems founded on the democratic philosophy, there is usually at least an *effort* to attain the professed aims of democratic education for the greater good of the nation and the individual.

Among the political ideologies and attitudes, account must also be taken of Socialist countries such as Sweden, which endeavour to organize their educational systems in such a way as to achieve maximum equalization of opportunity for all. Although this is also an objective of other democratic countries, in a dominantly socialist milieu it is likely to be accompanied, for example, by the elimination of all methods of selecting students that might lead to an elitist society in any form. Thus, in the years following World War II, when Sweden's Labour Party came to power, a concerted campaign was made to substitute the *comprehensive* secondary school (a school offering both academic and vocational curricula) for the dual types (separate schools for academic and vocational curricula), as well as to reduce the influence of the traditional private or independent school.

Religious and other factors. Until recently, many school systems derived their educational objectives and some of their content from religious ideologies. In many diverse countries, including West Germany, England, Denmark, and Norway, the study of religion is a regular part of the curriculum. As a general rule, religious instruction has tended to persist in Roman Catholic countries, such as Spain and Italy. Secularization of schools was undertaken in France in the late 19th century, with the result that religion is no longer taught in the public schools there. In the Soviet Union, by constitutional prohibition, religion may not be taught in schools, although, interestingly, another Communist country, Poland, permits the teaching of religion. In some countries, such as the United States, there exists a dual system of education, one with a secular orientation that is publicly supported and one that is composed of a private, often church-supported series of networks, both of them contributing to the culture and economy of the country. The United States, in line with the doctrine of separation of church and state implied in the First Amendment to the Constitution, does not support religion in education, but some states do allow some benefits to children attending church-administered schools, and some provide such schools with funds obtained from taxes.

A thorough analysis of an educational system would comprise, in addition to the miscellaneous factors already enumerated, the educational history of the country, together with foreign influences; the aims and ideals of education; the organization of the preschool, elementary, secondary, vocational, higher, adult, and professional levels; the provisions for the handicapped and other forms of special education; the varieties of administration, supervision, finance, and control; the types of curricula, methodology, and materials of instruction; the means of psychological and achievement testing; the types of youth groups and agencies; the measure of scholarly, educational, and professional organizations and publications; the influences of the mass media; and those problems and issues peculiar to the country, as well as those common to other countries. (W.W.B.)

II. Types of educational systems and their characteristics

CENTRALIZED SYSTEMS

France. The modern tradition of education in France can be traced to the Revolution of 1789, when the firm hold of the Roman Catholic Church was broken for the first time. The constitution of 1791 provided for the creation of a system of free, public, universal education. Centralization came with the Napoleonic period, especially with the establishment in 1806 of the "University of France," a rather abstract term that covered the entire educational system at elementary, secondary, and higher levels. During most of the 19th century, private and parochial schools were allowed, but, with the advent of the Third Republic in 1871, secularist trends again dominated. The laws of 1881–86 made French education free, compulsory, and secular, with teachers in state schools required to be laymen. The universities, which had been discontinued in 1808 in favour of a system of faculties of science and letters (see HIGHER EDUCATION), were reconstituted by the law of 1896. To maintain centralization, the national government assumed the sole right to grant certificates and degrees, even for students in private schools. There is now one significant exception: although the government confers the major doctorate, the *doctorat d'état*, each university is authorized to grant the *doctorat d'université*, a lesser degree, often to foreign students.

The French system as the classic example of centralization

Because the French system is the classic example of centralized administration, it will be worthwhile to detail its structure. The country is divided into political-educational units decreasing in size from *académies* to *départements* to *communes*. Each of 23 *académies* is headed by a university. The chief administrator of both the university and the *académie* is called a rector, who is appointed from the ranks of university professors by the president of France on recommendation of the national minister of education. He is responsible primarily for

secondary and higher education within his *académie;* he nominates candidates for administrative positions in his area, appoints examination committees, supervises examination content and procedures, and presides over a sort of personal cabinet called an academic council. Unlike political divisions in such countries as Australia, Canada, the United States, and West Germany, the *académies* have little power or authority of their own; rather, they are administrative arms of the Ministry of National Education; they represent a devolution rather than a delegation of power from the central government. (Since the law of 1968 the universities themselves, however, have been controlled by elected faculty-student councils.) The *départements* (from two to eight in each *académie*) are likewise administrative units governed by prefects who are appointed in Paris. The prefect heads the elementary school system, acting in a manner similar to the rector. The prefect, too, has an advisory council, appointive in nature, and by no means equivalent to local boards of education in such countries as the United States and Canada. *Départements* are empowered to levy subsidiary school taxes as approved by the national government; in the early 1960s about half the cost of *département* schooling was met by local taxation. *Départements* are also permitted to provide further school services, on approval of higher authorities, including additional emoluments for teachers. There are about 38,000 *communes,* varying in size from the smallest community to the city of Paris itself. The chief officer of the *commune* is the mayor, who is also the president of the *commune* school board. Again, these school boards are largely advisory in nature; they also see to it that national and *département* educational procedures and policies are carried out in their particular area; they have only minor administrative responsibilities and can initiate very few regulations or practices. Actually, the *département* prefect assumes much of the control over education in the *commune,* as such control is derived from higher authorities.

The French educational system is characterized by a pyramid of bureaus that reach their apex in the Ministry of National Education in Paris. As in all national systems, even including Great Britain, the minister of education is a political appointee of the prime minister and, as such, usually falls whenever his government falls. Practical control is exercised by a powerful and permanent civil service, which remains relatively unaffected by political changes. A corps of inspectors general, appointed by the Ministry of National Education, reports on progress and needs in all areas. Increasingly, control is being transferred to local governments, especially in matters involving purely internal conduct, but final responsibility still lies heavily with the central ministry.

Types of French schools

The French system of education is free, secular, and compulsory for children from age 6 to 16. Although most schools are public, there are a substantial number of private schools, most of them under Roman Catholic auspices. The state offers subsidies to private schools and exercises some control over their instruction and school sanitation. The school system as a whole consists of the following levels. For preschool children aged two to six there are the *école maternelle,* the *jardin d'enfants,* and the *classes enfantines* of the elementary schools. For children aged 6 to 11 there is the *école primaire élémentaire.* On the intermediate level, adolescents aged 11 to 15 go to either the *collège d'enseignement secondaire* or the university-bound *lycée.* On the higher secondary level adolescents aged 15 to 18 continue on one of three courses: the college of general and technical secondary education, the *lycée,* or the elementary normal (teacher-training) school. Finally, there are the universities (which admit those passing the *baccalauréat* examination) and the *grandes écoles*—such higher professional institutions as the École Polytechnique, the Conservatoire National Supérieur de Musique, and the École Normale Supérieure. The system also includes a variety of agricultural, vocational, and technical schools, as well as special schools for the handicapped.

French education, in general, aims at the cultivation of reason and the intellect—the objective being, in terms of content, to impart a general education (*culture générale*) stressing literary, historical, scientific, and technological phases more equally than before. (W.W.B./G.F.K.)

Italy. Italian education today is founded upon the constitution of January 1, 1948, which provides for a free, compulsory, eight-year elementary education and further secondary and higher education for students having means or meriting scholarships or other allowances. Beginning in the mid-19th century increasing laicization eventually diminished the control of the Roman Catholic Church and ended the teaching of theology in the state universities; nevertheless religious instruction has remained compulsory in the elementary schools (though pupils may be individually exempted on parental request) and voluntary in the secondary schools.

The minister of public instruction is the central director and supervisor for all forms of education, from preschool through university, and for artistic and cultural activities. The curriculum of all schools, including private schools desiring official recognition, must conform to that prescribed by the ministry. This central authority is represented in each provincial capital by a *provveditore agli studi,* or superintendent of education, who oversees all elementary and secondary schools within his area, both public and private, and who lends a certain note of "decentralization" to the system by being permitted to exercise some local discretion in implementing programs. Each province is itself divided into school districts (each headed by an inspector of schools), which are further subdivided into teaching groups or "circles" (*circoli didattici*), headed by a director and consisting of 60 to 100 teachers. Universities and higher institutes are supervised directly by the central ministry's directorate-general for higher education. An important consultative body in the ministry is the Higher Council of Public Education (Consiglio Superiore della Pubblica Istruzione), composed of 60 members chosen by teachers and professors, chaired by the minister, and authorized to offer advice, thus providing the central authority with an awareness of public and professional opinion.

Decentralizing elements in Italy's centralized system

In structure, Italian education begins with the preschool (*scuola materna*) for children aged three to six and the five-year elementary school (*scuola elementare*). There are two types of intermediate school—the *scuola media inferiore* leading to secondary education and the *scuola di avviamento professionale* which might be called "prevocational." There follows either the two- to five-year vocational school (*istituto professionale*) or various forms of secondary education for students aged 14 to 19, including the classical school or *ginnasio-liceo,* the *liceo-scientifico* or four-year institute for elementary school teachers and art students, and various five-year technical institutes. In higher education the state and private universities and higher institutes offer almost identical curricula leading to the *laurea di dottore* (doctor's degree). Except for the course in medicine and surgery, however, the Italian *laurea* is not regarded internationally as the equivalent of the Ph.D. as granted in the United States and some other countries. Unlike England and the United States, Italy has offered few opportunities for postgraduate study. The new law of 1969 on university reform may, however, change the status of higher education. The authorities hope that the new research doctorate (*dottorato di ricerca*) will win international recognition for the products of the higher educational system of Italy.

The aim of the Italian school system is to give a basic education to all, including religious and linguistic minorities, and to aid those lacking means to pursue secondary and higher studies. Apart from preparing skilled workers and professionals, the school attempts to educate citizens for living in a democratic society.

Scandinavian countries. Historically, Scandinavian education has been strongly influenced by Lutheran doctrines. Supervisory and administrative activities by the clergy were not discontinued until 1933 in Denmark and until 1958 in Sweden. The survival of some church influence in Norwegian education is evident from the ex-

istence of the Royal Department of Church and Education. Even though the administrative role of religion has been fading in recent decades, Lutheran religious instruction is still required in the public schools, with provisions made for abstention on the basis of conscience.

Lutheran influences in Scandinavian systems

Since the mid-19th century there have taken place several significant developments in education. The Danish movement in adult education, centring on the "folk high school" (*folkehøjskoler*) was founded in 1844 and became the model for similar efforts in other European countries, particularly Great Britain, and in North America. The Swedish *grundskola*, a nine-year comprehensive elementary school, introduced by stages since 1949, has encouraged other countries to try similar educational reforms combining classical, modern, and technical studies. Also gaining international notice has been the Swedish program of sex education in the schools.

In every Scandinavian country the educational ministry is the central and supreme authority and issues regulations and plans binding on local boards and councils. The local authorities are, however, granted extensive autonomy in educational matters. In Denmark the Ministry of Education (Undervisningsministeriet) shares decision making and financing with the municipal councils (such as the Municipal Council of Copenhagen) and, in rural parishes, with county school directorates. In all areas there are teachers councils entitled to be consulted on all important educational questions. The Ministry of Education also exercises supervisory relations with the various state institutions of higher learning, such as the University of Copenhagen. In Norway, the Department of Church and Education (Kirke- og Undervisningsdepartementet) exercises overall control of schools and direct supervision of teacher training and higher education, but local authorities have considerable latitude of action with regard to equipment, appointments, and finance. In Sweden the school system is under the authority of the ministerial Department of Education (Undervisningsdepartementet), but effective operational direction is exercised by the Royal Board of Education (Kungliga Skolöverstyrelsen) for elementary and nonvocational secondary schools and by the Royal Board of Vocational Education (Kungliga Överstyrelsen för Yrkesutbildning) for vocational and technical schools. Local school authorities share in the administration of elementary and secondary education; county school boards are concerned with supervision of elementary schools, teacher appointments, and the disbursement of state funds to the schools. In Scandinavia generally, the trend in recent years has been toward an increase in shared educational responsibility among the various levels of government, even if ultimate control and direction (especially with regard to curriculum) remains in the hands of the central ministry.

Like other school systems in Europe, the Scandinavian school systems have traditionally been dual-tracked: at the secondary or intermediate level the majority of students were channelled into vocational schools and students of merit or means were channelled into academic and pre-university schools. The Danish "academic" channel consists of a three- or four-year preschool (*børnehave*), a seven-year elementary school (*hovedskole*), a three-year course of lower secondary education (*realafdeling*) leading to the academic three-year secondary school (*gymnasieskole*), and then an institution of higher education. The Norwegian system consists of a four-year preschool, a seven-year compulsory elementary school (*barneskole*), a three-year general lower secondary school (*ungdomsskole*), a three-year lower academic secondary school (*realskole*), and a three-year higher academic secondary school (*gymnas*). Sweden has tended to de-emphasize the old dual plan and introduce comprehensive schools; it has a four-year preschool (*förskola*), a comprehensive nine-year elementary school (*grundskola*), and a three-year academic secondary school (*gymnasium*). In each country the transition from secondary to higher education is by way of an examination (*studentereksamen* in Denmark, *examen artium* in Norway, *studentexamen* in Sweden). Each country provides for vocational and technical training on various levels.

The aim of education in Denmark, as expressed in the Elementary Education Act of 1958 and other documents, is not only the dissemination of knowledge but also the development of character in line with Christian principles and of friendly attitudes toward other nations. According to the Norwegian law of 1959, education is a cooperative process between school and home for the attainment of civic virtue, Christian character, basic knowledge, and respect for other peoples, so that the individual can function as a useful human being. These aims are also shared by the Swedish educational system. (W.W.B.)

Latin American countries. Education in most Latin American nations is based on Spanish—or, in Brazil, on Portuguese—traditions as modified by later French, German, and American influences. As a rule, education is under the control of a national ministry of education of varying designation—such as the Ministry of Education and Justice (Argentina), the Ministry of Education and Culture (Brazil), the Ministry of Culture (El Salvador), or the Ministry of Education and Religion (Paraguay). The degree of centralization may be greatest in Cuba, which follows Marxist political principles, but educational authority is generally everywhere vested in a national ministry.

Uruguay's unique system of centralization

Uruguay, however, provides an interesting example of a different type of centralized control. Although nominally the highest administrative agency is the Ministry of Public Education and Social Welfare, the ministry actually concerns itself primarily with welfare services. Educational administration falls within the province of four national autonomous educational councils: the National Council for Primary and Teacher Education (Consejo Nacional de Enseñanza Primaria y Normal); the National Council for Secondary Education (Consejo Nacional de Enseñanza Secundaria); the Central University Council (Consejo Central Universitario); and the National Council for Industrial Education (Consejo de la Universidad del Trabajo), which occupies itself with agricultural, vocational, and industrial schools of all types, but which is divorced from the political and ideological overtones that characterize labour educational administration in such a country as Mexico. The coexistence of these four councils illustrates a form of decentralization; since the councils operate on a national basis, however, and since their functions interlock, the resulting organization might be conceived as a type of pluralistic centralized control. As for curricula, the policy regarding the teaching of religion varies—for example, in Brazil it is optional, whereas in Mexico it is forbidden in public schools. Among educational problems common to the Latin American nations are the instruction of the indigenous Indian population, the reduction of illiteracy, the changes resulting from unstable political conditions in some countries, the need for properly prepared teachers, the inequality between rural and urban schools, the education of vast numbers of poor people, and the perennial unrest among students in secondary schools and higher educational institutions.

Other centralized systems. National systems of education prevail in Belgium, the Netherlands, Greece, and other countries where centralized controls are geographically practical and strategic—such as the Middle East (Turkey, Iraq, Syria, Egypt, etc.)—and other political systems not federal in nature. In most traditionally centralized systems, however, local control over educational administration has increased since World War II.

(W.W.B./G.F.K.)

DECENTRALIZED SYSTEMS

Included in decentralized systems are state, provincial, county, and community types of school administration. Local emphases in administration are to be found in countries where public schooling originated in grass roots or separatist movements, or where political unity was achieved out of a confederation of sovereign states; though in some instances of federalization (Italy), a

central ministry has been granted administrative as well as advisory power. As a rule, however, cultural and political units have jealously guarded their control over education. Characteristics of local control may be generalized thus: (1) weak central ministries or bureaus of education, which are largely advisory in character; (2) financial support for education locally derived, aid from central authorities being mostly in the form of supplementary or emergency funds accepted with a minimum amount of control; and (3) administrative power lodged in local boards of education, sometimes appointed, sometimes elected. In such countries as Canada, West Germany, and Australia, where the separate states are sovereign, state ministers of education are appointed and state systems of inspection prevail. In the United States the state executive officer of education becomes either a commissioner of education (if appointed) or a state superintendent of schools, each with less power than his counterpart abroad.

United States. The United States is perhaps the outstanding example of local control on state, county, and community levels. The decentralization of education characteristic of U.S. schooling is a source of bewilderment to observers from other countries who seek to understand not only the many different types of state-controlled education but yet another system in the District of Columbia, administered by the federal government.

The whole is complicated by a myriad of local county, township, or community units of education, by types of schooling directly supported and controlled by the federal government, and by the great variation among the states as regards internal administration and relations with local organizations. Some states, like Delaware, tend toward centralization. Massachusetts is an example of decentralization. Even in the support of education, the amount of funds provided can vary from the practice in a state such as Nebraska, where in the 1960s, for example, over 75 percent of expenditures were underwritten by local authorities, to those like New Mexico, where state rather than local revenues paid for most public schooling. In addition to the various state public-school systems, there exist, outside these systems, parochial schools that are administered by church groups (especially Roman Catholics) and a large variety of privately supported schools.

Federal influence in the U.S.

Although no mention is made of education in the U.S. Constitution, the federal government has been mindful of the need to support formal schooling, exemplified in the Land Ordinance of 1785 and the Northwest Ordinance of 1787, by which part of the national domain was set aside for school purposes, and in the numerous federal statutes that have supported types of education (technical, military, agricultural, etc.) necessary to national welfare. The earliest federal statutes, enacted in accordance with traditional educational practices, gave control of lands and resources to local townships; but by 1804, with the Ohio Territory Act, the states accepted custodianship and assumed control of federal school support.

Education in the different colonies was variously conceived as a family, church, state, or charity matter. In Massachusetts, compulsory school attendance laws were enacted, but responsibility for education still lay with parents. In Virginia, voluntary charity groups were active. Later, however, the federal government, acting by authority of the "general welfare clause" of the Constitution (section 8, article i), required of all states joining the union that satisfactory provision for public education be stipulated in the state's constitution. By a series of Supreme Court decisions the federal government settled local educational disputes and in part guided the course of education in its national implications. The federal government has also supervised the educational welfare of territories, trusts, and dependencies and assumed the responsibility for educational agencies supplying the federal government with the kind of personnel and services it requires. Included in this category are the various academies and institutes that train officers and specialists for such areas as national defense, diplomacy, internal security, and technology. The federal government has also supported educational relief programs, such as the National Youth Administration and the Civilian Conservation Corps in the 1930s, the work-study and other educational programs set up under the Economic Opportunity ("War on Poverty") Act of 1964 and the school-aid bill of 1965.

The U.S. Office of Education was established by Congress in 1867. In 1939 it was transferred from the Department of the Interior to the Federal Security Agency. It became a branch of the Department of Health, Education, and Welfare when that department was established in 1953. It is not an administrative agency in essence, although it has assumed control over some projects supported by the federal government, such as certain programs of vocational education, exchange of students and teachers with other countries, certain aspects of land-grant-college administration, and the general promotion of educational thought and practice throughout the country. Specifically, its functions involve the collection, interpretation, publication, and dissemination of factual information on education, not only in the U.S. but abroad. Its specialists may be drawn upon for counsel by any educational group in the country. (Similar agencies exist in India, Canada, Australia, West Germany, and other nations where federated governments and local controls determine educational administration.)

State influence and responsibility in the U.S.

Responsibility for education in the United States rests with the individual states. The states delegate this responsibility to the local communities, not simply as a devolution of power, as in France and other centralized systems, but as a matter of self-determination under fairly liberal state school codes. It is commonly claimed that the three levels of educational organization and administration in the U.S.—federal, state, and local—are actually interdependent rather than hierarchic. Hence, a system of checks and balances is said to prevail in educational as well as other governmental matters. In practice, however, administration rests heavily on the shoulders of local communities and operates in accordance with local decisions, as long as such decisions accord with federal and state statutes.

On the state level, educational administration generally is headed by a state board of education, sometimes appointed by the governor or the state legislature, sometimes elected. As a rule, state boards of education are less powerful than state ministries of education in other countries, largely because local communities in the United States assume much more administrative responsibility. By way of exception, however, the board of education of the state of New York exercises powerful control over educational standards through a system of statewide examinations. The extensive ramifications of its operations are seen in the fact that its staff consists of over 500 full-time professionals. State boards range in size from 3 to 19. They are largely composed of laymen, but in some instances professional educators may become members.

Again, the duties of these state boards vary from state to state, but in general they all concern themselves with the distribution of state or federal funds (allotted to the states), the enforcement of educational statutes, the determination of basic courses of study, the adoption or recommendation of textbooks, the certification of teachers, the approval of building and maintenance standards, the provision of library services, and the operation of teachers' colleges. State boards likewise interest themselves in the improvement of education as a function of government, the equalization of educational opportunity, the provision of educational leadership, and the improvement of local operations. The inspection of local schools and school systems by inspectors responsible to state boards or other authorities is found in many states but is not as highly developed as in most other countries. School and classroom supervision is provided by local systems through specialists who do not function as inspectors but rather as "helping teachers." In one state they are officially so designated. The genius, but also the weakness, of the U.S. educational system lies in its more

Authority of local school districts in the U.S.

than 30,000 school districts, each exercising an amount of autonomy unequalled in any other country. Administrative authority is assumed by a local board of education, which in about 90 percent of cases is elected by the people. Boards of education are usually independent of other agencies of local government; in fact, school funds voted by boards of education are for the most part not subject to rejection by other bodies. In cases of dispute over budgets between boards of education and boards of finance, the courts have ruled almost exclusively in favour of the former. In most communities, school taxes have first priority and usually are independent of other sources of revenue.

The duties of a local board of education are limited to community needs. It selects the superintendent of schools, who in turn recommends personnel for appointment by the board. In nearly all cases, boards of education have the power directly to appoint nonteaching employees and to determine educational policy for the district, acting usually on professional advice, secured by way of the superintendent of schools and in accordance with the wishes of the community at large. The board assists the superintendent in the preparation of the budget, approval of which obligates the local board to see that tax funds are provided.

Because of the administrative decentralization, there is no uniform school organization or curriculum. Preschool education consists of nursery schools and kindergartens, although the latter in recent years have become more integrated with—usually a part of—the elementary schools. The elementary-secondary sequence overall is 12 years in length (not counting a one- or two-year kindergarten), but the subdivisions of these years are various: eight-four or six-six (elementary school and high school), six-three-three (elementary school, junior high school, and senior high school), and four-four-four (primary school, middle school, and high school), and some modifications of these basic patterns. In many states and local communities, there are also 2-year "junior" or "community" colleges, which provide terminal education (usually vocational or semiprofessional) for many students and the first two years of undergraduate college study for others, who customarily then transfer to full colleges or universities. Admission to a state college or university requires only a high-school diploma in many state institutions, but in other state colleges and universities, the requirements are either a superior high-school record or a satisfactory score on a national test administered by a group such as the College Entrance Examination Board, a nongovernmental organization, or both. Thus, the process of admission is less selective and more flexible than that of most other countries (see below *Sectarian systems*). (G.F.K.)

Federal Republic of Germany. During the 19th century and until the advent of Adolf Hitler and his Nazi dictatorship in 1933, Germany enjoyed an excellent reputation in education all over the world. After the debacle of World War II, Germany was split into four zones of occupation by the victorious powers, so that four systems of education resulted. With the establishment of the Federal Republic (West Germany) and the Democratic Republic (East Germany), in 1949, the organization of education in the West began to come under German direction. East Germany, too, controlled its own educational system, but it drew more inspiration from Soviet education than West Germany did from the English, French, and American systems.

According to Article Seven of the Basic Law (constitution) of 1949 of the German Federal Republic, education is a function of the 11 states (*Länder*), including West Berlin. In each state, the ministry of education exercises fairly complete control over all schools and close supervision of all major activities, although the local authorities (*Schulträger*) have responsibility for school buildings, equipment, and student health and welfare. Financing of schools is generally a function shared by the provincial and the local authorities. External university administration, including certain financial, academic, examination and other aspects, is exercised by the division of higher education of the state ministry of education, though the universities enjoy a large amount of autonomy.

Elements of centralization in West Germany's federal system

Even if West Germany does not have a centralized administration, it has been making some efforts, since 1948, to secure some sort of uniformity through the Permanent Conference of Ministers of Education (Kultusministerkonferenz). The periodic discussions of problems common to the various states has led to the adoption of agreements as to educational policy and content. Such decisions lack a legal status, however, unless they are adopted by the legislatures of the separate states. By way of example, the Hamburg Agreement of 1964 concerned such questions as examinations and grades, the period of compulsory attendance, the beginning of the school year, holidays, and terminology relating to schools.

The school systems of West Germany start with a three-year nursery school and kindergarten. Full-time compulsory attendance extends from age 6 to age 14 or 15, depending on the state; some education is compulsory until 18. The four-year elementary school (*Grundschule*) is the basis of the school system (in Hamburg, Bremen, and Berlin, elementary education lasts six years). Completion of the *Grundschule* leads to the five-year upper elementary school (*Hauptschule*) or to the six-year secondary school (*Realschule*) or to the nine-year academic secondary school (*Gymnasium*) preparatory to the university. In recent years it has become increasingly possible for pupils to demonstrate their ability at various stages and thus be enabled to transfer from one type of school to another. The passing of the *Abitur*, or school-leaving, examination gives the student access to higher education. Recently, through the "second chance" (*Zweiter Bildungsweg*), young adults may qualify for the university by attending a special preparatory school (*Kolleg*) or by passing special examinations. The university or professional college (*Hochschule*) confers a diploma or degree of *Doktor*, which requires a thesis. Formerly, for teaching in a higher educational institution, the presentation and defense of an advanced dissertation was necessary; the successful candidate received the degree of *Habilitation*. This degree, however, is no longer required for assuming a professorial post in a university.

The West German educational authorities are addressing themselves to various problems on the provincial and local levels. One problem concerns the facilitation of the transition between the terminal *Realschule* and the academic *Gymnasium*. Another issue is that of the provision of technical secondary schools parallel to the *Gymnasium*, so as to raise the level of vocational training. Also pertinent is the matter of educational television courses as supplements to instruction and as preparation for higher studies. A perennial question relates to how to teach the German child and adolescent the history of his country during the 20th century, with particular reference to World War I, the Weimar era, the Nazi era, World War II and its aftermath, the relations between West and East Germany, the importance of democratic government and society for the future of Germany, and other significant developments. Much has been done through conferences, teachers' publications, and new textbooks for pupils. (W.W.B.)

Ancient Indian and modern English influences in Indian education

India. In ancient and later times, India served as the educational inspiration for countries in Central and Southeast Asia. With respect to Europe, there was Indian influence, mainly through the Arabs, in the areas of grammar, literature, and mathematics. From the late 18th century to the mid-20th, Indian education came under the influence of the English educational tradition. The ideas of Mohandas K. Gandhi, Rabindranath Tagore, Jawaharlal Nehru, and other national intellectual leaders helped form the foundation of the school system of India, independent since 1947.

India may be taken as an example of local control strongly exercised by the various regional states under state ministries of education. Unlike the United States Constitution, the Indian constitution of 1947 asserts that

the federal government will endeavour to provide free, compulsory education for all children until they reach the age of 14. Partly because the central government did not have sufficient funds to implement this ideal and partly because of the natural diversity of India's population, considerable responsibility for education has been delegated to the states. The National Ministry of Education has advisory functions approximating those of the U.S. Office of Education, but whereas the U.S. commissioner of education is appointed by the president, the Indian minister of education is elected by Parliament and is directly responsible to that body.

Each state ministry of education is headed by an elected minister and has five divisions: (1) administration and cultural relations, (2) bureau of development, (3) general education, (4) technical education, and (5) scholarships. Delegation of authority by the states to the local communities is restricted to elementary education, as is the case in most countries, while secondary education is subject to state control. Indian state governments are responsible for educational policy, the local units being conceived largely as administrative and taxation units. Control over education is exercised through an inspection system.

Advising each state minister of education is the director of public education, who is a permanent executive head of the state ministry. He controls the inspectorial staff, the teachers, and the government-recognized private schools and is responsible for executing the policies of the Ministry of Education, principal among which is the inculcation of basic education through productive activity and local crafts to all children between the ages of 6 and 14. In the curricular system, the pupil may begin an eight-year elementary school at the age of six and possibly move on to a three-year secondary school and a three-year college, which comprise the higher scholastic organization. Other variations of the elementary school offer four- or six-year programs followed by a three-year junior secondary or middle school and a three-year secondary school.

Problems facing Indian education

There are great inequalities of educational opportunity in India, where population distribution, linguistic diversity, and the heritage of the caste system aggravate an already difficult educational situation. The Indian national government has confessed its inability to assist adequately in fostering school development, not for want of desire but from sheer financial inability. It has confined its efforts to supporting higher and technical education, research in the interest of national welfare, experiments in educational methods, the production of suitable literature, the training of specialized personnel, and the translation of important works into Indian languages. The administrative problem in India is therefore not one of where control ideally may be located but of securing the kind of educational organization and administration that will best serve a country that is as yet economically undeveloped, and where the many religions and sects insist upon tying public education to their particular beliefs.

Widespread, endemic poverty often prevents individuals from utilizing their full potential in acquiring an education or, if already educated, from using what they had learned. The problem of basic education, or the teaching of literacy, crafts, and skills, remains a pressing one, inasmuch as India still has some 300,000,000 illiterates.

Apart from the problems of literacy and the untouchables, the most publicized issue has been that of language. With some 12 major regional languages and about 200 other languages, it is clear that a significant barrier exists to effective communication. It is also obvious that a fully educated individual must know Hindi, the "official language of the Union," English as the international and common cultural language, and a regional language. In addition, some may have to acquire written knowledge of another language—their native tongue. The central government is endeavouring to have Hindi accepted as the national language. With political and other factors involved in the campaign, the future of linguistic education in India seems clouded.

Canada. With a population of native Indians and Eskimos and with immigrants from France, Britain, and the American colonies, Canada already had a complexion of cultures by the time the British North America Act of 1867 created the confederation of provinces. During the following century, Canada was to experience an influx of persons from Europe, Latin America, and Asia and thus became, like the United States, a product of many cultures. Under the terms of section 93 of the 1867 act, each of the ten Canadian provinces was given authority to direct educational activities within its own borders. However, the provincial legislatures were denied the right of enacting any law that "shall prejudicially affect any Right or Privilege with respect to Denominational Schools which any Class of Persons have by Law in the Province at the Union." In this way, the religious educational traditions existing at the time, particularly those of the Roman Catholics and the Protestants, were safeguarded (see below *Sectarian education* for further description of Canada's Catholic and Protestant school systems). Other religious denominations were also able to protect their own private schools according to this provision of the act.

In Canada there is neither a central ministry of education, as in all European countries, including England, nor an office of education, as in the United States. Instead, the Education Division of the Dominion Bureau of Statistics publishes an *Annual Survey of Education in Canada;* the Canadian Education Association collects and publishes research studies and disseminates information on procedures and practices in the various provinces; and a national research council assists materially in advanced research and investigation at the university level. The dominion government is not without authority, however, in educational matters. For one thing, it is empowered to overrule any act of a provincial government. It protects the educational rights of minorities; it has provided crown lands for the support of education; it is responsible for the operation of schools in the Yukon and Northwest Territories and for Indians and Eskimos; it grants a limited amount of aid for vocational and higher education; it supports research and other types of educational investigation; it provides special grants to increase staff salaries in universities; it supports school broadcasts. (Pressures upon the dominion government to support education at the federal level are similar to those in the United States.)

The structure of Canadian provincial education

Each provincial department of education is headed by a minister of education, who is a member of the provincial legislature and hence of the party in power and who is responsible to the legislature in matters involving broad educational policy. The deputy minister of education is the chief permanent officer of the provincial department of education. Educational administration within each province is a joint function of this provincial department of education and the local school districts. The functions of the province include curriculum, textbook selection, supervision, and finance, while the school district is concerned with teacher appointments, school maintenance, administration of funds, and general school policy. Some effort has been made, over the years, to achieve some federal coordination by such nongovernmental bodies as the Canadian Education Association, the Canadian Teachers' Federation, and other groups. Moreover, since 1968, the Council of Ministers of Education, which is made up of the educational directors of the ten provinces, has been meeting in order to further interprovincial cooperation. The most persistent administration problem in Canada is how to find the means of providing larger and more efficient units of organization and administration, at the same time preserving local control of schooling in a sparsely settled, widely diversified, heterogeneous nation (see below *Sectarian systems*).

Because of the decentralized nature of Canadian education, there are problems in the description of the school system. Preschool education consists of a two-year nursery school and a one-year kindergarten, but these are not attended everywhere. The elementary school, depending upon the province, may be eight, seven, or six years in

length. Secondary education is generally made up of a junior level, grades seven to nine, and the senior level, grades 10 to 12 or 13. Admission to colleges and universities is gained on the basis of a high school diploma, although an entrance examination may be required by some institutions in certain instances. The higher educational sequence offers programs leading to bachelor's, master's, and doctor's degrees in the humanities, social sciences, sciences, and various professional fields.

Australia. The underlying tradition of Australian education, as one would expect from the composition of its population, has been British. During the 20th century, as a consequence of various connections with foreign countries, other influences, especially American, have been added. As its history developed, Australia turned increasingly toward state involvement in education, equality of opportunity, and secularization. According to the constitution of the Commonwealth of Australia, promulgated in 1900, education was to be a function of the six states rather than that of the federal government. However, without the necessity of any constitutional change, there has taken place an increasing federal participation, mainly financial, in educational affairs.

Anglo-American influences in Australian education

The educational system of each state is centrally administered by a department of education operating under a minister of education. The director of the department, together with his staff and the inspectorate, deal with the elementary and secondary schools and with the teachers' colleges. Denominational and other independent schools, though they do not as a rule receive state aid, are subject to some state inspection and control. Since 1968 the Department of Education and Science of the Commonwealth of Australia undertook to administer education in the Australian Capital Territory and in the Northwest Territory, with teachers supplied by New South Wales and South Australia respectively.

Compulsory attendance begins at the age of 6 and extends to 15 in five states and to 16 in Tasmania. As a general rule, elementary and secondary education last six years each.

Australia may be taken as an example of a system in which administration may be decentralized in the states, but in which, nevertheless, there is a high degree of conformity. Many foreign observers and domestic specialists have been critical of what has been termed excessive centralization and authoritarianism in the administration of education in the states. The uniformity in the various schools within each state is praised by some as contributing to the equality of standards but criticized by others as lacking in creativity and in the opportunity for teachers and administrators to exercise initiative and to participate in the process of decision making. There is a likelihood of change as the tendency of parents to take part in school affairs increases. Laymen have hitherto not had any role of significance in Australian education. The formation of lay groups with a concern for educational issues is another signpost toward possible change in the future.

Financing of education remains a problem in Australia as in many other countries. Associated with it is the suspicion that the federal government would interfere in state school affairs if it gave financial aid. The issue of governmental contributions to denominational schools has been the subject of considerable debate; the arguments—double taxation on the one hand and the weakening of the public school system on the other—have been raised in much the same way as in the United States. The question of selectivity of students for higher education, together with difficulty of transfer from one kind of secondary school to another, has also been aired in conference and in print.

JOINT NATIONAL AND LOCAL SYSTEMS

Examples of national and local systems working cooperatively or in partnership are to be found in Great Britain and, to some extent, Japan.

England. The basic point about education in the United Kingdom is that the Department of Education and Science directs education in England and Wales only. The administration of education in Scotland is under the direction of the Scottish Education Department, which functions under the secretary of state for Scotland. In Northern Ireland, the school system is administered by the Ministry of Education in Belfast. The Department of Education and Science in London is also responsible for all the universities of Great Britain. Although Welsh schools are directed by the Education Office for Wales, located in Cardiff, this is a regional unit of the Department of Education and Science in London.

English education has been formed out of a long tradition going back to the Middle Ages and from a series of laws beginning with the Education Act of 1870, which established compulsory, government-financed elementary schools. Government-supported secondary schools date from the Education Act of 1902, which also set up a national system of education in cooperation with local school administration.

Britain's Local Education Authorities and Department of Education and Science

Present-day educational administration in England was fixed by the Education Act of 1944, which lodged final authority in the central Department of Education and Science (then called the Ministry of Education) but which placed immediate administrative responsibility upon Local Educational Authorities (LEAs). The Education Act of 1944 sets the structure of administration and control but allows local authorities to interpret policies in accordance with local need, preference, and internal school conditions. Central authority is vested in three bodies: the department, the Central Advisory Councils, and the civil service, representing the political, the advisory, and the professional elements of education respectively. The minister of education, appointed by the prime minister, is responsible for broad educational policy and practice within the framework of existing legislation as approved by Parliament. The two Central Advisory Councils (for England and Wales) advise the minister on internal matters of administration. The civil service is responsible for the communication and interpretation of broad policy at the local level. The permanent education secretary guarantees continuity of policy and operations. The "eyes and ears" of the ministry are Her Majesty's Inspectors, better known as HMIs, who are members of the civil service. The HMIs have four major duties: (1) to ensure that government appropriations are being spent wisely, (2) to see that educational legislation is being complied with, (3) to act as missionaries disseminating enlightened educational thought, and (4) to keep the ministry informed on what is being done at the grass roots of education.

The LEAs are not elected directly by popular vote but are appointed by county and borough councils, which are in turn elected directly by the people. Actual supervision is, however, delegated to education committees, which are appointed by the LEAs out of their own membership plus a minority of non-LEA members who are interested in education. One of the committee's functions, like that of a local board in the United States, is to appoint a chief education officer, usually called a director of education. The director of education enjoys less power and independence of action than the U.S. superintendent of schools, his chief function being to present proposals to his committee for decision and action, to record decisions and to carry them out. On the other hand, the headmaster of an English school, especially on the secondary level, is a more powerful and influential person than his U.S. colleague; he is given wide latitude in the formulation and execution of internal policy. As the power of the headmaster increases, it is natural that the power of the director of education decreases. Nevertheless, with the aid of a corps of local inspectors and advisers, the director of education commands a post of considerable weight, which takes on added importance in view of the director's responsibility to the national Department of Education and Science, as well as to his local education committee. Moreover, the national offices exercise additional influence through the distribution of official syllabi and teaching handbooks (though headmasters and

teachers are not constrained to follow them) and the distribution of examinations for the General Certificate of Education and the Certificate of Secondary Education. In general, it may be said that in contrast with other leading nations, England presents an educational administrative system that not only operates cooperatively on both local and national levels but also unites with private endeavours and voluntary organizations to spread the responsibility for control throughout the entire political and social structure of the nation.

Compulsory attendance in England begins at 5 and continues to age 15. The school system makes provisions for nursery schools for children aged two to five. Formal school attendance begins with age five, when the child enters the two-year infant school or department. Between seven and nine, the pupil attends the junior school. The seven-year secondary school may be the comprehensive type with three courses—general (modern), technical, and grammar (academic)—or bilateral with two of these courses. This school has been increasingly replacing the three separate secondary schools, into which children were firmly distributed upon taking a selective examination at the age of 11 (the 11-Plus examinations). Of the 146 local education authorities of England and Wales in 1968, 107 had reorganized or approved the reorganization of secondary education on the comprehensive pattern, while 24 others submitted plans for the approval of the Department of Education and Science. There is a strong likelihood that secondary education will be completely comprehensive in due time.

Admission to a university depends on the results on the Advanced Level of the General Certificate of Education, although Oxford and Cambridge insist on special entrance examinations.

Many of the problems of British education revolved about the determination of the government to extend educational opportunity as widely as possible. This policy not only involved finance but also opposition on the part of those who believe in a high degree of selectivity. Also related to this issue was the campaign to alter the complexion of the traditional, independent (public) schools by securing the enrollment of children from the less affluent elements of society. The status of the colleges of education remains a controversial question in higher educational circles. The growing population of Asiatic, African, and West Indian origins added a new dimension—the racial and ethnic—to the problems of British education. Local areas, such as Birmingham, experienced difficulties with newcomers who made unprecedented demands for special consideration of language, dress, and religion in the educational process.

Japan. The modern Japanese educational system dates from the beginning of the Meiji era in 1868 with the adoption of the policy that "knowledge shall be sought for all over the world and thus shall be strengthened the foundations of the Imperial polity." This was translated into a twofold campaign of importing foreign scholars and professional men to instruct the Japanese and of sending selected young men to study in European and American universities. The outcome was the radical transformation of Japan into a modern, technologically competent society that, in time, excelled the developed countries in several respects. However, the traditional Japanese morality, of Confucian origin, was not only retained but also made the ideological cornerstone of the educational system. The collapse of Japan in World War II brought about the introduction of Western, specifically American, educational ideas and organization. With the passing of the military occupation, however, many of these were modified or discarded and once again the question of moral instruction came to the fore.

Multiple influences on modern Japanese education

After World War II in Japan, educational administration tended to follow directives issued by the United States as the occupying power. Organization was established on three levels: national, prefectural, and local community. The central Ministry of Education, though stripped of much of its power, is still, as in England, the chief administrative authority. The ministry has a secretariat and five divisions: elementary and secondary education, higher education and science, social education, research, and administration. Advice on general policy is afforded by 18 councils on education, including a central council for education. The national Ministry of Education is entitled to draft laws in the interest of national educational standards, but its power to instruct local boards is limited by the Diet, Japan's parliament. Since the ministry allots much financial aid to local systems, especially for school buildings, considerable indirect control is added to that which may be exercised directly. In consideration also of the weakness of local boards, due in part to educational tradition in Japan and in part to insufficient funds for vigorous programs at the local level, the ideal of a weak central Ministry of Education advocated by the United States educational mission failed to materialize. Concrete evidence is seen in such developments as the assumption by the central ministry of responsibility for the nine years of compulsory schooling and the security of teachers' salaries and emoluments. Also, the recommended transfer of university control to local authorities has failed, so that since 1954 universities have fallen within the province of the national educational administration.

Unlike England, the roots of local self-determination, especially in educational matters, are not deep in Japan, so that local controls are in a stage of growth rather difficult to delineate. Local boards of education are legally responsible for the establishment and maintenance of elementary and lower secondary schools; while the prefectural board of education administers upper secondary and special schools, the certification of teachers, certain aspects of private schools, and the choice of textbooks. Originally, too, these boards were supposed to be elected, but because educated Japanese people did not wish to run for the office (or could not afford campaign expenses), and because the Japanese themselves did not feel a strong affinity for the elective process, board members became appointive. Another difficulty with implementing the ideal of local controls in Japan has been the traditional method of collecting taxes. It has not been possible to establish a system of special school taxes, so that local boards must depend upon general local revenues supplemented by allotments from the central ministry. Although the trend may be back to a centralized administration in Japan, local authorities still exercise far more control than before the war, so that a modified type of partnership in educational administration may therefore be said to prevail in Japan, in principle if not always in practice.

Problems of local control in Japan

The contemporary Japanese school system consists of a three-year kindergarten, a six-year elementary school, a three-year lower secondary school, and a three-year upper secondary school.

On the higher educational level, there are two- and three-year junior colleges in addition to universities and professional schools. Compulsory attendance is required from ages 6 to 15.

The educational objectives of the Allied occupation of Japan were to discredit and eliminate militarism and chauvinism and to modernize and democratize the school system. The Japanese constitution of 1946 was a step in this direction and was followed more specifically by the Fundamental Law of Education passed in 1947. It is clear, however, that the present system still makes it difficult for young Japanese who are less than brilliant to obtain opportunities for advancement in education and society; many Japanese educational leaders have felt that the democratic principles that had been pushed by the Americans must be supplemented by the inculcation of the ideals of Japanese morality and love of native land. Many Japanese are convinced that education is incomplete without courses in morals. This view, however, has been increasingly opposed by the teachers' organization (Nikkyoso) and by the Federation of Student Self-Government Associations (Zengakuren), which has frequently demonstrated in behalf of Marxist principles and causes.

SYSTEMS CONTROLLED BY POLITICAL PARTIES

The Soviet Union, the People's Republic of China, and Spain are among the prime examples of educational administration as controlled by a single political party or ideological group.

U.S.S.R. The right of all citizens to education is guaranteed by the constitution (article 121, revised), which provides for universal, compulsory education from ages 7 to 15 and for extensive development of secondary education. All forms of education, both secondary and higher, are free of charge; there is a system of state stipends payable to students who have distinguished themselves in higher schools; education in the schools is in the students' native language; and free technical, industrial, and cultural training for the working people is organized at plants, state farms, machine tractor stations, and collective farms.

The Soviet Union resembles India and China in its multiplicity of cultural, national, and linguistic groups. There are nearly 200 such groups incorporated into the union; more than 100 different languages are spoken. Within the Russian Soviet Federated Socialist Republic, by far the largest union republic, are no less than 100 cultural and linguistic groups. Under such circumstances, the U.S.S.R. could not help but organize as a federal state; indeed, one secret of Soviet success may be traced to the willingness of the central government to cater in part to national and cultural differences, including instruction in vernacular languages at primary and secondary levels, with Russian compulsory.

The administration of Soviet education

Until 1966 the Soviet Union had no central ministry for elementary and secondary education, relying instead upon the 15 separate ministries of education in the constituent republics to unify their activities in accordance with educational directives emanating from Communist Party headquarters and the central government in Moscow. In 1966 the Ministry of Education was set up to direct elementary and general secondary education with the same measure of authority carried by the already existing Ministry of Higher and Specialized Secondary Education, which controls all the universities, the vocational and technical secondary schools (*tekhnikums*), and most of the teachers institutes. Each of the republics now has dual ministries subordinate to these two in Moscow. In addition, there are urban, territorial (*kray*), regional (*oblast*), and district (*rayon*) levels of administration that direct educational work in accordance with the rules and regulations of the respective ministries. The underlying administrative principle is that of "democratic centralism," a hierarchical system whereby authority is exercised downward and information is sent upward through the various echelons. Policies may originate locally, but they have to be submitted up the line for approval. The philosophy underlying all policy and practice is that of the Communist Party, and party principles are the same for the entire nation.

Policy deliberations and decisions on all significant educational issues are made by the Central Committee of the Communist Party of the U.S.S.R. and the Council of Ministers of the U.S.S.R., with statements generally issued under joint auspices. The national legislature, the Supreme Soviet of the U.S.S.R., then enacts these decisions into law, which becomes the model for laws promulgated by the Supreme Soviets of the separate republics. The coordination of educational and economic planning is accomplished by the higher educational ministries together with the State Planning Committee (Gosplan) of the U.S.S.R. and of the republics. The outcome of such planning is the establishment of the number of persons to be trained in specific categories in the higher institutions and the specialized secondary schools of the entire country. The centralization of Soviet education is also illustrated by the existence of the Academy of Pedagogical Sciences of the U.S.S.R., which conducts research programs, prepares courses of study and textbooks, and develops methods of instruction.

Influence of the Communist party

The Communist Party not only sets policy but also determines who may be appointed or elected to educational office and exerts controls in a number of complex ways. Local school boards, for instance, are appointed by executive committees of the local soviet concerned. The local director of education is likewise appointed by the executive committee of the council. School principals are, however, named by local school boards; and principals in turn select the teachers, subject to approval of the school board. The school principal is also in charge of school purchases; but he must buy all supplies from government houses. The internal duties of local school boards resemble those of most countries, but they must look to higher authorities for the determination of overall policy and practice and have no authority to levy taxes. Each school is guided by a council consisting of parents, teachers, Communist Party members, and others connected with school life, including clerks and janitors. These councils occupy themselves vigorously with the ongoing activities of the school, uniting with parent-teacher associations in physical (as well as ideational) projects, such as cleaning and redecorating classrooms and providing benefits and perquisites not forthcoming from governmental sources.

In addition, there is an inspectorial system (under the control of each republic ministry), which interests itself in the improvement of teaching and classroom management. Although inspectors must see that party policies are carried out, in practice, they interest themselves primarily in modern pedagogical methods and in supervision; they also teach demonstration classes. Inspectors cooperate closely with school principals; they do not "rate" teachers, except to recommend awards or special honours for particular tasks well done.

The Soviet school system begins with a nursery or crêche (*yasli*) for children between the ages of three months and three years. The kindergarten (*detsky sad*) provides additional preschool education to children between three and seven. Then follows an eight-year school (*vosmiletnyaya shkola*), which covers the compulsory attendance period of 7 to 15. The first four years are considered elementary (*nachalnaya shkola*), whereas the remaining years constitute the first stage of the secondary school (*srednayaya shkola*), which is two years in length. Beginning with the fall of 1969, a new change in elementary education, first in Moscow and eventually through the Soviet Union, makes the elementary school three years long and the secondary school seven years and available to all. Admission to a higher educational institution is achieved after the completion of the ten-year school with a certificate of maturity (*attestat zrelosti*) and the passing of a highly competitive entrance examination. The possession of recommendations from officials testifying to service in the Communist Party of Komsomol, the trade unions, agriculture, industry, or the armed forces is a vital aid to acceptance to higher education, enabling the applicant to enter on the basis of a less rigorous selective process.

Finally, Communist education naturally has the task of preparing future leaders, for whom special party schools are established. In all schools the party exerts its influence in the organization of pupils in various types of political youth leagues, with Pioneer groups for ages between 10 and 15 and the Komsomol, or Young Communist's league, for youth over 14. Members of these groups are chosen for their loyalty to party principles and their actual or potential capacity for leadership. In sum, administrative policies in the U.S.S.R. may sometimes originate in local areas, but, even then, to become effective they must be approved at the top. Approved policy then sifts down through the various levels, and its enactment is overseen by party members or affiliates, who penetrate every phase of public and educational life.

People's Republic of China. Practices similar to those in the Soviet Union prevail in China, despite the present ideological split. Following World War II Chinese schools were reorganized under the aegis of Russian counsellors and modelled after the educational ideals of the Soviet Union. Articles 41 through 47 of chapter 5 of the Common Program contain the core of Chinese ed-

Centralization in Communist Chinese education

ucational policy. Education, according to this program, shall be "national, scientific, and cultural." It shall promote the five "loves": fatherland, people, labour, science, and public property. Education and politics become inseparable, so that educational leaders are either members of the Communist Party or pledged to support its principles. The entire nation is organized under such centralized ministries as finance, trade, labour, heavy industry, fuel industry, textiles, foods, light industries, railways, communications, agriculture, conservation, forestry, and land reclamation. The listing of these ministries is not merely a verbal duty; it reveals the nature of Communist China; and the school system is charged with providing trained personnel for all these ministries.

As in the Soviet Union, the administration is highly centralized under a Ministry of Education and a Ministry of Higher Education in Peking. The former has jurisdiction over the nursery, kindergarten, elementary, middle, part-time, and teacher-training schools; the latter, over the universities, colleges, higher technical schools, and research institutes. It is worthy of note that there was a single ministry controlling all education in 1949, two in 1951, one again in 1954, and two once more in 1964. As in the Soviet Union, private schools and colleges are not permitted to exist, thus assuring complete educational centralization.

The administrative structure in Chinese education is similar, in terms of hierarchy, to that of other countries: apart from the national ministry, there are educational authorities on the provincial (*sheng*) and the county (*hsien*) levels. School administration is said to be "democratized" by the organization of educational councils, in which teachers, students, and workmen enjoy equal voting status. The democratic way is also said to be practiced in all kinds of meetings and discussion groups, findings of which are transferred for action to higher authorities on the basis of "discussion by the many but decisions by the few." Heading these educational discussions, or influential in their operation, however, are members of the party. When one compares two such highly nationalized systems as China and France, political preference is readily discerned as perhaps the chief agent for ordering educational ends, even though the means of administration are similar.

School organization in Communist China does not differ substantially in structure from that of China prior to 1949 or from that of Taiwan. The major differences are in the underlying philosophy, the content and method of instruction, and the administrative control. The nursery is for children up to the age of four, while the kindergarten is for those aged four to seven. Formal elementary education extends for six years, but attendance is not compulsory by law, inasmuch as schooling, which is regarded as a privilege, is virtually universal. Two three-year middle schools provide general secondary education. Of special significance are the half-work and half-study schools, which seek to bring about the synthesis of the labourer and the intellectual. As in the Soviet Union, the children and the young people obtain political education from the Young Pioneers and the Communist Youth League. Students who are not admitted to either branch of the middle school may enroll in a vocational, technical, or normal school, or go directly to work. Admission to the university is on the basis of the national need of specific specialized personnel, an entrance examination, political status, and social (*i.e.*, proletarian) origin.

The aim of Communist Chinese education comprises the development of the new Chinese man who will be imbued with the Marxist-Leninist principles and who will use his knowledge and ability for the cultural advancement of the country and for the perpetuation of Communism. The Chinese authorities themselves often speak of education as aiming at the achievement of the status of "both red and expert," meaning one who is simultaneously a dedicated Communist and an accomplished specialist.

Spain. In recent decades Spain has attracted less notice in educational affairs, in part because of the powerful control exercised by the regime of Francisco Franco and by the conservative Spanish wing of the Roman Catholic Church. Recognizing that in an increasingly technological world persistent illiteracy, unequal educational opportunity, and other practices handicap the country, Spain's legislature, the Cortes, adopted in 1970 a new law to reform Spanish education. The tempo of educational change, evidently, is increasing: as recently as 1964 compulsory attendance was raised to age 14 (in law if not always in practice), and the Technical Education Act provided for advanced study of technical subjects.

Like the classic French system, Spain is divided into educational districts—12 of them—each with a university as a centre. All districts are subordinate to the central Ministry of Education and Science (Ministerio de Educación y Ciencia) in Madrid and liable to periodic or ad hoc inspection by its representatives. This central control applies to private as well as public schools. There are not only a hierarchy of inspectors—such as *inspectores generales* and provincial *inspectores centrales* and area *inspectores*—but also a complex of "technical advisers" who share in guiding teachers in the performance of their work. Although universities are outside the system of state inspectorates and in theory are autonomous, the rectors of all the universities are called to Madrid for regular meetings with the minister of education and science in order to arrive at important decisions.

The Spanish "academic" school system consists of a two-year nursery school (*escuela maternal*), a two-year kindergarten (*escuela de párvulos*), a four-year elementary school (*escuela primaria*), and a seven-year secondary school (the public *instituto* or the private *colegio*). Admission to an academic secondary school, which is not free, is possible after four years of elementary education if the pupil obtains a certificate (*certificado de estudios primarios*) after an examination. Secondary education is divided into three cycles: a four-year course leading to the diploma of *bachillerato general elemental;* a two-year course ending with the *bachillerato general superior;* and a one-year *curso preuniversitario*, the successful completion of which entitles the student to enter any faculty in a higher educational institution. In addition to the academic schools, Spain has a network of many vocational, trade, industrial, technical, commercial, and art schools on the different levels of instruction. Elementary school teachers are prepared in a secondary-level institution (*escuela de magisterio*) with a three-year program. Secondary school teachers must hold the university degree.

Political and religious influences in Spanish education

The fundamental aim of education in Spain, the development of a literate citizenry, is similar to that of modern nations generally. Special attention, however, is given to the formation of loyalty to the political ideology of the government and to the principles of religion as taught by the Roman Catholic Church. The Spanish state guarantees the teaching of the Catholic religion as a regular and compulsory subject in all educational institutions, whether state controlled or not, and no matter what their level and purpose. Any publication at all may be suppressed if it is contrary to Catholic dogma and morals. Also, in tests for the selection of teachers, for example, examining boards are composed of five members, three of whom, including the chairman, must be church officers. It must be said, however, that religious minorities have been receiving increased privileges in recent years with regard to education. In the early 1970s, moreover, the Spanish bishops had decided to separate themselves from politics and concentrate more on religious matters.

(W.W.B./G.F.K.)

SECTARIAN SYSTEMS

Religious and denominational influences control school administration in various nations and regions of the world—including Spain (as noted above), Ireland, certain provinces of Canada, and such Muslim countries as Pakistan. In some other countries, such as the United States, ecclesiastical and other religious bodies maintain elementary and secondary school systems separate from the public school systems.

Ireland. The aim of Irish education, as expressed in the republic's constitution of 1937, echoes the principles propounded by Pope Pius XI in the encyclical *Divini Illius Magistri* (1929), on the "Christian Education of Youth." Fundamentally, the government recognizes the natural right of parents to direct the education of their children either at home or in state schools. While the Roman Catholics, who comprise 97 percent of the population, are accorded a "special position," the state also recognizes the existing Protestant sects, the Jewish denominations, and the other religious groups. This recognition occurs not only in the freedom of religious instruction but also in the form of financial subsidies for the schools of minority religions. Religion is a component part of all curricula, including vocational education, of all institutions without regard to religious group, since there are no neutral or secular schools in Ireland. However, in line with article 44 of the constitution of 1937, the right to be absent from religious instruction is guaranteed.

Roman Catholic involvement in Irish education

Educational administration is centralized in the Ministry of Education, which issues a manual, *Rules and Regulations for National Schools*, and courses of study. The national inspectors, with the authority of the Ministry of Education, perform a critical and an instructive function in the effort to ensure compliance with the educational directives in the national or state-supported schools.

Compulsory school attendance extends between the ages of 6 and 14. After the preschool, the child enters the six-year elementary school, where he obtains the basic learning—Irish, English, arithmetic, and other subjects required for later education. Then follows the secondary educational sequence of four to six years, depending on the academic or vocational nature of the program. Entrance into a secondary school, until recently, depended on the possession of a Primary School Certificate or the passing of a special examination; now a school principal's report card is generally sufficient. Those pupils who do not go on to the secondary school complete their compulsory education requirement by attending extended primary classes, the "Secondary Top," and may even continue with secondary courses to the age of 18, at which time they may apply for the state examinations leading to the Leaving Certificate and admission to higher education. The secondary schools, many of them operated by Roman Catholic religious orders, though free, are under private control. (W.W.B.)

Canada: Quebec and Newfoundland. As noted in the discussion of Canada under *Decentralized systems* above, the British North America Act of 1867 delegated jurisdiction over education to the provinces. In the provinces of Quebec and Newfoundland, administration is sectarian in nature, though the schools are otherwise "public." Education is free and compulsory, but children must attend the school of their denomination, unless an "outsiders' fee" is paid. Quebec developed with two distinct groups: the Protestant English and the Roman Catholic French, each demanding the right to educate its members. There was no minister of education in Quebec until 1964 when such a ministry was established; the dual Catholic-Protestant system was retained. School districts may be organized by minorities in any community, if there are a sufficient number of pupils to justify the procedure and a sufficient number of taxpayers capable of assuming the responsibility. This principle of "dissent" is guaranteed by law. Individual taxpayers may also designate the school system to which their taxes must be paid. The French-speaking school districts in Quebec are generally coterminous with the Roman Catholic parishes; otherwise, it is obvious that political districts and school districts are not necessarily identical.

Denominational schools in Quebec and Newfoundland

In Newfoundland, where early settlement took place along denominational lines, nearly every local community has a majority of one or another of the various denominations. This situation has produced a sectarian system of education, whereby the churches of the various groups have built and in large part control their own schools. Directors of education are nevertheless responsible to the provincial ministry of education. Increasingly the schools of Newfoundland are becoming consolidated, however, especially at the secondary level, reflecting a national trend. (G.F.K.)

Pakistan. An Islāmic system of education can be exemplified by Pakistan, which, in its constitution of 1949, proclaimed that sovereignty and authority were derived from Allāh. Initially, a dominion in the British Commonwealth of Nations, Pakistan became an Islāmic republic in 1956.

Islamic influences in Pakistan

Education in Pakistan is dedicated to the perpetuation of the ideas and values of Muslim religion and society, a tradition dating from the rise of the Muslim faith in the 7th century. With over 88 percent of the population adherents of Islām, it is understandable that considerable attention is paid in the schools to the teachings of the Qur'ān and the Ḥadīth (oral tradition), which are the basis of orthodoxy (*sunnah*). At the same time, as a developing nation desiring to modernize itself as rapidly as possible, Pakistan has organized schools and higher education that would prepare personnel to help achieve this aim.

The responsibility for education is vested in the central Ministry of Education, which sets and proclaims policies and coordinates the activities of subordinate bodies. When Pakistan was divided into two provinces, East and West Pakistan, separated by more than 1,000 miles of Indian territory, it was necessary to set up two distinct provincial departments, each headed by a secretary of education. When the eastern province broke away and became the new state of Bangladesh, there remained only the single secretary in Karachi. Thus, Pakistan now has six regional directorates of education, which report to the director of public instruction in the provincial department. To some extent, the local school authorities share some of the educational power.

There are three forms of educational facilities in Pakistan: public schools and colleges, fully financed by the central or provincial authorities; public schools, supported jointly by central, provincial, and local governments; and private schools, which may be independently endowed or which may receive government aid for buildings and maintenance. The traditional schools, *maktab* (or primary school), *madrasah* (or general school), and *dār al-ʿulum* (or "house of sciences") did stress Islāmic religion and culture and did not favour the standard subjects of the humanities, mathematics, and the like, but even they are becoming more secularized. The secular component of Pakistani education consists largely of the various subjects taught in modern school systems, even though due recognition is still also given to the moral and spiritual values of Islām. The structure of the secular system is made up of a five-year elementary school for children aged five or six, a three-year middle school, a two-year high school, and a two-year intermediate college. Limited preschool education exists mainly in the larger cities. Pakistan has not yet passed a compulsory school attendance law.

Education in Pakistan underwent a crisis as a result of the exchange of populations with India during the establishment of the country. Among the pressing contemporary problems is that of illiteracy. As of 1961, 84.1 percent of the total population were illiterate. The highest rate of illiteracy was 96.6 percent in the Lasbela district of West Pakistan. Even with a reduction of illiteracy during the 1960s, the problem was far from being solved. Alongside this is the issue of the uncertainty of a national language, with implications of difficulties for education. The elementary school curriculum is composed of reading and writing a national mother tongue: this is usually Urdu; English is optional. (W.W.B.)

United States: parochial education. Many church parishes and other religious bodies in the United States maintain elementary and secondary schools that provide education based on sectarian principles and are separate from the public, or common, school system. Such schools for the children of members of a parish or congregation are supported principally by the voluntary offerings of the parishioners, augmented occasionally by modest fees; they are distinguished from public schools operated un-

der the direction of school boards made up of local citizens and supported by public taxes and from private schools maintained by private individuals or groups and supported principally by tuition fees. These parochial systems or schools exist for the three major religious divisions in the United States: the Roman Catholics, the Protestants, and the Jews.

Roman Catholic parochial schools. These schools in the United States are organized on a diocesan basis and are financed by the parish. In some areas they may serve the children of several parishes or a district defined by the bishop. Such schools are called diocesan schools and are financed by the several parishes involved and sometimes by diocesan funds. The administration of the schools of a diocese is the direct responsibility of the bishop. This work is usually delegated to a priest appointed by the bishop, who is given the title of superintendent of schools or secretary for education. His functions closely parallel those of the chief school officer in the public school system. He is responsible for the organization and supervision of the schools of the diocese, selection of textbooks, preparation of courses of study, policies and practices of all the schools, and enforcement of diocesan and state regulations. His office and authority are defined by the bishop.

Roman Catholic schools in the U.S.

Most dioceses have a school board made up of members appointed by the bishop. Usually this board has no administrative power but acts in an advisory capacity to the bishop and his chief school officer.

The Roman Catholic schools teach the same subjects as the public schools, but two important differences should be noted. In the Catholic school, a course in Christian doctrine is included that presents the truths of faith and morals that are the teachings of the Roman Catholic Church. Other courses, particularly in such areas as the social studies and humanities, are often given a Catholic orientation.

Oftentimes diocesan systems organize on a statewide or regional basis in order to share certain educational advantages and to exchange ideas and programs for the betterment of their teachers and schools. All diocesan systems are members of the National Catholic Educational Association, a nationwide organization established for the mutual benefit of all schools and teachers. The common interests of the Catholic bishops of the United States in education are represented by the department of education in the National Catholic Welfare Conference.

Protestant parochial schools. Among Protestant groups that maintain parochial school systems in the United States, the Lutheran bodies are by far the most active, with around 190,000 children in Lutheran schools. Smaller school systems are sponsored by the Seventh-day Adventists, the National Union of Christian Schools, the Protestant Episcopal Church, the National Association of Christian Schools (evangelical), and the Quakers, Mennonites, Baptists, Methodists, and Presbyterians.

Protestant schools in the U.S.

The administration of most Protestant schools is in the hands of a board of education elected by the sponsoring group—either individual congregation, association of congregations, or association of individuals. This board selects a principal and in cooperation with him develops the policies that control the school program. The clergy of the sponsoring groups have been active participants in all aspects of such school programs.

The teachers in Protestant schools are drawn from denominational teachers colleges and liberal arts colleges. In addition to meeting the general requirements set for teachers at the elementary and secondary levels, teachers in Protestant schools need training in Bible and doctrine to achieve the objectives of Christian education.

Protestant elementary and secondary schools strive to meet or exceed the educational requirements of the several states. While the course of study for each grade is substantially the same as given in the public schools, Christianity is made a unifying and integrating factor in the educational program.

Jewish parochial schools. The Jewish day school or "complete" school (*i.e.*, the Jewish parochial school) in the United States is an educational institution in which a combined program of Jewish and general studies is offered. There are four types of institutions that provide intensive Jewish religious education as well as a complete course of study in general subject matter. These are the Talmudic day school, the modern day school, the integrated day school, and the Hebrew-English private school.

Jewish day schools and private schools in the U.S.

The Talmudic day school makes provision for the study of the Hebrew texts of the Pentateuch, major portions of the Prophets and Scriptures, including the Masoretic commentaries, and places great stress on teaching of the Talmud. The secular studies meet the requirements of the city and state authorities. The school day ranges from eight to ten hours. Hebrew studies are given in the morning hours and the English in the afternoon.

The modern day school stresses the study of Hebrew language, the Bible, Jewish history, prayers, and selected portions from the Talmud. All instruction is in the Hebrew language. The secular subjects generally are given equal attention and equal time.

The integrated day school aims to achieve a blending of instruction in Judaism and secular subjects. To attain this end "both curricula are offered departmentally throughout the school day." The pupil may begin the school day with the study of the Bible, go next to a lesson in arithmetic, then proceed to the reading of a Hebrew text, then to geography, and so on.

The Hebrew-English private school emphasizes the study of general subjects and provides opportunities for Jewish learning for only five hours a week. Its aim is to be a modern and progressive school that satisfies those parents who seek private instruction of a high calibre and who are eager to solve the problem of a two-school system. A number of day schools are also sponsored by Yiddish culture groups. The program of these schools differs from the modern day school in that Yiddish language and literature are included as basic subjects of instruction.

The majority of day schools (85 percent) are sponsored by Orthodox groups. Some that fall in the category of modern day schools are conducted by Conservative and other groups. By the 1960s enrollment in the United States and Canada had increased to more than 50,000 pupils in about 300 schools, almost one-half of which were in New York City. The number of children in the day school constituted 8.5 percent of the total Jewish religious school enrollment. The overwhelming majority of Jewish children attend the public schools, and for them Jewish education is supplementary to their general education. Most day schools conduct classes on the elementary level; only a few operate high school departments. Torah Umesorah (National Society for Hebrew Day Schools) and the National Council for Torah Education help to establish day schools, place teachers and administrators, conduct seminars and workshops for educational personnel, and publish Hebrew text materials.

(Ed.)

III. Trends and prospects

Worldwide, many educational trends and changes are clearly evident, even though there remains a confusion of policies. Of particular international concern is the modernization and growth of the developing nations. The interest in, and responsibility for, the development of these countries has been manifested by the industrialized nations, by intergovernmental bodies, nongovernmental groups, and higher educational institutions. The newer countries are beset by problems that cannot be solved by any form of instantaneous action. One such problem is the "brain drain," whereby young students and professionals of underdeveloped countries trained in advanced countries find socioeconomic conditions there so much more favourable to their growth that they fail to return and help their own people to attain their national objectives. It is clear that the underdeveloped countries cannot rapidly become developed under such circumstances. In a number of instances, however, the changes under way hold out promise for the better utilization of native talent

Developing nations

educated abroad. The growth of indigenous universities, together with the multiplication of suitable employment opportunities and the recognition of the academic and professional attainments of natives, may operate, in the years to come, in favour of the development of new countries so that they will be decreasingly dependent on the industrialized nations.

Science and humanism

The emphasis from the 1950s to the 1970s on science and technology in government and society had its repercussions on all levels of education from the primary through the postgraduate. The spectacular advancements in the various branches of theoretical and applied science contributed to the growth of interest in such studies. At the same time, some observers in different countries began to call attention to the relative neglect of and lack of interest in the areas of the humanities and the social sciences. Concern was expressed in many quarters at the decline of values and morals among young and old; the rise of violence on local, national, and international scales; and the growing prevalence of pollution of the environment, with the resultant dangers to man everywhere. Some have questioned the governmental policies whereby large shares of the budget were allocated to science and technology, while insignificant sums were granted for culture and general education. The malaise that seemed to have gripped the economically advanced nations also appeared to have engulfed the developing countries, which desired to take a giant step forward and enjoy very quickly what they conceived to be the benefits of modernized living. It was obvious to critics of the contemporary scene that some reassessment would have to take place. Just as the scientific and technological revolution took place as a reaction to a rather fixed society and order of things, it seemed desirable, perhaps inevitable, that there would be a resurgence of the humanistic values in life and education. Many hoped for and looked to a better balance of what C.P. Snow called the two cultures, both in the older and in the newer nations.

Early childhood education

One of the noteworthy trends in recent years has been the increase of attention to, and of facilities for, preschool education, especially prior to the kindergarten period. Some countries, such as the Soviet Union, have placed considerable weight, for social and economic as well as educational reasons, on the years of early childhood. Although the idea was not really new—one thinks of the experiments in nursery education in Britain and elsewhere in the 1920s—it is only relatively recently that a trend along these lines has developed on a wider geographical scale. To some extent, this is a response to the socioeconomic conditions that require the entrance of women into the labour force. But one cannot overlook the cumulative effect of the research findings and public statements by educators and child-development specialists—psychologists, psychoanalysts, psychiatrists, sociologists, and social workers. It is fair to speculate that, in spite of the added cost to all countries, the nursery school will be increasingly recognized as a basic component of national educational systems in various parts of the world. There may be some opposition on the part of parents and professionals who feel that the family is the fundamental educational institution, and such a view cannot be discounted. However, it is possible that the values of a socialized educational experience for every child, for part of the day at least, will find favour among all types among the population (see also PRESCHOOL EDUCATION).

The new media

The development of mass communications through new media has made more people than ever aware of the educational potentialities of radio, television, and devices for rapid reproduction of printed and visual materials. Some advantage has already been taken of the newer mass media for the wider circulation of information and ideas. Workers in the field of adult education in many countries have made use of them in the teaching of literacy, liberal arts, cultural experiences, and vocational work. In schools, even in developing countries, one can nowadays find records, tapes, projectors, screens, films, and various other devices that serve as aids in teaching the sciences, social studies, and other areas of knowledge. In many countries, teaching machines and computers have made their appearance. It is easy to understand that the new devices demand an increased expenditure, a factor that might inhibit adoption in the more impoverished areas. There are some educators and others who look upon the mass media as something less than an unmixed blessing, especially since these methods might facilitate the dissemination of untruth as well as truth, misunderstanding along with understanding. The likelihood, in general, is that the newer mass media will persist as helps to the educational process, just as the book and the blackboard became recognized as fixtures in the schools of the past.

Some note has already been taken of the increasing impact of science and technology on the mind and world of modern man. The advent of automation in industry and in other aspects of contemporary life has also given rise to reflection in many circles all over the world. There is little doubt that large numbers of persons are enjoying what they regard as benefits of living, travelling, and consuming brought about by the introduction of automation in industry and society. The reduction of the amount of time spent at work by the average man in advanced countries has resulted in an increasing awareness of the importance of the wise use of leisure, as well as of the freeing of the mind and soul of man from the uniformity and mechanization of life. Many educators and others who followed this development with concern gave more thought to the importance of adult education and free-time activities. The availability of time, due to a shorter work day, an outcome of labour-union activity, industrial automation, and other factors, made possible the planning of constructive and meaningful recreational activities and educational programs. Such a situation was conceivable, as a rule, in an economically advanced country, rather than in one which was still undergoing development. Yet, in the developed countries, there did not seem to develop a mass movement toward the maximum utilization of time released from labour for benevolent recreation, cultural development, and educational growth. In far too many cases, in different parts of the world, leisure time was employed in an aimless, monotonous pursuit of nonconstructive and superficial activity. The process of reorienting the general public toward a voluntary acceptance and appreciation of the value of applying the spare time, or part of it at least, toward enhancing one's educational and cultural level will remain an arduous one.

The uses of leisure

In view of the fact that automation will doubtless continue and even develop further, there is a definite obligation for a greater expenditure of effort on the part of educators and others concerned with a stable society to convince men and women of the necessity of the wise and constructive use of leisure time.

BIBLIOGRAPHY. O.C. CARMICHAEL, *Universities: Commonwealth and American: A Comparative Study* (1959), an analysis of the backgrounds, theory, administration, and problems of higher and professional education in English-speaking countries; J.F. CRAMER and G.S. BROWNE, *Contemporary Education: A Comparative Study of National Systems*, 2nd ed. (1965), an analysis of the educational administration, organization, supervision, operation, and finance of schools in ten countries in Europe, Asia, North America, and Australia; M.A. ECKSTEIN and H.J. NOAH (eds.), *Scientific Investigations in Comparative Education* (1969), essays by various specialists illustrating the methods of research, with examples of comparative studies of socioeconomic-political issues in education; A.H. HALSEY, J. FLOUD, and C.A. ANDERSON (ed.), *Education, Economy, and Society* (1965), a collection of scholarly essays on the relation of economic and social factors and issues to education in various countries; N.A. HANS, *Comparative Education*, 3rd ed. (1958), a study of how racial, linguistic, geographic, economic, religious, philosophical, and other factors affect education, especially in England, the U.S., France, and the U.S.S.R.; B. HOLMES, *Problems in Education: A Comparative Approach* (1965), a study of socioeconomic issues in education in international perspective, and case studies of educational problems in the U.S., Britain, the U.S.S.R., and Japan; T. HUSEN (ed.), *International Study of*

Achievement in Mathematics, 2 vol. (1967), a unique, cooperative, international project evaluating mathematics achievement of pupils in 12 countries; I.L. KANDEL, *Comparative Education* (1933), a classic, influential work, on the theory of comparative education and on the comparative analysis of educational administration and organization and teacher preparation in elementary and secondary education in England, France, Germany, Italy, the U.S.S.R., and the U.S.; A.M. KAZAMIAS and E.H. EPSTEIN (eds.), *Schools in Transition: Essays in Comparative Education* (1968), an anthology of studies of social, political, economic, and cultural questions in education in comparative perspective; E.J. KING, *World Perspectives in Education* (1962), an analysis of secondary and higher education, religious education, and other educational issues in an international context; A.E. MEYER, *The Development of Education in the Twentieth Century*, 2nd ed. (1949), detailed surveys of 20th-century trends in progressive education in Europe and the U.S., as well as historically oriented overviews of education in modern England, France, Germany, Italy, the U.S.S.R., and the U.S.; H.J. NOAH and M.A. ECKSTEIN, *Toward a Science of Comparative Education* (1969), the history and methodology of research in comparative education with selected exemplary studies; R. POIGNANT, *L'Enseignement dans les pays du Marché Commun* (1965; Eng. trans., *Education and Development in Western Europe, the United States, and the U.S.S.R.: A Comparative Study*, 1969), a lucid comparative analysis making frequent use of statistical data; R.M. THOMAS, L.B. SANDS, and D.L. BRUBAKER (eds.), *Strategies for Curriculum Change: Cases from 13 Nations* (1968), studies of the process and direction of change in courses of study in Asia, Europe, Africa, and the Americas; I.N. THUT and D. ADAMS, *Educational Patterns in Contemporary Societies* (1964), historical and contemporary analyses of educational development in developed and developing nations in Europe, Asia, Latin America, and Africa; R. ULICH, *The Education of Nations: A Comparison in Historical Perspective*, rev. ed. (1967), historical traditions from the Middle Ages which have had a deep impact on the educational systems of England, France, Germany, the U.S., and the U.S.S.R.; UNESCO, *World Survey of Education*, vol. 4, *Higher Education* (1966), a significant reference work which furnishes the basic data necessary to understand the workings of universities all over the world.

Periodical literature: International Bureau of Education Bulletin (quarterly), useful for annotated bibliographies of educational works published in several languages; *Comparative Education Review* (3/yr.), an important journal containing various types of scholarly studies in comparative education; *World Year Book of Education* (annual), an indispensable record of almost all aspects of educational development in the various parts of the globe.

(W.W.B.)

Education of the Exceptional

In the context of education, the term exceptional children refers to children who deviate socially, mentally or physically from the average to such an extent that they require major modifications of usual school practices in order to develop to their full capacity. Approximately 10 to 15 percent of all school children are considered exceptional in these terms.

Special education

To promote their maximum development or to lessen their dependence on society, handicapped and gifted children are often provided with programs of "special education." Special education involves adaptations of the curriculum, the teaching techniques, or the organizational program, rather than a necessarily totally different program. It refers only to those aspects of education that are unique or are added to the regular program. The amount and kind of adaptation needed depend upon many factors, including the degree of deviation and the type of deviation.

TYPES OF VARIATION AND HANDICAP

The gifted. Children who are superior in intellectual ability often function two or more years ahead of the grade level in which they are normally placed. It is not unusual to find a gifted child of 10 who is reading at the level of age 14 and who is interested in science and history well beyond his age level.

The gifted have been variously defined as those having IQ's on an individually administered intelligence test above 125, 130, or 140, but since the IQ definition alone is not usually considered to be conclusive, other criteria are used.

The mentally retarded. Among intellectual deviates are children variously called mentally retarded, intellectually limited, or subnormal. (The older terms feebleminded and mentally deficient are not generally used in education.) These children appear to be substantially below the average child in learning ability and find it difficult to cope with the program of the regular grade levels.

The mentally retarded are generally classified according to the degree of mental subnormality they exhibit and their potential for learning and adjustment. For educational purposes the majority are grouped in school in classes for either the "educable" or the "trainable" mentally retarded. The classification "educable mentally retarded" covers those children whose IQ's are roughly between 50 and 75 and who can be educated (1) academically, to the level of an average child of 9 to 11 years of age; (2) socially, in adjustment to the home and community; and (3) occupationally, in being able to support themselves totally or partially at the adult level. The classification "trainable mentally retarded" covers those children whose IQ's are roughly between 25 and 50, but who can be trained to learn (1) self-care skills such as dressing, feeding, and toiletting; (2) social adjustment in the home or immediate neighbourhood; and (3) some degree of economic usefulness in the home and neighbourhood or at routine tasks in a sheltered environment under supervision. The "totally dependent" retarded child (IQ below 25) is generally under health or nursing care auspices.

Educable and trainable retarded

Sensory handicaps. The two sensory handicaps most affecting children's education are impaired vision and impaired hearing.

The visually impaired. A person is defined as "blind" if his vision (after correction) is 20/200 in his better eye. This degree of visual acuity is in general inadequate for education through the eyes, and special techniques have been devised to make possible education through tactual and auditory channels. A person is defined as being "partially sighted" if the vision in his better eye (after correction) is less than 20/70 but better than 20/200. Such children need special equipment and are often taught in special classes or resource rooms that provide special methods and materials. In many cases they can be educated in a regular class if special materials and equipment are provided.

Children with impaired hearing. Like children with impaired vision, the deaf and the hard of hearing fall into separate categories. Those designated as "deaf" are so impaired in hearing that they have to be educated exclusively through other senses. The "hard of hearing," on the other hand, have sufficient hearing to be educated through the auditory channel though amplification devices such as hearing aids are needed.

Orthopedic and neurological handicaps. *Orthopedic and special health problems.* Children who are crippled or orthopedically handicapped and those with special health problems include many categories. The crippled child is one who has an orthopedic impairment interfering with the normal functions of the bones, joints, or muscles to such an extent that special arrangements must be made by the school to transport him and to adapt him to a school environment. Included in the category of the crippled are children born with such handicaps as dislocated hips or joints, club feet, and spina bifida and children born normal who acquire a crippling condition through accidents or through such infections as poliomyelitis, osteomyelitis, and tuberculosis of bones or joints.

Some children have special health problems that do not result in motor handicaps. These include cardiac conditions, tuberculosis, delicate constitution, and chronic illnesses that may nonetheless require special adjustment of school programs.

The neurologically impaired. Neurological impairment probably affects more children than any other cause

of exceptionality. It occurs in many forms and in many degrees of involvement. Not a disease, it is a condition affecting the nervous system. It may involve intelligence, the neuro-motor functions, perceptual abilities, the senses, bodily balance, or any combination of neurological involvements. Cerebral palsy, the most commonly diagnosed neurological impairment, results in a disturbance in the development of neuromotor functions and is sometimes associated with deficiencies in vision, hearing, and speech. Intelligence is also affected in half these children. Usually they are either spastic or suffering from extremely jerky movements of the limbs (athetoids).

Cerebral palsy

Epilepsy is a specific form of neurological impairment resulting in recurring motor, sensory, or psychic malfunctioning of varying frequency and varying degrees of intensity. It has been stated that 50 percent of all epileptics can be kept free of convulsions with the proper administration of drugs and that 80 percent can lead normal lives and even engage in competitive work. The control of epilepsy by drugs has, of course, resulted in a decrease in the number of schools for epileptics, since most of these can be educated in regular grades and school situations.

Epilepsy

Communication disorders. *Speech defects.* Speech is considered defective when it deviates from the average enough to draw unfavourable attention to the speaker, whether through unpleasant sound, lack of intelligibility, or unappropriateness to the age level of the speaker. Such defects include disorders of articulation, disorders of voice, delayed speech, stuttering, and speech disorders resulting from cleft palate, cerebral palsy, or hearing loss. A large percentage of speech problems respond to therapy, and many may be completely corrected.

Learning disability. Learning disability as an educational term refers to a disorder or delayed development in the processes of communication or academic achievement and has been defined by the National Advisory Committee for Handicapped Children as follows:

> Children with special learning disabilities exhibit a disorder in one or more of the basic psychological processes involved in understanding or in using spoken or written language. These may be manifested in disorders of listening, thinking, talking, reading, writing, spelling, or arithmetic. They include conditions which have been referred to as perceptual handicaps, brain injury, minimal brain dysfunction, dyslexia, developmental aphasia, etc. They do not include learning problems which are due primarily to visual, hearing, or motor handicaps, mental retardation, emotional disturbances, or to environmental disadvantage.

Social and emotional disorders. Some children are either unable or unwilling to conform to accepted norms of behaviour, and if these deviations are serious or extreme or persistent, they are considered behavioural disorders. Basically, such problems fall into two categories: (1) emotional maladjustment, in which the child shows inner tensions, anxiety, neuroticism, or psychotic behaviour, and (2) social maladjustment, in which the child shows defiance and disregard for constituted authority and general antisocial behaviour. Although there is sometimes an overlap between the emotionally disturbed and the socially maladjusted, emotional disturbance basically refers to the psychological characteristics of the child, whereas social maladjustment refers to the relations of the child to society as a whole. In many instances a child may be well adjusted, conforming to his peer group or gang but quite maladjusted with regard to the larger society.

PATTERNS OF SPECIAL EDUCATION

Historical background. Although some attention was given to the care and protection of handicapped children shortly after the beginning of the Christian era, special education in Western society did not truly begin until the 16th century. Pedro Ponce de León (1520?–84) successfully taught some deaf pupils in Spain to speak, read, and write; and it is assumed that his methods were followed by Juan Pablo Bonet, who, in 1620, published the first book on the subject. This gave rise to a wider interest in the education of the deaf in Europe. In England John Bulwer (fl. 1654) wrote on teaching the deaf to speak and read the lips, and in France, similar work was later carried on by Charles-Michel, abbé de l'Epée (1712–89), who made a most profound contribution in developing the natural sign language of the deaf into a systematic and conventional language to be used as a medium of instruction. His work was developed by the abbé Sicard (1742–1822) and gave rise to the manual system or silent method of teaching the deaf. In Germany Samuel Heinicke (1727–90) educated deaf children orally, and later Friedrich Moritz Hill (1805–74), perhaps one of the greatest educators of the deaf, developed this method in relation to the concept that education must relate to the "here and now" of the child—known as the natural method. Thus arose the oral method of instruction that continued to influence the teaching of the deaf and in time became an accepted practice throughout the world.

No real attempt was made to educate or train the blind until the close of the 18th century. Valentin Haüy (1745–1822), the man to whom the title "father and apostle of the blind" afterward was given, opened the Institution Nationale des Jeunes Aveugles in Paris in 1784, with 12 blind children as his first pupils. News of Haüy's success in teaching these children to read soon spread to other countries. Schools for the blind were opened in Liverpool (1791), London (1799), Vienna (1805), Berlin (1806), Amsterdam and Stockholm (1808), Zürich (1809), and Boston and New York (1832).

Scientific attempts to educate mentally retarded children began with the efforts of Jean-Marc-Gaspard Itard (1775–1838), a French physician and otologist connected with an institution for the deaf. In his classic book, *The Wild Boy of Aveyron* (1801), he related his five-year effort to train and educate an 11-year-old boy found running naked and wild in the woods of Aveyron. Later Edouard Séguin (1812–80), a student of Itard's who emigrated to the United States in 1848, devised an educational method using physical and sensory activities to develop mental processes. Séguin's published works influenced Maria Montessori (1870–1952), an Italian pediatrician who became an educator and the innovator of a unique method of training young mentally retarded and culturally deprived children in Rome in the 1890s and early 1900s. Self-education through the use of specially designed "didactic materials" for sensorimotor training was the keynote of the system.

In advanced countries today, special education of the handicapped has become relatively universal; and, in the process, two concepts of individual differences have evolved: (1) the concept of "interindividual differences," which compares one child with another, and (2) the concept of "intraindividual differences," which is concerned with how the child's abilities in one area compare with his abilities in other areas. The grouping of children in special classes rests on the concept of *inter*individual differences, but the instructional procedures for each child are determined by *intra*individual differences—that is, by his abilities and disabilities.

Concepts of individual differences

Modern patterns of organizing educational programs. *Diagnostic patterns.* Because children with a particular kind of handicap do not form a homogeneous group, diagnosis must go beyond merely classifying the children according to their major deviation. A child suffering from cerebral palsy, for example, has a motor handicap, but in intelligence he may be superior or average or mentally retarded. Hence children with certain labels of cerebral palsy or deafness or blindness must be carefully assessed before placement in a particular group.

For the gifted and the mentally retarded, the primary criterion of identification is an individually administered intelligence test. If the child is high (for the gifted, generally above 125 IQ) or low (for the mentally retarded, below 75), he is considered for the special program. The determination is made by psychologists who in most cases must certify the child for eligibility for special programs. In making their assessment, psychologists also take into account other criteria such as school achieve-

ment, personality, and an evaluation of the adjustment of the child in the regular grades.

For children with sensory handicaps the evaluation of the problem is made by medical personnel. Orthopedically and neurologically impaired children are assigned to special programs upon the results of orthopedic or neurological examination by physicians, with supplementary assessment by psychological and educational diagnosticians. Children with learning disabilities are assessed primarily by psycho-educational diagnosticians, who through educational and psychological diagnostic tests determine the child's potential for learning and his level of achievement. Ancillary diagnoses by medical, psychological, and other personnel are considered in the determination of eligibility for special programs and in the kind of programs required. Socially and emotionally maladjusted children are generally referred by parents or by regular teachers when they find it difficult to manage the child at home or in school. Psychiatrists and clinical psychologists, together with social workers and educational personnel, determine eligibility for assignment to special programs.

Patterns of instructional adaptation. The goals of special education are similar to the goals for ordinary children, only the techniques for attaining them are different. An effort is made, for example, to teach all exceptional children (except those unable to profit at all from school experience) to read. Most gifted children have little difficulty in learning to read, many of them having learned to read at home before entering the first grade. The mentally retarded, on the other hand, obviously require prolonged periods of more intensive and more individualized instruction, which includes more techniques to maintain interest, more active participation, and much repetition of similar material in varied form. Children with severe sensory handicaps (the deaf and the blind) must learn to read through other sense modalities. The deaf learn to read through a visual method not involving the sense of hearing. The blind learn to read Braille through the tactile and kinesthetic sense.

Techniques for teaching the deaf and the blind

The most specialized teaching techniques are required for deaf children who, because of their hearing loss and their inability to develop auditory language, are inhibited in the ability to speak and to learn to read. The inability to read well inhibits other educational endeavours. Special techniques involve learning to speak by kinesthetic and visual methods, learning to "hear" others through lip reading, and learning to read by special visual techniques to compensate for their deficiency in oral language. General information is sometimes obtained by the deaf through manual language and finger spelling. For several centuries, the major controversy in the education of the deaf has been whether to use manual or oral methods to educate the deaf. There are thus some schools that teach primarily by the oral method (speech and speech reading) without the use of manual language and some that educate by a combination of oral and manual methods.

There is no major controversy about the education of the hard-of-hearing. These children are generally educated in regular classes with ordinary children but receive special instruction from visiting teachers in auditory training and the use of hearing aids, speech correction, and speech reading.

The blind are educated through the sense of hearing and the use of the tactile sense. Books are prepared in Braille, a code of embossed dots over which the blind person runs his fingers and translates the sequences of dots into meaning—much in the same manner as the average person translates sequences of seen letters into meaning. Many textbooks, much literary material, and quantities of information are also transcribed onto tapes and records for the blind. Mobility training has been developed and is being widely used to help the blind move about comfortably in their environment by using senses other than vision.

Partially seeing children, on the other hand, are educated primarily through the visual, auditory, and haptic senses, using magnification of their limited sight, training of visual perception, and utilization of large-type print.

For children with motor handicaps, required academic adjustments are minimal. Unless the child has other problems such as learning disabilities, mental retardation, or speech problems (which are often found among the cerebral palsied), crippled children learn like other children and can follow the same classroom materials. Special techniques are necessary, however, to help such children to adapt to their environment and to adapt the environment to their disability. Wheel chairs, modified desks, and other apparatus aid the child in mobility and the manipulation of classroom materials. One of the most important aspects of the education of the orthopedically handicapped is attitudinal—that is, preparing the child for adapting to the world outside the classroom and facing his handicap in such a way as to permit him to lead a relatively normal life.

Children with learning disabilities and those with speech defects require highly specialized techniques, usually on an individual basis.

For children with social and emotional problems, special therapeutic and clinical services may be provided. Psychotherapy and behaviour therapy by clinical psychologists, social workers, and psychiatrists are generally a part of the educational program. Academic teachers in these classes stress personality development, social adjustment, habits of interpersonal relations. With this group of children, these factors are prerequisite to academic achievement. Academic work is, however, sometimes therapeutic in itself and is promoted as much as possible.

Grouping patterns. The various types of organization by which exceptional children are grouped for educational purposes are determined by the extent of the exceptionality as well as by the type of exceptionality.

Special classes

The most common organization for the majority of handicapped children has been the self-contained "special class" in either state or privately operated school systems. Special classes for the gifted, the mentally retarded, the deaf and the hard of hearing, the blind and the partially sighted, and for other handicapped children are found in school systems throughout the world. This type of organization allows the child to live at home and often to attend a neighbourhood school that has organized a special class.

Within the regular school, children with speech problems, learning disabilities, and special handicaps are often given tutoring by specially trained personnel. Such children attend regular classes but receive special help for short periods each day or several times a week to help correct their special disability. Children who are homebound and cannot attend school are taught at home by special teachers.

The "resource room" is a specially equipped room with a specially trained teacher who serves children sent to her for part of the day. Children assigned to a resource room attend regular classes with ordinary children but go to the resource room for special help such as tutoring or remedial work.

Less common than these programs are special "day schools" organized for some particular kind of handicapped child. These differ from special classes or rooms in that the whole school serves crippled children or mentally retarded or emotionally disturbed, the deaf, or the blind. Children attending these schools also live at home but very often are transported to and from school by bus.

"Residential schools" enroll handicapped children for 24 hours a day and are usually attended by children who cannot obtain services in their community for any one of various reasons. Sometimes the services are simply not available; sometimes the severity of the handicap makes it difficult to adjust in a day school or special class; sometimes the children are dependent, or the parents find that keeping the child at home creates problems for the rest of the family, or the environment does not provide sufficient help for the child himself. State and private residential schools for the deaf and for the blind, and for the severely mentally retarded are the most common. In some

European countries such as France and the Soviet Union, areas outside the large cities are preferred as sites for boarding schools or residential schools for the educable mentally retarded.

Children confined to hospitals for varying lengths of time are frequently offered educational services individually or in small groups while they are in the hospital. The goal is to assist them in keeping up with their education so that they will not be retarded educationally when they return to their home and school.

PROBLEMS OF THE PERSONAL ADJUSTMENT OF EXCEPTIONAL CHILDREN

Research on the personal adjustment of gifted and handicapped children (other than the socially or emotionally disturbed) has not shown the probability of maladjustment to be as great as "common sense" would indicate. It was once assumed, for example, that a gifted child would show exceptional talent in some area but would be abnormal in social relations. This myth has long been discarded; studies over a period of years have shown that the extremely gifted (children with IQ's of 140 and above) were better adjusted than the average, personally and socially, in school and in later life.

Personal adjustment of the handicapped is contingent upon many factors, among which is excessive frustration of the child in his growing years when there is a discrepancy between his ability to perform and the requirements of the environment. During the preschool years the handicapped child is a major concern of his parents. Their anxiety leads them to seek diagnosis and cures from many sources. How well the parents learn to accept the handicap and to assist the child in developing without undue pressure, rejection, or overprotection determines to a large extent the child's personal adjustment.

In his early school years the handicapped child meets obstacles to development when in competition with the average child. Studies have shown that when mentally retarded or crippled children are assigned to regular grades without special provisions being made for them, they tend to be isolated and rejected by their normal classmates. Special classes and other special provisions have been organized to avoid these problems and to decrease the discrepancies between the requirements of the environment and the ability of the child to perform.

The adolescent level is difficult for handicapped children, as it is for all children. At this age level more personal and social problems appear. Special provisions are less common than in schools for younger children. Because many handicapped children (the mentally retarded, the cerebral palsied, the deaf, and the blind) are unable to cope with the more abstract and academic requirements of the school at this age level, special provisions must be made for them.

Prevocational training and instruction in the practical tasks of life in a nonacademic way assist exceptional children in adjusting to their school, their peers, and their families. Vocational rehabilitation programs in European countries and the United States have been instrumental in educating handicapped adolescents and youth in vocations in which they have a possibility of becoming self-sufficient and independent at the adult level.

One of the major goals of special education (beyond developing skills and imparting information) is the ultimate adjustment of the handicapped to the society of nonhandicapped individuals. This goal has created a problem in the organization of special classes and schools. One criticism of special education programs is that they tend to segregate children with their own kind. Currently, efforts are being made to organize programs that integrate the deviant child with average children. For the gifted, special programs of enrichment and acceleration are increasingly preferred to special classes. Resource rooms for those impaired in sight or hearing make it possible to retain the child in the regular grades for part of the day. The older educable mentally retarded can be assigned to regular workshops, physical education classes, home economics classes, and other nonacademic classes for part of the day. Similar adjustments in programs are used for other handicapped children. The goal is to adjust these children to the larger society of nonhandicapped before they leave school.

BIBLIOGRAPHY. One of the first authoritative textbooks to be published on handicapped children and their education, including research findings to 1924, was J.E.W. WALLIN, *The Education of Handicapped Children* (1924). More recent textbooks include; W.M. CRUICKSHANK and G.O. JOHNSON (eds.), *Education of Exceptional Children and Youth*, 2nd ed. (1967), prepared by specialists in each area of exceptionality, with emphasis on educational progress; L.M. DUNN (ed.), *Exceptional Children in the Schools* (1963), includes research findings on the characteristics and general educational adaptations for various types of exceptionality; and S.A. KIRK, *Educating Exceptional Children* (1962), utilizes the concept of intra-individual differences as the central theme and discusses special educational techniques. Another work that discusses special methods of instruction used by teachers is N.G. HARING and R.L. SCHIEFELBUSCH (eds.), *Methods in Special Education* (1967).

(S.A.K.)

Edward I of England

King of England from 1272 to 1307, Edward I presided over a period of rising national consciousness within the country. During his reign, the English Parliament began to develop its distinctive character, and the monarchy extended its effective control over Wales. His attempt to extend English suzerainty over Scotland initiated two and a half centuries of sanguinary warfare between the two countries. His reign is particularly noted for administrative efficiency and legal reform. He introduced a series of statutes that did much to strengthen the crown in the feudal hierarchy. His definition and emendation of English common law has earned him the name of the "English Justinian."

By courtesy of the trustees of the British Museum

Edward I, watercolour, 15th century. In the British Museum (Cotton MS. Julius E. IV).

Early life. Edward was born at Westminster on June 17, 1239, the eldest son of King Henry III and Eleanor of Provence. In 1254 he was given the duchy of Gascony, the French Oléron, the Channel Islands, Ireland, Henry's lands in Wales, and the earldom of Chester, as well as several castles. Henry negotiated Edward's marriage with Eleanor, half sister of Alfonso X of León and Castile. He married Eleanor at Las Huelgas in Spain (October 1254) and then travelled to Bordeaux to organize his scattered appanage. He now had his own household and officials,

chancery and seal, with an exchequer (treasury) at Bristol Castle; though nominally governing all his lands, in Gascony and Ireland he merely enjoyed the revenues. He returned to England in November 1255 and attacked Llewelyn ap Gruffydd, prince of Gwynedd, to whom his Welsh subjects had appealed for support when Edward attempted to introduce English administrative units in his Welsh lands. Edward, receiving no help from either Henry or the marcher lords, was defeated ignominiously. His arrogant lawlessness and his close association with his greedy Poitevin uncles, who had accompanied his mother from France, increased Edward's unpopularity among the English. But after the Poitevins were expelled, Edward fell under the influence of Simon de Montfort, his uncle by marriage, with whom he made a formal pact. Montfort was the leader of a baronial clique that was attempting to curb the misgovernment of Henry.

Role in barons' revolt

Edward reluctantly accepted the Provisions of Oxford (1258), which gave effective government to the barons at the expense of the king. On the other hand, he intervened dramatically to support the radical Provisions of Westminster (October 1259), which ordered the barons to accept reforms demanded by their tenants. In the dangerous crisis early in 1260 he supported Montfort and the extremists, though finally he deserted Montfort and was forgiven by Henry (May 1260). He was sent to Gascony in October 1260 but returned early in 1263. Civil war had now broken out between Henry and the barons, who were supported by London. Edward's violent behaviour and his quarrel with the Londoners harmed Henry's cause. At the Battle of Lewes (May 14, 1264) his vengeful pursuit of the Londoners early in the battle contributed to Henry's defeat. Edward surrendered and became a hostage in Montfort's hands. He escaped at Hereford in May 1265 and took charge of the royalist forces, penned Montfort behind the River Severn, and, by lightning strategy, destroyed a large relieving army at Kenilworth (August 1). On August 4 he trapped and slew Montfort at Evesham and rescued Henry. Shattered and enfeebled, Henry allowed Edward effective control of government, and his extreme policy of vengeance, especially against the Londoners, revived and prolonged rebel resistance. Finally, the papal legate Ottobuono, Edward's uncle Richard, earl of Cornwall, and other moderates persuaded Henry to the milder policy of the Dictum of Kenilworth (October 31, 1266), and after some delay the rebels surrendered. Edward took the cross (1268), intending to join the French king Louis IX on a crusade to the Holy Land, but was delayed by lack of money until August 1270. Louis died before Edward's arrival; and Edward, after wintering in Sicily, went to Acre, where he stayed from May 1271 to September 1272, winning fame by his energy and courage and narrowly escaping death by assassination but achieving no useful results. On his way home he learned in Sicily of Henry III's death on November 16, 1272.

Crusading activity

Accession and character. Edward had nominated Walter Giffard, archbishop of York, Philip Basset, Roger Mortimer, and his trusted clerk Robert Burnell to safeguard his interests during his absence. After Henry's funeral, the English barons all swore fealty to Edward (November 20, 1272). His succession by hereditary right and the will of his magnates was proclaimed, and England welcomed the new reign peacefully, Burnell taking charge of the administration with his colleagues' support. The quiet succession demonstrated England's unity only five years after a bitter civil war. Edward could journey homeward slowly, halting in Paris to do homage to his cousin Philip III for his French lands (July 26, 1273), staying several months in Gascony and reaching Dover on August 2, 1274, for his coronation at Westminster on August 19. Now 35 years old, Edward had redeemed a bad start. He had been arrogant, lawless, violent, treacherous, revengeful, and cruel; his Angevin rages matched those of Henry II. Loving his own way and intolerant of opposition, he had still proved susceptible to influence by strong-minded associates. He had shown intense family affection, loyalty to friends, courage, brilliant military capacity, and a gift for leadership; handsome, tall, powerful, and tough, he had the qualities men admired. He loved efficient, strong government, enjoyed power, and had learned to admire justice, though in his own affairs it was often the letter, not the spirit of the law that he observed. Having mastered his anger, he had shown himself capable of patient negotiation, generosity, and even idealism; and he preferred the society and advice of strong counsellors with good minds. As long as Burnell and Queen Eleanor lived, the better side of Edward triumphed, and the years until about 1294 were years of great achievement. Thereafter, his character deteriorated for lack of domestic comfort and independent advice. He allowed his autocratic temper full rein and devoted his failing energies to prosecution of the wars in France and against Scotland.

Parliament and statutes. Shrewdly realistic, Edward understood the value of the "parliaments," which since 1254 had distinguished English government and which Montfort had deliberately employed to publicize government policy and to enlist widespread, active support by summoning representatives of shires and boroughs to the council to decide important matters. Edward developed this practice swiftly, not to share royal power with his subjects but to strengthen royal authority with the support of rising national consciousness. From 1275 to 1307 he summoned knights and burgesses to his parliaments in varying manners. The Parliament of 1295, which included representatives of shires, boroughs, and the lesser clergy, is usually styled the Model Parliament, but the pattern varied from assembly to assembly, as Edward decided. By 1307, Parliament, thus broadly constituted, had become the distinctive feature of English politics, though its powers were still undefined and its organization embryonic.

The Model Parliament

Edward used these parliaments and other councils to enact measures of consolidation and reform in legal, procedural, and administrative matters of many kinds. The great statutes promulgated between 1275 and 1290 are the glory of his reign. Conservative and definitory rather than original, they owed much to Burnell, Edward's chancellor. With the vast developments and reorganization of the administrative machine that Burnell coordinated, they created a new era in English government. The quo warranto inquiry, begun in 1275, the statutes of Gloucester (1278) and of Quo Warranto (1290) sought with much success to bring existing franchises under control and to prevent the unauthorized assumption of new ones. Tenants were required to show "by what warrant" or right they held their franchises. Edward strove, unsuccessfully, to restore the feudal army and strengthen local government institutions by compelling minor landowners to assume the duties of knighthood. His land legislation, especially the clause *de donis conditionalibus* in the miscellaneous Second Statute of Westminster (1285) and the statute Quia Emptores (Third Statute of Westminster, 1290), eventually helped to undermine feudalism, quite contrary to his purpose. By the Statute of Mortmain (1279) the crown gained control of the acquisition of land by ecclesiastical bodies. The Statute of Winchester (1285) codified and strengthened the police system for preserving public order. The Statute of Acton Burnell (1283) and the Statute of Merchants (1285) showed practical concern for trade and merchants. These are but the most famous of many statutes aimed at efficiency and sound administration.

Decline of feudalism

Wars. Meanwhile, Edward destroyed the autonomous principality of Wales, which, under Llewelyn ap Gruffydd, had expanded to include all Welsh lordships and much territory recovered from the marcher lords. Domestic difficulties had compelled Henry III to recognize Llewelyn's gains by the Treaty of Shrewsbury (1267), but Edward was determined to reduce Llewelyn and used Llewelyn's persistent evasion of his duty to perform homage as a pretext for attack. He invaded Wales by three coordinated advances with naval support (1277), blockaded Llewelyn in Snowdonia, starved him into submission, and stripped him of all his conquests since 1247. He then erected a tremendous ring of powerful castles encircling Gwynedd and reorganized the conquered dis-

tricts as shires and hundreds. When English rule provoked rebellion, he methodically reconquered the principality, killing both Llewelyn (1282) and his brother David (1283). By the Statute of Wales (1284) he completed the reorganization of the principality on English lines, leaving the Welsh marchers unaffected. A further Welsh rising in 1294–95 was ruthlessly crushed, and Wales remained supine for more than 100 years.

After 1294, matters deteriorated. Queen Eleanor had died in 1290, Burnell in 1292, and Edward never thereafter found such good advisers. The conquest and fortification of Wales had badly strained his finances; now endless wars with Scotland and France bankrupted him. He quarrelled bitterly with both clergy and barons, behaving as a rash and obstinate autocrat who refused to recognize his limitations. Philip III and Philip IV of France had both cheated him of the contingent benefits promised by the Treaty of Paris (1259). By constant intervention on pretext of suzerainty they had nibbled at his Gascon borders and undermined the authority of his administration there. After doing homage to Philip IV in 1286, Edward visited Gascony to reorganize the administration and restore authority. On returning to England in 1289 he had to dismiss many judges and officials for corruption and oppression during his absence. In 1290, having systematically stripped the Jews of their remaining wealth, he expelled them from England. French intervention in Gascony was now intensified; affrays between English and French sailors inflamed feelings; and in 1293 Philip IV tricked Edward's brother Edmund, earl of Lancaster, who was conducting negotiations, into ordering a supposedly formal and temporary surrender of the duchy, which Philip then refused to restore. The Welsh rising and Scottish troubles prevented Edward from taking action, and when at last, in 1297, he sailed to attack France from Flanders, his barons refused to invade Gascony, and Wallace's rising forced him to return. He made peace with Philip (1299), and by Boniface VIII's persuasion, married Philip's sister Margaret, and eventually recovered an attenuated Gascon duchy.

Renewal of quarrels with Scotland

For more than 100 years relations between England and Scotland had been amicable, and the border had been remarkably peaceful. Edward inaugurated 250 years of bitter hatred, savage warfare, and bloody border forays. The deaths of Alexander III of Scotland (1286) and his granddaughter Margaret, the Maid of Norway (1290), whom Edward planned to marry to his heir, Edward of Caernarvon (afterward Edward II), ended the line of succession. Many dubious claimants arose, and the Scottish magnates requested Edward's arbitration. Edward compelled the nobles and the claimants to recognize his suzerainty, and only then adjudged John de Balliol king (1292). Balliol did homage and was crowned, but Edward's insistence on effective jurisdiction, as suzerain, in Scottish cases eventually provoked the Scottish nobles to force Balliol to repudiate Edward's claims and to ally with France (1295). Edward invaded and conquered Scotland (1296), removing to Westminster the coronation stone of Scone. William Wallace led a revolt in 1297, and Edward, though brilliantly victorious at Falkirk (July 22, 1298), could not subdue the rebellion despite prolonged campaigning (1298–1303).

Last years. The strain of these years provoked heavy collisions between Edward and his magnates. He had quarrelled violently with his archbishops of Canterbury, John Peckham (1279–92) and Robert Winchelsey (1293–1313), over ecclesiastical liberties and jurisdiction. In 1297 Winchelsey, obeying Pope Boniface VIII's bull *Clericis Laicos* (1296), rejected Edward's demands for taxes from the clergy, whereupon Edward outlawed the clergy. His barons now defied his orders to invade Gascony and, when Edward went to Flanders, compelled the regents to confirm the charters of liberties, with important additions forbidding arbitrary taxation (1297), thereby forcing Edward to abandon the campaign and eventually to make peace with France. Although Pope Clement V, more pliant than Boniface, allowed Edward to exile Winchelsey and intimidate the clergy (1306), the barons had exacted further concessions (1301) before reconciliation. Edward renewed the conquest of Scotland in 1303, captured Stirling in 1304, and executed Wallace as a traitor in 1305; but when Scotland seemed finally subjected, Robert I the Bruce revived rebellion and was crowned in 1306. On his way to reconquer Scotland, Edward died at Burgh by Sands, near Carlisle, on July 7, 1307.

BIBLIOGRAPHY. E.M. BARRON, *The Scottish War of Independence*, 2nd ed. (1934), although not wholly objective, is a clear and readable account of the wars. J.E. MORRIS, *The Welsh Wars of Edward I* (1901, reprinted 1968), makes clear the impact of these wars on Edward's military organization and technique as well as on constitutional development in England. D. PASQUET, *Essai sur les origines de la Chambre des Communes* (1914; Eng. trans., *An Essay on the Origins of the House of Commons*, 1925), takes the view that the English Parliament owes more to the initiative and needs of the King than to any sense of community among his subjects. There is a vast body of more recent literature on parliamentary origins. This is summarized and selections from it are presented in *Historical Studies of the English Parliament*, ed. by E.B. FRYDE and EDWARD MILLER, 2 vol. (1970). T.F.T. PLUCKNETT, *Legislation of Edward I* (1949), is an analysis of Edward I's statutes, showing that, far from being antifeudal, Edward's measures were intended to support his own position within the feudal hierarchy. F.M. POWICKE's two works on the 13th century, *King Henry III and the Lord Edward*, 2 vol. (1947), and *The Thirteenth Century, 1216–1307* (1953), present the reign in terms of social forces at work in politics, particularly those springing from the baronage. G. TEMPLEMAN, "Edward I and the Historians," in the *Cambridge Historical Journal*, 10:16–35 (1950), summarizes the changing views of Edward I from Polydore Vergil to Powicke and Plucknett. T.F. TOUT, *Chapters in the Administrative History of Mediaeval England*, vol. 2 (1920, reprinted 1967), demonstrates Edward's aims with regard to administrative efficiency and royal control. Tout's *Edward the First* (1893) is still the best biography of the King; L.F. SALZMAN, *Edward I* (1968), adds nothing substantially new to our understanding of the King or the man.

(R.F.Tr.)

Edward III of England

King of England from 1327 to 1377, Edward III led England into the Hundred Years' War with France. His victories of Crécy, Sluis, and Calais were much celebrated by the 14th-century historian Jean Froissart, but their long-term significance is debatable. The eldest son of Edward II and Isabella of France, Edward III was born at Windsor on November 13, 1312. He was summoned to Parliament as earl of Chester (1320) and was made duke of Aquitaine (1325), but, contrary to tradition, he never received the title of prince of Wales.

Edward III grew up amid struggles between his father and a number of barons who were attempting to limit the king's power and to strengthen their own role in governing England. His mother, repelled by her husband's treatment of the nobles and disaffected by the confiscation of her English estates by his supporters, played an important role in this conflict. In 1325 she left England to return to France to intervene in the dispute between her brother, Charles IV of France, and her husband over the latter's French possessions, Guyenne, Gascony, and Ponthieu. She was successful; the land was secured for England on condition that the English king pay homage to Charles. This was performed on the King's behalf by his young son.

The heir apparent was secure at his mother's side. With Roger Mortimer, an influential baron who had escaped to France in 1323 and had become her lover, Isabella now began preparations to invade England to depose her husband. To raise funds for this enterprise, Edward III was betrothed to Philippa, daughter of William, count of Hainaut and Holland.

Within five months of their invasion of England, the Queen and the nobles, who had much popular support, overpowered the king's forces. Edward II, charged with incompetence and breaking his coronation oath, was forced to resign, and on January 29, 1327, Edward III, aged 15, was crowned king of England.

Regency

During the next four years Isabella and Mortimer governed in his name, though nominally his guardian was Henry, earl of Lancaster. In the summer of 1327 he took

Edward III, watercolour, 15th century. In the British Museum (Cotton MS. Julius E. IV).
By courtesy of the trustees of the British Museum

part in an abortive campaign against the Scots, which resulted in the Treaty of Northampton (1328), making Scotland an independent realm. Edward was deeply troubled by the settlement and signed it only after much persuasion by Isabella and Mortimer. He married Philippa at York on January 24, 1328. Soon afterward, Edward made a successful effort to throw off his degrading dependence on his mother and Mortimer. While a council was being held at Nottingham, he entered the castle by night, through a subterranean passage, took Mortimer prisoner, and had him executed (November 1330). Edward had discreetly ignored his mother's liaison with Mortimer and treated her with every respect, but her political influence was at an end.

Edward III now began to rule as well as to reign. Young, ardent, and active, he sought to remake England into the powerful nation it had been under Edward I. He still resented the concession of independence made to Scotland by the Treaty of Northampton; and the death of Robert I, the Bruce, king of Scotland, in 1329 gave him a chance of retrieving his position. The new king of Scots, his brother-in-law, David II, was a mere boy, and Edward took advantage of his weakness to aid the Scottish barons who had been exiled by Bruce to place their leader, Edward Balliol, on the Scottish throne. David II fled to France, but Balliol was despised as a puppet of the English king, and David returned in 1341.

Hundred Years' War. During the 1330s England gradually drifted into a state of hostility with France, for which the most obvious reason was the dispute over English rule in Gascony. Contributory causes were France's new King Philip VI's support of the Scots, Edward's alliance with the Flemish cities—then on bad terms with their French overlord—and the revival, in 1337, of Edward's claim, first made in 1328, to the French crown. Edward twice attempted to invade France from the north (1339, 1340) but the only result of his campaigns was to reduce him to bankruptcy. In January 1340 he assumed the title of king of France. At first he may have done this to gratify the Flemings, whose scruples in fighting the French king disappeared when they persuaded themselves that Edward was the rightful king of France. But his pretensions to the French crown gradually became more important, and the persistence with which he and his successors urged them made stable peace impossible for more than a century. This was the struggle famous in history as the Hundred Years' War. Until 1801 every English king also called himself king of France.

Battle of Sluis

Edward was present in person at the great naval battle off the Flemish city of Sluis in June 1340, in which he all but destroyed the French navy. Despite this victory his resources were exhausted by his land campaign, and he was forced to make a truce (which was broken two years later) and return to England. During the years after 1342 he spent much time and money in rebuilding Windsor Castle and instituting the Order of the Garter, which became Britain's highest order of knighthood. A new phase of the French war began when Edward landed in Normandy in July 1346, accompanied by his eldest son, Prince Edward, later known as the Black Prince (born 1330). At first the king showed some lack of strategic purpose, engaging in little more than a large-scale plundering raid to the gates of Paris. The campaign was made memorable by his decisive victory over the French at Crécy in Ponthieu (August 26), where he scattered the army with which Philip VI sought to cut off his retreat to the northeast. Edward laid siege to the French port of Calais in September 1346 and received its surrender in August 1347. Other victories in Gascony and Brittany, and the defeat and capture of David II at Neville's Cross near Durham (October 1346), gave further proof of Edward's power, but Calais was to be his only lasting conquest. He ejected most of its French inhabitants, colonizing the town with Englishmen and establishing there a base from which to conduct further invasions of France. Nevertheless, in the midst of his successes, want of money forced him to make a new truce in September 1347.

Years in England

Edward returned to England in October 1347. He celebrated his triumph by a series of splendid tournaments. In 1348 he rejected an offer to become Holy Roman emperor. In the same year the bubonic plague known as the Black Death first appeared in England and raged until the end of 1349. Its horrors hardly checked the magnificent revels of Edward's court, and neither the plague nor the truce stayed the slow course of the French war, though the fighting was indecisive and on a small scale. Edward's martial exploits during the next years were those of a gallant knight rather than of a responsible general. Although the English House of Commons was now weary of the war, efforts to make peace came to nothing and large-scale operations began again in 1355 when Edward led an unsuccessful raid out of Calais. He harried the Lothians, part of southeast Scotland, in the expedition famous as the Burned Candlemas (January and February 1356), and in the same year he received a formal surrender of the Kingdom of Scotland from Balliol. His exploits were, however, eclipsed by those of his son, Edward, whose victory at Poitiers (September 19, 1356), resulting in the capture of the French king John II (who had succeeded Philip VI in 1350), forced the French to accept a new truce. Edward entertained his captive magnificently, but forced him by the Treaty of London (1359) to surrender so much territory that the agreement was repudiated in France. In an effort to compel acceptance, Edward landed at Calais (October 28) and besieged Reims, where he planned to be crowned king of France. The strenuous resistance of the citizens frustrated this scheme and Edward marched into Burgundy, eventually returning toward Paris. After this unsuccessful campaign he was glad to conclude preliminaries of peace at Brittany (May 8, 1360). This treaty, less onerous to France than that of London, took its final form in the Treaty of Calais, ratified by both kings (October 1360). By it, Edward renounced his claim to the French crown in return for the whole of Aquitaine, a rich area in southwest France.

Capture of the French king

The years of decline: 1360–77. The Treaty of Calais did not bring rest or prosperity to either England or France. Fresh visitations of the Black Death in England in 1361 and 1369 intensified social and economic disturbances, and desperate but not very successful efforts were made to enforce the Statute of Labourers (1351), which was intended to maintain prices and wages as they had been before the pestilence. Other famous laws enacted during the 1350s had been the Statutes of Provisors (1351) and Praemunire (1353), which reflected popular hostility against foreign clergy. These measures were fre-

quently re-enacted, and Edward formally repudiated (1366) the feudal supremacy over England still claimed by the papacy.

Resumption of fighting

When the French king Charles V, son of John II, repudiated the treaty of Calais, Edward resumed the title of king of France, but he showed little of his former vigour in meeting this new trouble, leaving most of the fighting and the administration of his foreign territories to his sons Edward and John of Gaunt, duke of Lancaster. While they were struggling with little success against the rising tide of French national feeling, Edward's want of money made him a willing participant in the attack on the wealth and privileges of the church. Meanwhile, Aquitaine was gradually lost, Prince Edward returned to England in broken health (1371), and John of Gaunt's march through France from Calais to Bordeaux (1373) achieved nothing. Edward's final attempt to lead an army abroad himself (1372) was frustrated when contrary winds prevented his landing his troops in France. In 1375 he was glad to make a truce, which lasted until his death. By it, the only important possessions remaining in English hands were Calais, Bordeaux, Bayonne, and Brest.

Edward was now sinking into his dotage. After the death of Queen Philippa in 1369 he fell entirely under the influence of his greedy mistress, Alice Perrers, while Prince Edward and John of Gaunt became the leaders of sharply divided parties in the royal court and council. John of Gaunt returned to England in April 1374 and with the help of Alice Perrers obtained the chief influence with his father, but his administration was neither honourable nor successful. At the famous so-called Good Parliament of 1376 popular indignation against John of Gaunt's ruling party came at last to a head. Alice Perrers was removed and some of Gaunt's followers were impeached. Before the Parliament had concluded its business, however, the death of the Prince Edward (June 8, 1376) robbed the Commons of its strongest support. John of Gaunt regained power, and the acts of the Good Parliament had been reversed when Edward III died on June 21, 1377, at Sheen (now Richmond).

Edward's character

Edward III possessed extraordinary vigour and energy of temperament; he was an admirable tactician and a consummate knight. His court was the most brilliant in contemporary Europe, and he was himself well fitted to be the head of the gallant knights who obtained fame in the French wars. Though his main ambition was military glory, he was not a bad ruler of England, being liberal, kindly, good-tempered, and easy of access. His need to obtain supplies for carrying on the French wars made him favourable to his subjects' petitions and contributed to the growing strength of Parliament. His weak points were his wanton breaches of good faith, his extravagance, his frivolity, and his self-indulgence. His ambition ultimately transcended his resources, and before he died even his subjects had sensed his failure.

BIBLIOGRAPHY. W. STUBBS, *The Constitutional History of England*, vol. 2 (1866 and many later editions), dated but still very valuable; M. MCKISACK, *The Fourteenth Century, 1307–1399* (1959), contains the best modern account; G. UNWIN (ed.), *Finance and Trade Under Edward III* (1918), still useful; E. PERROY, *La Guerre de cent ans* (1945; Eng. trans., *The Hundred Years War*, 1951), good for the French point of view.

(T.F.T./J.R.L.H.)

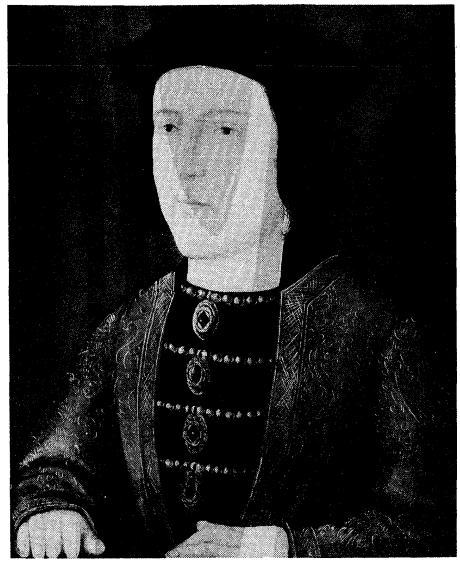
Edward IV of England

Edward IV was king of England from 1461 until October 1470, and again from April 1471 until his death in 1483, and was one of the leading participants in the Yorkist–Lancastrian civil conflict known as the Wars of the Roses. Edward was the eldest surviving son of Richard, duke of York, by Cicely, daughter of Ralph Neville, earl of Westmorland; he was born at Rouen, France, on April 28, 1442. His father was descended from two sons of the 14th-century king Edward III and, in the 1450s, led a revolt against Henry VI; in 1460, Richard's supporters declared him Henry's successor. When his father was killed in December of that year, Edward gathered an army in Wales, and defeated Henry's supporters (called Lancastrians because of Henry's descent from John of Gaunt, duke of Lancaster). Edward was crowned as King Edward IV in London on June 28, 1461.

Edward's claim to the throne

Edward's struggle with Warwick. Edward at this time showed little promise, apparently caring only for fighting, drinking, women, and pageantry. He owed his throne largely to his cousin Richard Neville, earl of Warwick, who was in the first years of Edward's reign the most powerful man in England. Warwick crushed Lancastrian resistance in the far north of England between 1462 and 1464 and conducted England's diplomacy. Edward, however, was winning many friends (especially in London) by his comeliness and charm and was determined to assert his independence. On May 1, 1464, he secretly married a young widow, Elizabeth Woodville, of no great rank, offending Warwick and other Yorkist nobles who were planning to marry him to a French princess. By showering favours on Elizabeth's two sons by her first husband and on her five brothers and her seven sisters, Edward began to build up a group of magnates who would be a counterpoise to the Nevilles. Gradually Warwick lost all influence at court, and when he was negotiating an alliance with France, Edward humiliated him by revealing that he had already concluded an alliance (1467) with France's enemy Burgundy. Edward's sister Margaret was married July 1468 with great pomp to Duke Charles the Bold of Burgundy, and the brothers-in-law planned a joint invasion of France.

By courtesy of the National Portrait Gallery, London



Edward IV, portrait by an unknown artist. In the National Portrait Gallery, London.

Warwick, in a countermove encouraged by Louis XI of France, seized Edward and made him a prisoner in July 1469. But Edward had by now too many supporters (especially in London) for him to be kept under tutelage for long. He regained his freedom in October; Warwick fled to France, allied himself with the Lancastrians and with Louis, and invaded England in September 1470.

Surprised, Edward fled with a few faithful supporters to the Netherlands in October. Aided by Charles of Burgundy, he and his brother, Richard, duke of Gloucester, returned to England in March 1471. Taking London, he defeated and killed Warwick at Barnet on April 14. On the same day, Queen Margaret (Henry VI's wife) belatedly landed in Dorset from France with her only son, Edward, prince of Wales. Her advisers hoped to gain Lancastrian support in Wales, and it became a race for time between Edward IV's forces and hers as to whether she could get there before he overtook her. At Tewkesbury, after some remarkable forced marches (one of over 40 miles at a stretch), he caught up with her army on May 4. There he won another crushing victory. Nearly all the remaining Lancastrian leaders were killed on the field or executed afterward and, after murdering Henry VI

The Battle of Tewkesbury

(May 21–22) and repelling an attack on London, Edward was secure for the remainder of his life.

The second half of Edward's reign. He was now able to revive the project of an invasion of France in concert with the duke of Burgundy. He made great preparations in 1474 and obtained a large grant from Parliament. In 1475 he invaded France with the largest army, it was said, that had ever left England, but he found the Duke of Burgundy very ill-prepared and the French formidable and willing to buy him out. Hence the Treaty of Picquigny was made by which Edward agreed to withdraw from France in return for 75,000 gold crowns down and a pension of 50,000 gold crowns a year. These sums helped to free Edward from dependence on parliamentary grants. As he grew older, he showed considerable ingenuity in raising money by reviving obsolescent rights and using doubtfully legal devices. Commercial treaties with France (1475), Burgundy (1468), and the Hanseatic League (1474) combined with external peace and growing internal order to revive trade strikingly after 1475, and this benefitted the customs duties and other revenues. Edward became a trader himself, transporting goods in his own ships and those of foreign merchants. He began a reorganization of the revenues from the crown estates, experimenting with methods of improving yields and promoting more efficient auditing under officials of the flexible royal household treasury instead of the unadaptable exchequer. These and other measures enabled him to leave behind a fortune; some of his improved financial administration was continued and developed by his successors Richard III and Henry VII.

The last decade of Edward's reign also saw an improvement in law enforcement. One especially disturbed area was Wales and the Welsh marches; Edward used the royal estates there as a foundation on which to base a council that acted in the name of his infant heir, the Prince of Wales, and employed the royal prerogative to make a start in repressing disorder. It was the forerunner of the council of Wales and the marches that subjugated the area to English rule.

Achievements

Modern research has emphasized these administrative achievements of Edward IV, and contemporary and Tudor historians viewed his later years as a time of prosperity and success. He rebuilt St. George's Chapel, Windsor, and collected illuminated Flemish manuscripts. He was also a friend and patron of the printer William Caxton, and his book collection became the foundation of the Old Royal library, later one of the glories of the British Museum.

Edward's promiscuity enabled Richard of Gloucester, after his death, to question the validity of his marriage and so to ruin his sons. As a young man Edward had been trustful and openhanded, but his experiences made him increasingly suspicious, leading him in 1478 to execute his brother George, duke of Clarence, who in former years had sided with Warwick against him. In 1482, Louis XI, in order to come to terms with the rulers of Burgundy, tacitly repudiated the Treaty of Picquigny and the annual tribute that it provided. Edward contemplated a fresh invasion of France, but before it could be carried out he fell ill and died at Westminster on April 9, 1483, at the age of only 40. By Elizabeth Woodville he had seven children who survived him: two sons, Edward (afterward Edward V) and Richard, duke of York, who were probably murdered in the Tower of London in August 1483, and five daughters, of whom the eldest, Elizabeth, married Henry VII.

BIBLIOGRAPHY. C.L. SCOFIELD, *The Life and Reign of Edward IV*, 2 vol. (1923), is extremely detailed, accurate, and based on original sources; but the detail often obscures the argument, and the general outlook is now somewhat out of date. E.F. JACOB, *The Fifteenth Century, 1399–1485* (1961), provides a more up-to-date interpretation with an annotated bibliography. J.R. LANDER, *Conflict and Stability in Fifteenth Century England* (1969), puts the reign of Edward IV in its setting and has a useful annotated bibliography. A.R. MYERS, *English Historical Documents, 1327–1485* (1969), provides a large number of documents, translated into English, for the reign of Edward IV.

(A.R.M.)

Edward, the Black Prince

Son and heir of the English king Edward III (reigned 1327–77), Edward of Woodstock, the "Black Prince," was one of the outstanding commanders of the Hundred Years' War and is often considered the embodiment of the ideal of chivalry. His sobriquet, said to have come from his wearing black armour, has no contemporary justification and is found first in Richard Grafton's *Chronicle of England* (1568).

Edward, the Black Prince, electrotype from effigy in Canterbury Cathedral, c. 1376. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

Born at Woodstock on June 15, 1330, to Edward III and Philippa of Hainault, Edward was created earl of Chester (March 1333), duke of Cornwall (February 1337)—the first appearance of this rank in England—and prince of Wales (May 1343); he was prince of Aquitaine from 1362 to 1372. His first campaign was served under his father in northern France (1346–47) and at the Battle of Crécy (August 26, 1346) he won both his spurs and the famous ostrich plumes and with them the mottoes used by himself and subsequent princes of Wales, *houmout*; *ich dene* ("Courage; I serve"; the words are here spelled as Edward himself wrote them; later variants include *houmout* and *ich dien* or *ich diene*). One of the original Knights of the Garter, he was sent to France with independent command in 1355, winning his most famous victory over the French at Poitiers on September 19, 1356. The French king John II, brought captive to England, was treated by the Prince with a celebrated courtesy, but he was obligated to pay a ransom of 3,000,000 gold crowns and to negotiate the treaties of Brétigny and Calais (1360) by which Aquitaine was ceded to the English.

Victory at Poitiers

Edward married his cousin Joan, the divorced and widowed countess of Kent, in October 1361. He was created prince of Aquitaine in July 1362 and left England in 1363 to take up his duties. His powers and his opportunities were great, but his rule was a failure, and he himself was largely to blame. His court at Bordeaux, that of a foreign conqueror, was extravagant; the 13 *sénéchaussées* into which the principality was divided administratively followed their earlier French pattern and allowed local French loyalites to subsist; his relations with the many

bishops were unfriendly, while the greater nobles, Arnaud-Amanieu, sire d'Albret, Gaston II, comte de Foix, and Jean I, comte d'Armagnac, were hostile. He summoned several estates, or parliaments, but always to levy taxes. In 1367 he undertook to restore Peter the Cruel of Castile to his throne, and though he won a classic victory at Nájera on April 3, 1367, the campaign ruined his health, his finances, and any prospect of sound rule in Aquitaine, where, in 1368, the nobles and prelates appealed against him to Charles V of France as suzerain. Edward's reply to the French king's citation to answer the appellants before the *parlement* of Paris in May 1369 is well known—he would appear with 60,000 men at his back. He had, however, alienated the towns and peasantry as well as the nobles; and by March 1369 more than 900 towns, castles, and strong places had declared against him. Relying on mercenaries whom he could not afford to pay, he was powerless to quell the revolt, and the terrible sack of Limoges (October 1370) merely redounded to his discredit. He returned to England a sick and broken man in January 1371 and formally surrendered his principality to his father in October 1372, alleging that the revenues of the country were insufficient to defray his expenses. He had no successor as prince of Aquitaine.

His life in England

Edward's position in England, where, throughout his life, he was heir apparent, was that of a typical 14th-century magnate. The registers of his household from 1346 to 1348 and from 1351 to 1365 have survived and add to what is known of him from the chroniclers and from his biographer, the herald of Sir John Chandos. In one important respect all of these sources paint the same picture, that of a man constantly living beyond his means. His generosity, however, extended to his tenants as well as to his knightly companions, and faithful service was rewarded, as in 1356 when the ferry of Saltash was granted to William Lenche, who had lost an eye at Poitiers.

The Prince paid two visits to Chester—in 1353 and again in 1358. Cheshire furnished many of his archers, who wore a rudimentary uniform of a short coat and hat of green and white cloth with the green on the right. Despite his title, however, Edward did not visit Wales.

He appears to have shared the interests of his class—jousting, falconry, hunting, gaming. He was literate and conventionally pious, substantially endowing a religious house at Ashridge (1376). He had the customary fine presence of the Plantagenets and shared their love of jewels. The Black Prince's ruby in the present imperial state crown may or may not have been given to him by King Peter after the Battle of Nájera, but he would certainly have prized it, as a connoisseur. Similar artistic interest is shown in his seals, adorned with their ostrich feathers, and in the elegant gold coins that he issued as prince of Aquitaine.

The last five years of the Prince's life are obscure. Some contemporaries suggest that he supported the Commons when political discontent culminated in the Good Parliament of April 1376; but he knew he was dying, and he was probably seeking the best means to ensure the succession of his second, but only surviving son, Richard of Bordeaux (afterward Richard II). He died at Westminster on June 8, 1376, and was buried at Canterbury, where his tomb with his accoutrements, recently restored and renovated, still stands.

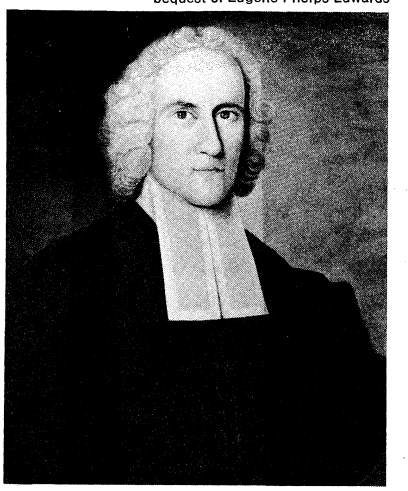
BIBLIOGRAPHY. M.K. POPE and E.C. LODGE (eds.), *The Life of the Black Prince, by the Herald of Sir John Chandos* (1910), written in French, probably in 1397, is a metrical biography, not at all critical, dealing with the knightly exploits of the Prince. The works of Robert of Avesbury and Thomas Walsingham in the "Rolls Series" are useful for general reference. T.F. TOUT, *Chapters in Administrative History*, vol. 5 (1930), is good on the central administrative system of the Black Prince's household. The most interesting and valuable records printed are *The Black Prince's Registers*, 4 vol. (HMSO, 1930–33); the first deals with the years 1346–48, the other three with the years 1351–66. For the affairs of Aquitaine, vol. 4 of R. DELACHENAL, *Histoire de Charles V* (1928), should be consulted.

(I.P.S.)

Edwards, Jonathan

Jonathan Edwards, the greatest theologian and philosopher of American Puritanism, stimulated the religious revival known as the Great Awakening (1740–42) by his sermons and reports and became its most able literary spokesman. His writings and example helped usher in the great age of Protestant missionary expansion in the 19th century. In *A Careful and Strict Inquiry into the Modern Prevailing Notions of that Freedom of Will* and other works, he attempted a philosophical defense of orthodox Calvinism but in so original a fashion as to found a distinct school of American theology. More highly esteemed today are his *Treatise Concerning Religious Affections* and *The Nature of True Virtue*, as major works in theological ethics.

By courtesy of Yale University Art Gallery
bequest of Eugene Phelps Edwards



Edwards, oil painting by Joseph Badger.
In the Yale University Art Gallery.

Early life and ministry. Edwards was born October 5, 1703, at East Windsor, Connecticut. His father Timothy was pastor of the church there; his mother Esther was a daughter of Solomon Stoddard, pastor of the church at Northhampton, Massachusetts. Jonathan was the fifth child and only son among 11 children; he grew up in an atmosphere of Puritan piety, affection, and learning. After a rigorous schooling at home, he entered Yale College in New Haven, Connecticut, at the age of 13. He was graduated in 1720 but remained at New Haven for two years, studying divinity. After a brief New York pastorate (1722–23), he received the M.A. degree in 1723; during most of 1724–26 he was a tutor at Yale. In 1727 he became his grandfather's colleague at Northampton. In the same year, he married Sarah Pierrepont, who combined a deep, often ecstatic, piety with personal winsomeness and practical good sense. To them were born 11 children.

Early intellectual interests

The manuscripts that survive from his student days exhibit Edwards' remarkable powers of observation and analysis ("Of Insects"), the fascination that the English scientist Isaac Newton's optical theories held for him ("Of the Rainbow"), and his ambition to publish scientific and philosophical works in confutation of materialism and atheism ("Natural Philosophy"). Throughout his life he habitually studied with pen in hand, recording his thoughts in numerous hand-sewn notebooks; one of these, his "Catalogue" of books, demonstrates the wide variety of his interests.

Edwards did not accept his theological inheritance passively. In his "Personal Narrative" he confesses that, from his childhood on, his mind "had been full of objections" against the doctrine of predestination—*i.e.*, that God sovereignly chooses some to salvation but rejects others to everlasting torment; "it used to appear like a horrible doctrine to me." Though he gradually worked through his intellectual objections, it was only with his conversion (early in 1721) that he came to a "delightful

conviction" of divine sovereignty, to a "new sense" of God's glory revealed in Scripture and in nature. This became the centre of Edwards' piety: a direct, intuitive apprehension of God in all his glory, a sight and taste of Christ's majesty and beauty far beyond all "notional" understanding, immediately imparted to the soul (as a 1734 sermon title puts it) by "a divine and supernatural light." This alone confers worth on man, and in this consists his salvation. What such a God does must be right; hence, Edwards' cosmic optimism. The acceptance and affirmation of God as he is and does and the love of God simply because he is God, became central motifs in all of Edwards' preaching.

Under the influence of Puritan and other Reformed divines, the Cambridge Platonists, and British philosopher-scientists such as Newton and Locke, Edwards began to sketch in his manuscripts the outlines of a "Rational Account" of the doctrines of Christianity in terms of contemporary philosophy. In the essay "Of Being," he argued from the inconceivability of absolute Nothing to the existence of God as the eternal omnipresent Being. It was also inconceivable to him that anything should exist (even universal Being) apart from consciousness; hence, material things exist only as ideas in perceiving minds; the universe depends for its being every moment on the knowledge and creative will of God; and "spirits only are properly substance." Further, if all knowledge is ultimately from sensation (Locke) and if a sense perception is merely God's method of communicating ideas to the mind, then all knowledge is directly dependent on the divine will to reveal; and a saving knowledge of God and spiritual things is possible only to those who have received the gift of the "new sense." This grace is independent of human effort and is "irresistible," for the perception of God's beauty and goodness that it confers is in its very nature a glad "consent." Nevertheless, God decrees conversion and a holy life as well as ultimate felicity; and he has so constituted things that "means of grace" (*e.g.*, sermons, sacraments, even the fear of hell) are employed by the Spirit in conversion, though not as "proper causes." Thus, the predestinarian preacher could appeal to the emotions and wills of men.

Pastorate at Northampton, Massachusetts. At Stoddard's death in 1729, Edwards became sole occupant of the Northampton pulpit, the most important in Massachusetts outside of Boston. In his first published sermon, preached in 1731 to the Boston clergy and significantly entitled *God Glorified in the Work of Redemption, by the Greatness of Man's Dependence upon Him, in the Whole of It*, Edwards blamed New England's moral ills on its assumption of religious and moral self-sufficiency. Because God is the saints' whole good, faith, which abases man and exalts God, must be insisted on as the only means of salvation. The English colonists' enterprising spirit made them susceptible to a version of Arminianism (deriving from the Dutch theologian Jacobus Arminius), which was popular in the Anglican Church and spreading among dissenters; it minimized the disabling effects of original sin, stressed free will, and tended to make morality the essence of religion. Against these ideas Edwards also delivered a series of sermons on "Justification by Faith Alone" in November 1734. The result was a great revival in Northampton and along the Connecticut River Valley in the winter and spring of 1734–35, during which period more than 300 of Edwards' people made professions of faith. His subsequent report, *A Faithful Narrative of the Surprising Work of God* (1737), made a profound impression in America and Europe, particularly through his description of the types and stages of conversion experience.

The Great Awakening

In 1740–42 came the Great Awakening throughout the colonies. George Whitefield, a highly successful evangelist in the English Methodist movement, and Gilbert Tennent, a Presbyterian minister from New Jersey, drew huge crowds; their "pathetical" (*i.e.*, emotional) sermons resulted in violent emotional response and mass conversions. Edwards himself, though he held his own congregation relatively calm, employed the "preaching of terror" on several occasions, as in the Enfield sermon, "Sinners in the Hands of an Angry God" (1741). The Awakening produced not only conversions and changed lives but also excesses, disorders, and ecclesiastical and civil disruptions. Though increasingly critical of attitudes and practices associated with the revival, to the extent of personally rebuking Whitefield, Edwards maintained that it was a genuine work of God, which needed to be furthered and purified. In defense and criticism of the Awakening he wrote *The Distinguishing Marks of a Work of the Spirit of God* (1741), *Some Thoughts Concerning the Present Revival of Religion in New England* (1742), and *A Treatise Concerning Religious Affections* (1746). In the *Affections*, Edwards insisted, against the revival critics' ideal of sober, "reasonable" religion, that "the essence of all true religion lies in holy love," a love that proves its genuineness by its inner quality and practical results. In 1749 he edited, with "Reflections," the memoirs of David Brainerd, a young New Light revivalist who became a Presbyterian missionary to the Indians and died in 1747. The volume became a highly influential missionary biography. Edwards' *Humble Attempt to Promote Explicit Agreement and Visible Union of God's People in Extraordinary Prayer* (1747), written in support of a proposed international "concert of prayer" for "the Revival of Religion and the Advancement of Christ's Kingdom on Earth," helped to remove a major ideological barrier to missionary activity by arguing that the worst of the "great tribulations" (prophesied in the book of Revelation to John as preceding the millennium) were already past and that the church could look forward to an increasing success of the gospel among men.

Dismissal from the Northampton pulpit

Meanwhile, Edwards' relations with his own congregation had become strained; one reason for it was his changed views on the requirements for admission to the Lord's Supper. In the Halfway Covenant, baptized but unconverted children of believers might have their own children baptized by "owning the covenant"; Stoddard had instituted the subsequently widespread practice of admitting to the Eucharist all who were thus "in the covenant," even if they knew themselves to be unconverted. Edwards gradually came to believe that the profession required for admission to full communion should be understood to imply genuine faith, not merely doctrinal knowledge and good moral behaviour. The public announcement of his position in 1749 precipitated a violent controversy that resulted in his dismissal. On July 1, 1750, Edwards preached his dignified and restrained "Farewel-Sermon." In this controversy he wrote two books, *Qualifications for Communion* (1749) and *Misrepresentations Corrected, and Truth Vindicated, in a Reply to the Rev. Mr. Solomon Williams's Book* (1752), one to convince his congregation, the other to correct what he considered misrepresentations of his views by a kinsman, the pastor at Lebanon, Connecticut. Though Edwards himself was defeated, his position finally triumphed and provided New England Congregationalism with a doctrine of church membership more appropriate to its situation after disestablishment.

Pastorate at Stockbridge, Massachusetts. In 1751, Edwards became pastor of the frontier church at Stockbridge, Massachusetts, and missionary to the Indians there. Hampered by language difficulties, illness, Indian wars, and conflicts with powerful personal enemies, he nevertheless discharged his pastoral duties and found time to write his famous work on the *Freedom of Will* (1754). The will, said Edwards, is not a separate, self-determining faculty with power to act contrary to the strongest motives, as he understood the Arminians to teach. Rather, it is identical with feelings or preference, and a volition is simply the soul's "prevailing inclination" in action; the will "is as the greatest apparent good." Men are free to do as they please, and God therefore rightly holds them morally responsible for the quality of their volitions as expressions of their desires and intentions.

By 1757 Edwards had finished his *Great Christian Doctrine of Original Sin Defended* (1758), which was mainly a reply to the English divine, John Taylor of Norwich, whose works attacking Calvinism (based on the thought of the 16th-century Protestant Reformer John Calvin)

had "made a mighty noise in America." Edwards defended the doctrine not only by citing biblical statements about the corruption of man's heart but by arguing that the empirical evidence of men's universal commission of sinful acts points to a sinful predisposition in every man. In answering Arminian objections to the notion that God "imputed" Adam's guilt to his posterity, Edwards proposed a novel theory of identity by divine "constitution" to account for men's unity with Adam and suggested that their innate corruption is not a judicial punishment for Adam's guilt but is really their own because of their participation (being one with him) in the sinful inclination that preceded Adam's sinful act. Edwards' was the first major contribution to the long debate about human nature in American theology and helped set the terms of that debate.

Edwards' defense of Calvinism

Edwards perceived the threat in Taylor's notion of man's innate goodness and autonomy; the whole Christian conception of supernatural redemption seemed to be at stake. He therefore planned further treatises, of which he completed two posthumously published dissertations: *Concerning the End for Which God Created the World* and *The Nature of True Virtue* (1765). God's glory, not human happiness, is his end in creation; but this is because God in his all-sufficient fullness must communicate himself by the exercise of his attributes. God can be said to aim at the creature's happiness, but it is a happiness that consists in contemplating and rejoicing in God's glory manifested in creation and redemption. Edwards defines true virtue as disinterested love (benevolence) toward God as Being in general and toward all lesser beings according to their degree of being. True virtue, therefore, does not spring from self-love or from any earthbound altruism (two prime 18th-century views); love to self, family, nation, or even mankind is good only if these lesser systems of being do not usurp the place of highest regard that belongs to God alone.

Edwards also projected books on other subjects, notably *A History of the Work of Redemption* (he had preached a series of sermons—posthumously published—on that subject in 1739), which was to be a complete theology combining biblical, historical, and systematic materials "in an entire new method." Late in 1757, however, he accepted the presidency of the College of New Jersey (later Princeton University) and arrived there in January. He had hardly assumed his duties when he contracted smallpox and died at Princeton on March 22, 1758.

Influence. Edwards' immediate disciples, Joseph Bellamy (pastor at Bethlehem, Connecticut), Samuel Hopkins (pastor at Great Barrington, Massachusetts, later at Newport, Rhode Island), and Jonathan Edwards, Jr., developed some of his "improvements" into a distinct theological school; it was first called "Hopkinsianism" and later the "New England Theology." These men and their successors, in their effort to defend Calvinism against Arminians, Unitarians (those who denied the doctrine of the Trinity), and "infidels," made important modifications in some of its doctrines and thus prepared the way for later 19th-century evangelical liberalism.

Edwards' influence on American Protestantism for a century after his death was very pronounced, and he was widely read in the British Isles. In a general revolt against Puritanism and Calvinism after the U.S. Civil War (1861–65), Edwards' prestige declined, and he was remembered mainly as a hell-fire preacher or as an abstruse, absent-minded metaphysician. In the 1930s and after, he was rediscovered by theologians reacting against liberalism and by secular scholars seeking to delineate the "American mind." Edwards' ability to combine religious intensity with intellectual rigour and moral earnestness, the cosmic sweep of his theological vision, his emphasis on faith as an "existential" response to reality, his insistence that love is the heart of religion, and his uncompromising stand against all forms of idolatry are some of the reasons his life and writings are again being seriously studied.

BIBLIOGRAPHY. PERRY MILLER and JOHN E. SMITH (eds.), *The Works of Jonathan Edwards*, 4 vol. to date (1957–), will be a complete critical edition. OLA ELIZABETH WINSLOW, *Jonathan Edwards, 1703–1758* (1940), containing a delineation of Edwards the man; and ARTHUR C. MCGIFFERT, JR., *Jonathan Edwards* (1932), containing a good understanding of his thought, are regarded as the best biographies. PERRY MILLER, *Jonathan Edwards* (1949), an intellectual biography, is a brilliant interpretation that has been accused of unduly modernizing Edwards. Additional sources of information may be found in CLARENCE H. FAUST and THOMAS H. JOHNSON (eds.), *Jonathan Edwards: Representative Selections*, rev. ed. (1962), the best handbook, including an extensive annotated bibliography; and NELSON R. BURR, *A Critical Bibliography of Religion in America*, vol. 2, pp. 976–987 (1961), a bibliographical essay. THOMAS H. JOHNSON, *The Printed Writings of Jonathan Edwards, 1703–1758: A Bibliography* (1940), lists editions, reprints, and variants. Other interpretive studies include: DOUGLAS J. ELWOOD, *The Philosophical Theology of Jonathan Edwards* (1960), which stresses the Platonic and Augustinian elements; CONRAD CHERRY, *The Theology of Jonathan Edwards: A Reappraisal* (1966), an excellent theological study; and ROLAND A. DELATTRE, *Beauty and Sensibility in the Thought of Jonathan Edwards: An Essay in Aesthetics and Theological Ethics* (1968).

(T.A.S.)

Egg Production, Commercial

While the primary role of the egg obviously is to reproduce the species (for zoological information on eggs, see BIRD), most eggs produced by domestic fowl, except those specifically set aside for hatching purposes, are not fertilized, but are sold mainly for human consumption. They may be purchased by the consumer; or, in liquid, dehydrated, or frozen form, by quantity buyers for use in food service operations or for inclusion in commercially manufactured food products. Eggs produced in quantity come from chickens, ducks, geese, turkeys, guinea fowl, pigeons, pheasants, and quail. This article deals with world egg production and consumption; the selection of laying hens; the cleaning, oiling, sizing, and grading of eggs; and their preparation and use in the human diet.

PRODUCTION AND CONSUMPTION STATISTICS

Egg-producing fowl

Chicken eggs represent the bulk of egg production in the United States and Europe. Though some laying hens are known to produce an egg every day, in commercial installations the average for layers is about 220 eggs per year. In several parts of the world, the duck is important as an egg producer and the egg-laying ability of certain types excels that of chickens. Some ducks produce from 300 to 360 eggs a year, but the average is much less. Duck eggs are larger and heavier than chicken eggs, and are used as food in western Europe and the Orient. Geese are raised extensively in Poland and some of the western European countries, and their eggs are often used for foods. Commercial production of turkey and pigeon eggs is almost entirely confined to those used for producing turkey poults and young pigeons (squabs). Pheasant and quail eggs provide birds for hobby or sport use.

World production. Annual world production of hen eggs is about 390,000,000,000. The leading egg-producing country is China, with an annual production of 73,000,000,000 eggs. The United States ranks second, with some 314,000,000 laying birds whose yearly egg production of about 70,000,000,000 eggs yields a gross income of over $2,200,000,000. The Soviet Union ranks third as an egg-producing nation, and Japan fourth. Other countries in order of production are the United Kingdom, West Germany, France, Brazil, Italy, Spain, Poland, and Canada. In the United States during 1970, production for each of the ten major egg-producing states ranged from over 2,500,000,000 to over 8,300,000,000. Annual production in California alone equalled that of Italy.

European production. In Europe, egg production in the 1960s increased steadily at a rate of about 3 percent a year. The rate was somewhat higher for eastern than for western Europe; *i.e.*, 3.3 percent versus 2.7 percent. In the Soviet Union production jumped considerably, being 40 percent higher in 1970 than in 1960. There was a phenomenal increase in egg production of Japan, where during the period 1948–52, production averaged 2,245,000,000 eggs, rising in 1970 to 28,571,000,000.

U.S. production. In the U.S. production rose about 14 percent from 1960 to 1970, but a levelling off after 1967 was attributed to a lack of incentive for increasing production because of low farm prices and a reduction of exports. In 1958, the United States had exported about five times as many eggs as it had a decade later. Several countries that had formerly imported eggs attained self-sufficiency. This change in dependence on importation of eggs also affected other countries. For example, Denmark, traditionally an important exporter of eggs, cut back its output substantially because of reduced export outlets.

Imports and exports

Consumption. During the 20th century, use of eggs in the United States increased annually until, in 1951, it reached 392 eggs per person per year, after which the figure fell steadily until, by 1966, it was only 314. In the United Kingdom the per capita use of 270 eggs per year is somewhat higher than the average for other countries of western Europe (210). The number of eggs consumed in Japan per capita in 1960 was 98; in 1964, it was 180; and in 1969, it was 269. The increase in consumption reflects a rising standard of living in Japan. People of West Germany and Austria each consume about 250 eggs per year. Per capita egg consumption in the U.S.S.R. and eastern Europe is much lower: 140 and 190 respectively.

FACTORS IN PRODUCTION AND QUALITY OF EGGS

Selection of laying hens. Chickens are the world's most widely distributed domestic fowl. Commercially, the major chicken breeds for egg production include several of the varieties of the Leghorn in the Mediterranean class and such varieties as white Plymouth Rock and New Hampshire breeds in the dual-purpose, or American class. Breeders have selected birds with high rates of egg production and by inbreeding, single crossing of inbred lines, or multiple crossing of lines, have evolved pullets (immature hens) that can maintain good performance for a year or more. Ordinarily, pullets are placed in production at 5 to 6 months of age and are kept as layers until they are about 18 months old. While older layers produce large eggs, the shell usually does not have good quality and production is reduced after one year of laying.

Cleaning and oiling. In the United States shell eggs are shipped from the producer to the packing station in cardboard containers holding either 15 dozen or 30 dozen eggs separated by fillers and flats. Such eggs are unsorted as to size and grade. When received at the plant, eggs are placed into a cool 50°–60° F (10°–16° C) humidified room.

Soiled eggs

Soiled eggs cause heavy financial losses. Dirty eggs bring less per dozen because the cost of cleaning them is high and their keeping quality is poor. Such eggs are cleaned by washing or by dry cleaning the dirty or stained area with an abrasive emery cloth or fine sandpaper.

When eggs are washed, a solution at least 20° F warmer than the egg and having a temperature of 90° F (32° C) or more is used so that the egg's interior will not contract and allow polluted water to be drawn into the shell. The washed egg is rinsed immediately with warm water containing a sanitizer and dried before packing. Dry cleaning is little used because it is slow, does not do a thorough job, and often results in excessive breakage.

Oiling eggs reduces evaporation of moisture and replaces the cuticle lost during washing. As soon as possible after the eggs are laid or washed, a fine film of lightweight mineral oil, which is colourless, odourless, and tasteless, is sprayed over the egg shell to seal the pores. Oiling also cuts weight loss and halts air cell expansion.

The keeping quality of cleaned eggs also can be improved by holding them in water at 130° F (55° C) for 15 minutes or at slightly higher temperatures for shorter periods of time. This form of thermostabilization causes a slight thickening of the albumen adjacent to the shell membranes and acts as a seal. Thermostabilizing eggs is not common.

Egg weights. Eggs are graded on the basis of size and quality. There are four general weight classes: extra large, large, medium, and small. The minimum net weights per dozen for eggs in these grades are respectively 27, 24, 21, and 18 ounces. In addition to these classes, two others occasionally are sold: *i.e.*, jumbo (30 ounces per dozen) and peewee (15 ounces per dozen).

Both exterior and interior quality are used in determining the grade of an egg. Criteria used for this assessment include: shell condition and cleanliness, conformity and depth of air cell, clearness and firmness of white, and definition and shape of yolk. For these determinations, the interior of the egg is viewed by transmitted light, a process called candling.

Egg candling. When eggs are candled by hand, the operator supports the egg between the thumb and index finger with the small end toward the palm of the hand. Then by a twist of the hand and wrist the egg is quickly rotated through a 180° arc so that the contents are sent spinning. At that instant the large end is held before a candling lamp so that the light shines through and reveals the air cell and yolk, albumen, and shell conditions. Most commercial grading operations take place in cool, darkened rooms and hand candling is used only occasionally to check the accuracy of grading. Usually the quality of the egg is determined by mass scanning of several rows of eggs at a time as they move automatically over a light source that shines through them. During the same operation the eggs pass over tripping devices calibrated to sort them into various weight classes. Many types of automatic and semiautomatic candling and grading devices are available.

While most candling of eggs is done using incandescent light, the use of "black light," or ultraviolet candling, has proved valuable in detecting spoilage caused by the bacterium *Pseudomonas*, since many of these organisms produce a characteristic green or blue-green fluorescence when illuminated by ultraviolet light.

Egg quality

Grades. Four edible grades of shell eggs are recognized: AA (or fresh fancy), A, B, and C. For AA grade eggs, the shell eggs must be clean, unbroken, of good even texture and strength, free from visible stains, discolorations, rough areas, or thin spots. The air cell must be ⅛ inch (3.2 millimetres) or less in depth, the white clear and firm, and the outline of the yolk only slightly defined and practically free of defects. In shelled eggs, egg white thickness is indicated numerically in Haugh units. When broken out, the egg must cover a small area, have a firm albumen, or thick white portion (72 Haugh units minimum), with much thick white surrounding a round and upstanding yolk. In A grade eggs, the air cell may be 3/16 inch (4.8 millimetres) in depth and the yolk fairly well defined; when such eggs are broken out, they cover a moderate area due to a medium amount of thin white present (60 Haugh units minimum). The shell of a B grade egg may be slightly faulty in texture or strength and slightly stained (not over 1/16 of the shell surface); the air cell may be as much as ⅜ inch (9.5 millimetres) in depth. The white may be slightly weak and the yolk outline well defined, slightly enlarged and flattened, and have defects not considered serious. The broken-out B grade egg covers a wide area. The yolk is flattened and enlarged and the albumen slightly weak (31 Haugh units minimum). C grade eggs may be moderately stained (less than ¼ of the shell surface), misshapen, or faulty in texture or strength. The air cell may be over ⅜ inch in depth, bubbly or free moving. The white may be weak and watery and small blood clots or meat spots may be present. The yolk outline, which is plainly visible, shows that the yolk is enlarged or flattened; there may be germ development but no blood. When broken out, the egg covers a very wide area. There is no thick white (less than 31 Haugh units) but a very large amount of thin white and a very flat and enlarged yolk.

Substandard eggs. In addition to these four salable grades of shell eggs, dirty and check eggs are permitted for use in egg products providing there is no leakage of the interior egg content; *i.e.*, they may have moderate stains covering more than ¼ of the shell surface or even prominent stains or adhering dirt and, with check eggs, the shell may be broken, but there must be an intact membrane (as demonstrated by candled appearance).

Eggs are designated as inedible if black-, white-, or mixed-rots are present. Also, moldy, musty, sour, or

adulterated eggs are not permitted for sale as food nor are those with green or bloody whites or with stuck yolks, incubator rejects, or eggs containing embryo chicks at or beyond the blood ring condition, or any in such condition that washing and breaking result in contamination.

Shell colour. The shell varies in colour from white to brown but colour does not affect the quality of the egg. A pigment imparted to the shell while the egg is in the oviduct is laid down in the spongy layer of the shell and is derived from the blood. Leghorn hens lay much whiter-shelled eggs than do most American classes. In some areas, white eggs are preferred, while in other regions, those with brown shells are chosen. Usually, more difficulty is experienced in detecting the interior quality of brown shelled eggs; small meat or bloodspots are harder for the candler to see at the packing station. White eggs and brown eggs may be colour sorted. When eggs with mixed shell colours are cartoned together, the consumer may have the erroneous impression that these lots are heterogeneous in other respects as well.

Packaging. For retail use in the United States, eggs are repackaged in dozen and half-dozen paperboard cartons. In some other countries they are packed with straw or excelsior in long wooden boxes. In many parts of the world, they are marketed in baskets or boxes and the individual eggs are sold by weight. Several European countries stamp each egg with a date and number to meet the import restrictions of other nations.

USES AND PROCESSING OF EGGS AND EGG PRODUCTS

Use of sub-standard eggs

The bulk of egg production is sold directly to the consumer for table use. Those that are not, including most of the smaller sizes and undergrades, go to egg breaking firms where they are broken out for use as frozen or dried egg or egg products. A large percentage of the liquid egg is frozen for use by bakers and confectioners.

Uses. Some breakers use refrigerated bulk tank trucks for shipping broken out white, yolks, or a mixture to their freezing or drying plants. In addition, such refrigerated bulk tank shipments are often made directly to bakeries or confectioners.

Eggs are widely used as food product ingredients because of their leavening, emulsifying, binding, taste appeal, and nutritive values. In addition, eggs are used for purposes other than food, as in paints, varnishes, ink, soap, shampoos, animal feeds, fertilizers, and vaccines. The foaming ability and heat coagulability of the proteins in egg white have made this product invaluable to the candy and bakery trades. Foam is formed in egg white when air is incorporated during beating and a solid film results as the protein denatures at the surface. When heat is also applied, such as in cooking and baking, a solid structure is set up that supports other ingredients as well as the egg white foam. Excellent foam development and coagulation is also achieved when egg yolk is used in making some kinds of cakes and biscuits (cookies). Egg yolk is a stable emulsion that has unique value in the emulsification of products such as mayonnaise.

Pasteurization. Egg pasteurization is the process of heating liquid or whole egg products for sufficient time to destroy possible pathogens and organisms of fecal origin. In 1966, regulations of the U.S. Food and Drug Administration required that processed egg products in interstate commerce be pasteurized or otherwise treated to destroy all viable *Salmonella*, organisms that cause food poisoning and acute intestinal disorders in man. Efficient pasteurization is accomplished by what is called the "flash," or high temperature-short time (HTST) method or by the holding, or vat, procedure.

In the HTST process, liquid whole egg flows through a pipe at such speed that it is maintained at 140° F (60° C) for 3.5 minutes before the product is cooled, frozen, or dried. For plain liquid yolk, the temperature of pasteurization required is 142° F (61° C) and for salted or sugared yolk it is 146° F (63° C).

Egg white is more easily damaged by heat than liquid yolk or whole egg and cannot be pasteurized by ordinary methods without great loss of its functional properties. By the addition of lactic acid and aluminum salts, however, the albumen proteins are stabilized and the damage greatly reduced. Efficient pasteurization of liquid albumen also can be accomplished by adding 0.075–0.10 percent hydrogen peroxide to egg white heated for 1½ minutes and then holding for two more minutes at the same temperature. Albumen can be freed of *Salmonella* by storing the dried product in a "hot room" (122–125° F, 50°–52° C) for 5–7 days.

Protein stabilization

For the holding or vat method of pasteurization, liquid whole egg or egg white is held at 135° F (57° C) for 15 minutes, plain liquid yolk at 136° F (58° C) for 20 minutes, and salted or sugar yolk at 138° F (59° C) for 35 minutes.

Dried albumen. Since the turn of the century, dried albumen has been produced in China by allowing the organisms normally present in the broken-out liquid to ferment spontaneously and change the small amount of glucose to water, carbon dioxide, and cells before drying the product on pans, trays, or drums. It was recognized that when there was residual glucose, the dried product deteriorated quite rapidly during storage because of a "browning" reaction between the aldehyde group of the glucose and the amino groups of the proteins. Although solubility and other properties of the Chinese dried albumen were satisfactory, the sanitary shortcomings of the product made it necessary to replace the procedure by a controlled bacterial or yeast fermentation or by an enzymatic process utilizing a glucose oxidase to convert the glucose to gluconic acid.

Several procedures are employed by industry for the drying of albumen, including drying on trays, in cabinets or drying tunnels, on drums or belts, foam drying, and spray drying.

Dried whole egg. Whole egg solids are usually produced by spray drying. Glucose in whole-egg also was found to be responsible for the development of a glucose-cephalin reaction with the phospholipid fraction of the egg resulting in an off-flavoured product. Removal of the glucose prevented this reaction as well as the browning phenomenon described above for egg white. The addition of sucrose (a non-reducing sugar) to a level so that the dried product contains 20–30 percent sucrose has been found to decrease drying damage.

Dried yolk. Egg yolk solids are produced by spray drying. While this form of drying produces some damage to the residual sugar, the percentage is much less than with albumen or whole egg, and other considerations such as "oiling-off" and rancidification are of greater concern. Drying to a low moisture content and packing in an inert gas help to prevent rancidification.

BIBLIOGRAPHY. Much comprehensive literature on the production of eggs exists. General discussions of domesticated fowl are: M.A. JULL, "The Races of Domestic Fowl" and "Fowls of Forest and Stream Tamed by Man," *Natn. Geogr. Mag.*, 51:379–452 (1927), and 51:327–371 (1930).

Economic considerations and marketing information are presented in the following: "The European Egg Markets in the Sixties," *Monthly Bulletin of Agricultural Economics and Statistics*, Food and Agricultural Organization, Rome, vol. 17 (1968); and "Chickens and Eggs, April 1969," pub. by the Crop Reporting Board Statistical Reporting Service, United States Department of Agriculture (U.S.D.A.). Current information may be found in recent reports from these agencies.

Technical information on egg grading is given in the U.S.D.A. Agriculture Handbook No. 75, *Egg Grading Manual*, rev. ed. (1965).

Scientific background on factors affecting the use of eggs in the food industry, functional properties of eggs and egg products, and advances in egg processing technology is included in: R.H. FORSYTHE, "Some Factors Affecting the Use of Eggs in the Food Industry," *Cereal Sci. Today*, 2:211–216 (1957), "Recent Advances in Egg Processing Technology and Egg Products," *The Bakers' Dig.* 38:52–58 (1964); R.E. FEENEY, "The Properties of Eggs in Relation to Baking and Premixes," *ibid.*, 29:23–26 (1955); and G.J. BANWART and J.C. AYRES, "The Effect of High Temperature Storage on the Content of Salmonella and on the Functional Properties of Egg Whites," *Fd. Technol.*, 10:68–73 (1956).

The periodical *Poultry Science* (1922–), is devoted to the scientific study of poultry and eggs and provides much additional information on commercial production of eggs.

(J.C.A.)

Egypt, Arab Republic of

Egypt, or the Arab Republic of Egypt, to use its current title, consists of a rectangle of territory occupying the northeast corner of the African continent between the 22nd and 32nd parallels of latitude. It has a total area of about 386,900 square miles (1,002,000 square kilometres). Its land frontiers border Libya in the west, The Sudan in the south, and Israel in the northeast. In the north, its Mediterranean coastline is about 620 miles, and, in the east, its coastline on the Red Sea and the Gulf of Aqaba is about 1,835 miles. In 1971 the population was almost 34,130,000. The capital is Cairo.

Egypt's strategic location

Egypt has always been a nodal point for routes—westward along the coast of North Africa, northwest to Europe, northeast to the Levant, south along the Nile to Africa, and southeast to the Indian Ocean and the Far East. This natural advantage was enhanced in 1869 by the opening of the Suez Canal. The concern of the European powers to safeguard this link for strategic and commercial reasons is probably the most important single factor influencing the history of Egypt since the 19th century. The increasing presence of the United States and the Soviet Union in the Mediterranean since World War II has kept Egypt in the spotlight of world concern. It is not, however, simply in the context of the balance of power in the Mediterranean but also in Africa and in the Indian Ocean that its significance must be assessed.

Israeli forces have occupied the Sinai Peninsula and the Gaza Strip in eastern Egypt since the cease-fire following the Arab-Israeli War of 1967. (See the articles on ALEXANDRIA and CAIRO; for associated physical features, see NILE RIVER; RED SEA; SAHARA (DESERT); SINAI DESERT; and SUEZ CANAL; for historical aspects and for relations with Israel and Syria see EGYPT, HISTORY OF; and SYRIA AND PALESTINE, HISTORY OF.)

This article is divided into the following sections:

I. The landscape
  Description of the natural landscape
    Relief features
    Drainage and soils
    Climate
    Vegetation and animal life
  Traditional regions
  The landscape under human settlement
    Rural settlement
    Urban settlement
II. People and population
  Groups historically associated with the contemporary country
    Linguistic groups
    Ethnic groups
    Religious groups
  Demography
III. The national economy
  A comparative view of the country's regional and worldwide standing
  The extent and distribution of resources
  Sources of national income
    Agriculture and fishing
    Manufacturing
    Energy
    Financial services
    Foreign trade
  Management of the economy
  Transportation and communications
IV. Administration and social conditions
  The structure of government
    The constitutional framework
    Local government and administration
    The political process
  Justice
  Administrative services
    Educational services
    Health and welfare services
    Housing
    Police services
  Social conditions
  Cultural life and institutions
    The cultural milieu
    The state of the arts
    Cultural institutions
    Press and broadcasting
    Prospects for the future

I. The landscape

DESCRIPTION OF THE NATURAL LANDSCAPE

Relief features. The topography of Egypt is dominated by the Nile. For about 750 miles of its northward course through the country, the river cuts its way through bare desert—a sharply delineated strip of green, abundantly fecund in contrast to the desolation that surrounds it. From the Aswān High Dam in the south as far as Cairo in the north, it is hemmed into its trenchlike valley by bordering cliffs, but at Cairo these disappear, and the river begins to fan out into its delta. Although totally flat, apart from an occasional mound projecting through the alluvium, the delta is far from featureless; it is crisscrossed by a maze of canals and drainage channels, the banks of which project above the surface of the ground like brown veins.

The Nile divides the desert plateau through which it flows into two unequal sections—the Western Desert (Arabic aṣ-Ṣaḥrā' al-Gharbīyah), between the river and the Libyan frontier; and the Eastern Desert (Arabic aṣ-Ṣaḥrā' ash-Sharqīyah), extending to the Suez Canal, the Gulf of Suez, and the Red Sea. Each of them has its own different character, as does the third and smallest of the Egyptian deserts, that of Sinai. The Western (Libyan) Desert is arid and without wadis (dry beds of seasonal rivers), while the Eastern Desert is extensively dissected by wadis and fringed by rugged mountains in the east. The desert of central Sinai is open country, broken by isolated hills and scored by wadis.

The mountains

Contrary to popular belief, Egypt is not an unrelievedly flat country. Mountainous areas occur in the extreme southwest of the Western Desert, along the Red Sea coast, and in southern Sinai. The high ground in the southwest is associated with the Jabal al-'Uwaynāt mountain mass, which lies just outside Egyptian territory. A number of peaks in the Red Sea Hills (Ītbāy) rise to over 6,000 feet, and the highest, Jabal Shā'ib al-Banāt, reaches 7,173 feet (2,187 metres). The sharply serrated crests of the mountains of southern Sinai reach elevations of more than 8,000 feet; among them is Mt. Catherine, Egypt's highest mountain, which reaches an elevation of 8,668 feet (2,642 metres).

The coastal regions of Egypt, with the exception of the delta, are everywhere hemmed in either by desert or by mountain; they are arid or of very limited fertility. The coastal plain, in both the north and east, tends to be narrow; it seldom extends to more than 30 miles and is on average much less. With the exception of the cities of Alexandria, Port Said, and Suez and a scattering of small ports and resorts, the coastal regions are sparsely populated and underdeveloped.

Drainage and soils. Apart from the Nile, the only natural perennial surface drainage consists of a few meagre streams in the mountains of southern Sinai. Most of the valleys of the Eastern Desert drain westward to the Nile. They are water eroded but normally dry; only after heavy rainstorms in the Red Sea Hills do they carry torrents. The shorter valleys on the eastern flank of the Red Sea Hills drain toward the Red Sea; they, too, are normally dry. Drainage in the Sinai mountains is toward the gulfs of Suez and Aqaba; as in the Red Sea Hills, torrent action has produced deeply eroded valleys, normally dry.

The central plateau of Sinai drains northward toward Wādī al-'Arīsh, a depression in the desert that occasionally carries surface water. One of the features of the Western Desert is its aridity, as shown by the absence of drainage lines.

Outside the areas of Nile silt deposits, the nature of such cultivable soil as exists depends upon the availability of the water supply and the type of rock in the area. Thirty percent of the total land surface of Egypt consists of Nubian sandstone, which extends over the southern sections of both the Eastern and Western deserts. Limestone deposits of the Eocene Epoch (*i.e.*, from 38,000,000 to 54,000,000 years old) cover a further 20 percent of the land surface, including central Sinai and the central portions of both the Eastern and Western deserts. The north-

MAP INDEX

Political subdivisions

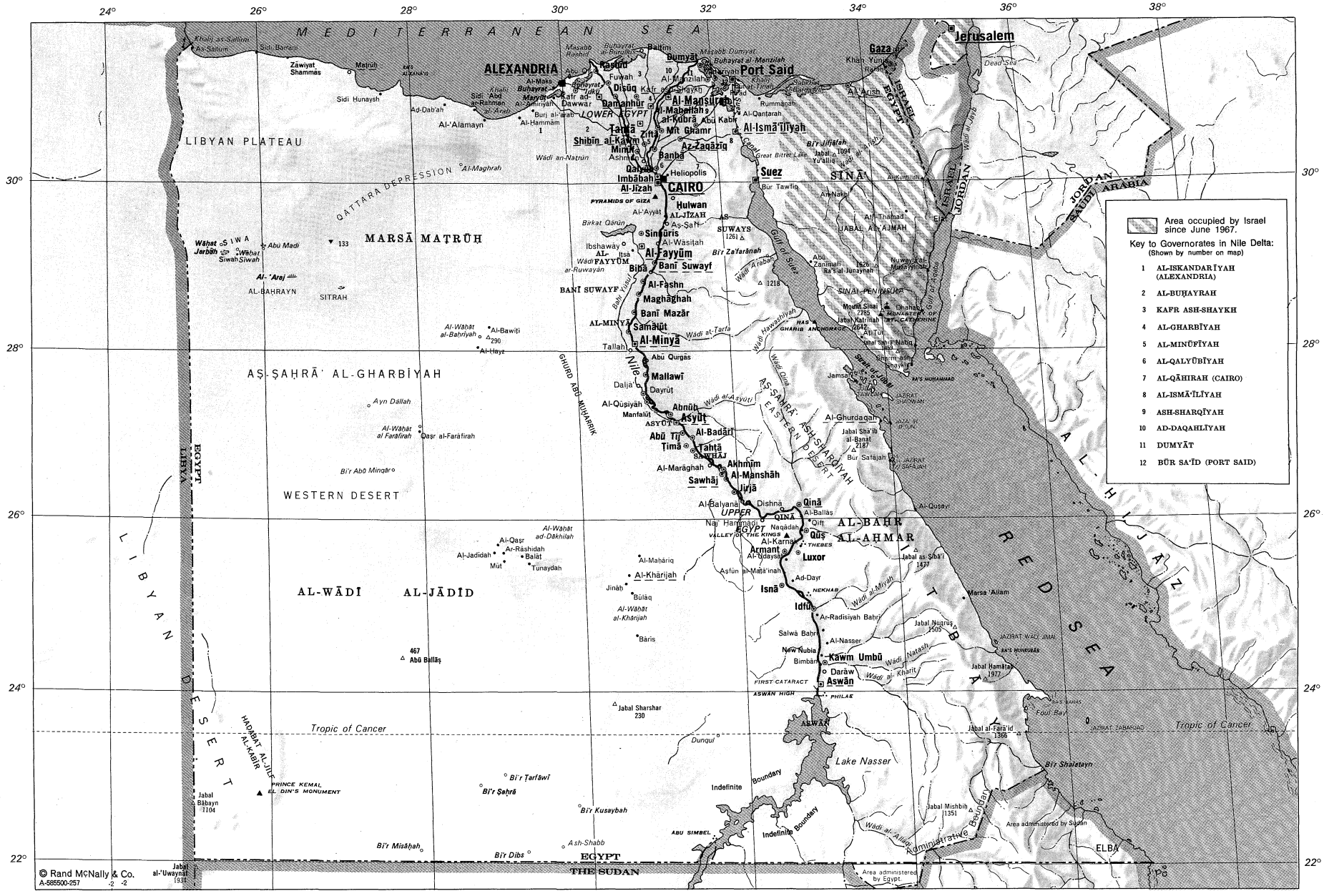
ad-Daqahlīyah...30·55n 31·30e
al-Baḥr al-Aḥmar 28·45n 32·00e
al-Buḥayrah.....30·59n 30·12e
al-Fayyūm.....29·19n 30·48e
al-Gharbīyah.....31·08n 31·00e
al-Iskandarīyah (Alexandria)....31·10n 29·53e
al-Ismā'īlīyah....30·30n 32·15e
al-Jīzah.....29·46n 31·18e
al-Minūfīyah.....30·30n 31·00e
al-Minyā.....28·26n 30·42e
al-Qāhirah (Cairo).....30·02n 31·13e
al-Qalyūbīyah....30·18n 31·18e
al-Wādī al-Jadīd .25·00n 28·00e
ash-Sharqīyah...30·48n 31·48e
as Suways.....29·59n 32·33e
Aswān.....23·05n 32·40e
Asyūṭ.....27·20n 30·50e
Banī Suwayf.....29·03n 31·02e
Bur Sa'īd (Port Said).....30·36n 32·15e
Dumyāṭ.....31·20n 31·45e
Kafr ash-Shaykh.31·15n 30·50e
Marsā Maṭrūḥ....28·00n 28·00e
Qinā.....25·50n 32·45e
Sawhāj.....26·20n 31·50e
Sīnā'.....30·15n 32·40e

Cities and towns

Abnūb.....27·16n 31·09e
Abū Kabīr.....30·44n 31·40e
Abū Madi (oasis).29·14n 26·00e
Abū Qīr.....31·19n 30·04e
Abū Qurgās.....27·56n 30·50e
Abū Tīj.....27·02n 31·19e
Abū Zanīmah....29·03n 33·06e
aḍ-Ḍab'ah.....31·02n 28·26e
ad-Dayr.....25·20n 32·35e
Akhmīm.....26·34n 31·44e
al-'Alamayn (el-'Alamein)....30·49n 28·57e
al-'Āmirīyah.....31·01n 29·48e
al-'Arīsh.....31·08n 33·48e
al-'Ayyāṭ.....29·37n 31·15e
al-Badārī.....26·59n 31·25e
al-Ballāṣ.....26·01n 32·46e
al-Balyanā.....26·14n 32·00e
al-Bawīṭī.....28·21n 28·52e
Alexandria.....31·12n 29·54e
al-Fashn.....28·49n 30·54e
al-Fayyūm.....29·19n 30·50e
al-Ghurdaqah....27·14n 33·50e
al-Ḥammām.....30·50n 29·23e
al-Ḥayz.....28·02n 28·39e
al-Ismā'īlīyah....30·35n 32·16e
al-Jadīdah.....25·34n 28·51e
al-Jafadīdah.....25·34n 28·51e
al-Jīzah.....30·01n 31·13e
al-Karnak.....25·43n 32·39e
al-Khārijah.....25·26n 30·33e
al-Kuntillah.....30·00n 34·41e
al-Maḥallah al-Kubrā.....30·58n 31·10e
al-Maḥārīq.....25·37n 30·39e
al-Maks (Mex)...31·09n 29·51e
al-Manshāh.....26·28n 31·48e
al-Manṣūrah.....31·03n 31·23e
al-Manzilah.....31·09n 31·56e
al-Marāghah.....26·42n 31·36e
al-Maṭarīyah.....31·11n 32·02e
al-Minyā.....28·06n 30·45e
al-Nasser.....24·32n 32·55e
al-Qanṭarah.....30·52n 32·19e
al-Qaṣr.....25·42n 28·53e
al-Quṣayr.....26·06n 34·17e
al-Qūṣīyah.....27·26n 30·49e
al-'Udaysāt.....25·35n 32·29e
al-Wāḥāt ad-Dākhilah (oasis).....25·30n 29·05e
al-Wāḥāt al-Bahrīyah (oasis).....28·15n 28·57e
al-Wāḥāt al-Farafirah (oasis).....27·15n 28·10e
al-Wāḥāt al-Khārijah (oasis).25·20n 30·35e
al-Wāsiṭah.....29·20n 31·12e
an-Nakhl.....29·55n 33·45e
Armant.....25·37n 32·32e
ar-Radīsīyah Baḥrī.....24·57n 32·53e
ar-Rāshidah.....25·35n 28·56e
Aṣfūn al-Maṭā'inah.....25·23n 32·32e
Ashmūn.....30·18n 30·58e
aṣ-Ṣaff.....29·34n 31·17e
as-Sallūm.....31·34n 25·09e
Aswān.....24·05n 32·53e
Asyūṭ.....27·11n 31·11e
ath-Thamad.....29·41n 34·18e
aṭ-Ṭūr.....28·14n 33·37e
az-Zaqāzīq.....30·35n 31·30e
Balāṭ.....25·33n 29·16e
Balṭīm.....31·33n 31·05e
Banhā.....30·28n 31·11e
Banī Mazār.....28·30n 30·48e
Banī Suwayf.....29·05n 31·05e
Bārīs.....24·40n 30·36e
Bibā.....28·55n 30·59e
Bimbān.....24·26n 32·53e
Būlāq.....25·12n 30·32e
Būr Fu'ād.....31·15n 32·19e
Burj al-'Arab....30·55n 29·32e
Būr Safājah.....26·44n 33·56e
Būr Tawfīq.....29·57n 32·34e
Cairo.....30·03n 31·15e
Daljā'.....27·39n 30·42e
Damanhūr.....31·02n 30·28e
Damietta, see Dumyāṭ
Darāw.....24·25n 32·56e
Dayrūṭ.....27·33n 30·49e
Dhahab.....28·29n 34·32e
Dishnā.....26·07n 32·28e
Disūq.....31·08n 30·39e
Dumyāṭ (Damietta).....31·25n 31·48e
el-'Alamein, see al-'Alamayn
Fuwah.....31·12n 30·33e
Gaza (Ghazzah)..31·30n 34·28e
Heliopolis (Miṣr al-Jadīdah).....30·06n 31·20e
Ḥulwan.....29·51n 31·20e
Ibshawāy.....29·22n 30·41e
Idfū.....24·58n 32·52e
Idkū.....31·18n 30·18e
Imbābah.....30·04n 31·13e
Isnā.....25·18n 32·33e
Iṭsā.....29·15n 30·48e
Jamsah.....27·38n 33·35e
Jināḥ.....25·20n 30·31e
Jirjā.....26·20n 31·53e
Kafr ad-Dawwār.31·08n 30·07e
Kafr ash-Shaykh.31·15n 30·50e
Kawm Umbū....24·28n 32·57e
Khān Yūnus.....31·21n 34·19e
Luxor.....25·41n 32·39e
Maghāghah.....28·39n 30·50e
Mallawī.....27·44n 30·50e
Manfalūṭ.....27·19n 30·58e
Marsa 'Allam....25·05n 34·56e
Maṭrūḥ (Marsa Matruh).....31·21n 27·14e
Mex, see al-Maks
Minūf.....30·28n 30·56e
Miṣr al-Jadīdah, see Heliopolis
Mīt Ghamr.....30·43n 31·16e
Mūt.....25·29n 28·59e
Nabq.....28·04n 34·25e
Naj' Ḥammādī...26·03n 32·15e
Naqādah.....25·54n 32·43e
New Nubia.....24·28n 32·54e
Nuwaybi' al-Muzayyinah....28·58n 34·39e
Port Said.....31·16n 32·18e
Qalyūb.....30·11n 31·12e
Qaṣr al-Farāfirah.27·03n 27·58e
Qift.....26·00n 32·49e
Qinā.....26·10n 32·43e
Qūṣ.....25·55n 32·45e
Rafah.....31·18n 34·15e
Rashīd.....31·24n 30·25e
Rummānah.....31·01n 32·40e
Salwā Baḥrī.....24·44n 32·56e
Samālūt.....28·18n 30·42e
Sawhāj.....26·33n 31·42e
Sharm ash Shaykh.....27·51n 34·17e
Shibīn al-Kawm..30·33n 31·01e
Sīdī 'Abd ar-Raḥmān.....30·58n 28·44e
Sīdī Barrānī.....31·36n 25·55e
Sīdī Ḥunaysh....31·10n 27·37e
Sinnūris.....29·25n 30·52e
Sīwah.....29·12n 25·31e
Suez.....29·58n 32·34e
Ṭahṭā.....26·46n 31·30e
Tallah.....28·05n 30·44e
Ṭanṭā.....30·47n 31·00e
Ṭīmā.....26·54n 31·26e
Tunaydah.....25·31n 29·21e
Wāḥat Jarbah (oasis).....29·21n 25·20e
Wāḥat Sīwah (oasis).....29·12n 25·31e
Zāwiyat Shammās.....31·31n 26·24e
Ziftā.....30·43n 31·15e

Physical features and points of interest

Abū Ballāṣ, *hill*..24·26n 27·39e
Abū Minqār, Bi'r, *well*.....26·30n 27·35e
Abū Muḥarrik, Ghurd, *dunes*...27·50n 29·40e
Abū Simbel, *ruins*.....22·22n 31·38e
Ahrāmāt al-Jīzah, *historic site*.....29·59n 31·08e
'Ajmah, Jabal al-, *mountains*.....29·12n 34·02e
al-'Araj, *marsh*...28·56n 26·24e
al-'Ayyināt, *upland*.....22·08n 32·45e
al-Baḥrayn, *physical region*..28·40n 26·32e
al-Fayyūm, *depression*.....29·17n 30·50e
'Allāqī, Wādī al-, *watercourse*.....23·07n 32·47e
al-Maghrah, *well*.30·15n 28·56e
Aqaba, Gulf of...29·00n 34·40e
'Arab, Khalīj al-, *bay*.....30·55n 29·05e
'Arabah, Wādī, *watercourse*.....29·07n 32·39e
'Arīsh, Wādī al-, *watercourses*....31·09n 33·49e
ash-Shabb, *well*..22·19n 29·46e
Aswān High Dam.....23·55n 32·50e
Asyūṭī, Wādī al-, *watercourse*.....27·11n 31·16e
Bābayn, Jabal, *mountain*.....22·38n 25·00e
Banās, Ra's, *point*.....23·54n 35·48e
Bardawīl, Sabkhat al-, *marsh*.....31·10n 33·10e
Burullus, Buḥayrat al-, *lake*.....31·30n 30·50e
Catherine, Mount, see Katrīnah, Jabal
Dab'ah, Ra's aḍ-, *point*.....31·05n 28·26e
Dāllah, 'Ayn, *well*.....27·19n 27·20e
Dibs, Bi'r, *well*...22·12n 29·32e
Dumyāṭ, Maṣabb, *river mouth*.....31·32n 31·52e
Dunqul, *oases*....23·26n 31·37e
Eastern Desert (Sharqīyah, aṣ-Ṣaḥrā' ash-).28·00n 32·00e
Elba, *mountains*.22·02n 36·26e
Farā'id, Jabal al-, *mountain*.....23·31n 35·20e
First Cataract, *rapids*.....24·01n 32·52e
Foul Bay.....23·30n 35·39e
Gharbīyah, aṣ-Ṣaḥrā' al-, see Western Desert
Gīza, Pyramids of.....29·50n 31·05e
Great Bitter Lake.30·20n 32·23e
Halfayah Pass...31·30n 25·11e
Ḥamāṭah, Jabal, *mountain*.....24·12n 35·00e
Ḥaṣa, Bi'r al-, *well*.....22·58n 35·40e
Ḥawashīyah, Wādī, *watercourse*.....28·31n 32·58e
Ḥunkurāb, Ra's, *point*.....24·34n 35·10e
Idkū, Buḥayrat, *lake*.....31·15n 30·15e
Itbāy (Red Sea Hills), *mountains*.....22·00n 35·30e
Jifjāfah, Bi'r, *well*.....30·27n 33·11e
Jiftūn, Jazā'ir, *islands*.....27·13n 33·56e
Jilf al-Kabīr, Haḍabat al-, *plateau*.....23·27n 26·00e
Jūbāl, Strait of...27·40n 33·55e
Junaynah, Ra's al-, *mountain*....29·28n 34·02e
Kanā'is, Ra's al-, *point*.....31·15n 27·51e
Katrīnah, Jabal (Mount Catherine), *mountain*.....28·31n 33·57e
Kharīt, Wādī al-, *watercourse*.....24·26n 33·03e
Kusaybah, Bi'r, *well*.....22·41n 29·55e
Libyan Desert....24·20n 25·00e
Libyan Plateau...30·30n 25·30e
Lower Egypt, *historic region*...31·00n 31·00e
Manzilah, Buḥayrat, *lake*...31·15n 32·00e
Maryūt, Buḥayrat, *lake*.....31·08n 29·54e
Mediterranean Sea.....31·50n 28·30e
Misāḥah, Bi'r, *well*.....22·12n 27·57e
Mishbih, Jabal, *mountain*.....22·38n 34·44e
Miyāh, Wādī al-, *watercourse*.....25·00n 33·23e
Muḥammad, Ra's, *point*.....27·44n 34·15e
Mūsá, Jabal, see Sinai, Mount
Nasser, Lake.....23·00n 32·50e
Natash, Wādī, *watercourse*.....24·25n 33·26e
Naṭrūn, Wādī an-, *depression*.....30·25n 30·13e
Nekhab, *ruins*....25·10n 32·48e
Nile, *river*.....30·10n 31·06e
Nuqruṣ, Jabal, *mountain*.....24·49n 34·36e
Philae, *ruins*.....24·01n 32·53e
Prince Kemal el-Dīn's Monument.....22·50n 25·50e
Qārūn, Birkat, *lake*.....29·28n 30·40e
Qattara Depression (Qaṭṭārah, Munkhafaḍ al-)..30·00n 27·30e
Qinā, Wādī, *watercourse*.....26·12n 32·44e
Ras Gharib Anchorage.....28·21n 33·06e
Rashīd, Maṣabb, *river mouth*.....31·30n 30·20e
Red Sea.....25·00n 36·00e
Red Sea Hills, see Itbāy
Ruwayān, Wādī *ar-, depression*..29·07n 30·10e
Safājah, Jazīrat, *island*.....26·45n 33·59e
Ṣaḥrā', Bi'r, *well*.....22·52n 28·37e
Saḥrā', Jabal, *mountain*.....28·00n 34·08e
Saint Catherine, Monastery of...28·29n 34·01e
Sallūm, Khalīj as-, *gulf*.....31·41n 25·21e
Shadwān, Jazīrat, *island*.....27·30n 33·59e
Shā'ib al-Banāt, Jabal, *mountain*.....26·59n 33·29e
Shalatayn, Bi'r, *well*.....23·08n 35·36e
Sharqīyah, aṣ-Ṣaḥrā' ash-, see Eastern Desert
Sharshar, Jabal, *hill*.....23·52n 30·20e
Shaykh, Sharm ash-, *bay*.....27·51n 34·17e
Sibā'ī, Jabal as-, *mountain*.....25·43n 34·09e
Sināfīr, Jazīrat, *island*.....27·55n 34·40e
Sinai, Mount (Mūsá, Jabal), *mountain*.....28·32n 33·59e
Sinai Peninsula..29·30n 34·00e
Sitrah, *oases*.....28·42n 26·54e
Siwa, *depression*.29·10n 25·40e
Suez, Gulf of.....29·00n 32·50e
Suez Canal.....29·55n 32·33e
Tarfā, Wādī at-, *watercourse*.....28·25n 30·50e
Ṭarfāwī, Bi'r, *well*.....22·55n 28·53e
Ṭawīlah, Juzur, *islands*.....27·35n 33·46e
Thebes, *ruins*....25·42n 32·37e
Tīh, Jabal at-, *upland*.....29·30n 34·00e
Tīnah, Khalīj aṭ-, *bay*.....31·08n 32·40e
Tīrān, Jazīrat, *island*.....27·56n 34·34e
Tiran, Strait of...27·58n 34·28e
Upper Egypt, *historic region*...26·00n 32·00e
Valley of the Kings, *site*.....25·45n 32·37e
Wādī Jimāl, Jazīrat, *island*...24·40n 35·10e
Western Desert (Gharbīyah, aṣ-Ṣaḥrā' al-)...26·20n 27·00e
Yu'alliq, Jabal, *mountain*.....30·22n 33·31e
Yūsuf, Bahr, *canal*.....29·18n 30·50e
Zabarjad, Jazīrat, *island*...23·37n 36·12e
Za'farānah, Bi'r, *well*.....29·07n 32·33e


MEDITERRANEAN SEA
ALEXANDRIA
Port Said
Jerusalem
Gaza
Suez
CAIRO
Al-Minyā
Asyūṭ
Sawhāj
Qinā
Luxor
Aswān
Idfū
Kawm Umbū
Isnā
Armant
Banhā
Damanhūr
Tanṭā
Al-Maḥallah al-Kubrā
Al-Manṣūrah
Dumyāṭ
Az-Zaqāzīq
Al-Ismāʿīlīyah
Shibīn al-Kawm
Minūf
Qalyūb
Imbābah
Al-Jīzah
Ḥulwān
Banī Suwayf
Al-Fayyūm
Maghāghah
Banī Mazār
Samālūt
Mallawī
Abnūb
Abū Tīj
Ṭimā
Akhmīm
Al-Manshāh
Jirjā
Naj Ḥammādī
Qift
Al-Karnak
Ad-Dayr
Kafr ash-Shaykh
Damietta
Rosetta
Buḥayrat al-Burullus
Buḥayrat al-Manzilah
Khalīj as-Sallūm
Sīdī Barrānī
Zāwiyat Shammās
Maṭrūḥ
Sīdī Ḥunaysh
Ad-Dabʿah
Al-ʿAlamayn
Al-Ḥammām
Burj al-ʿArab
Khalīj Abū Qīr
Buḥayrat Idkū
Buḥayrat Maryūṭ
Kafr ad-Dawwār
Rosetta Mouth
Masaʾab Rashīd
Masaʾab Dumyāṭ
Buḥayrat al-Manzilah
Khalīj as-Suways
Qanāṭir al-Khayrīyah
Ismailia
Great Bitter Lake
Lake Timsah
Suez Canal
Gulf of Suez
Gulf of Aqaba
Ras Muhammad
Sharm ash-Shaykh
Jazīrat Shadwān
Jazīrat Tīrān
Jazīrat Sanāfīr
Ras Ghārib
Ras Anchorage
Monastery of St. Catherine
Mount Sinai
Jabal Katrinah
Abū Zanīmah
Ra's al-Junaynah
Nuwaybaʿ al-Muṣayyinah
Ra's Sudr
Ra's al-Ḥikmah
Ra's Kanas
Foul Bay
Jazīrat Tabarjad
Jabal al-Fārāʿid
Bīr Shalaṭeyn
Jabal Ḥamāṭah
Jabal Nugrus
Jabal as Sbaʿī
Jabal Mishbih
Wārsa ʿAllam
Wādī Natash
Wādī al-Kharīt
Wādī al-Miyah
Wādī al-Ḥabīl
Wādī al-Ajal
Wādī Ḥawashiyah
Wādī at-Tarfā
Wādī al-Asyūṭī
Wādī Qena
Wādī Ḥabbān Yūsuf
Wādī ar-Ruwayān
Wādī an-Naṭrūn
Wādī Arabah
Birkat Qārūn
Lake Nasser
Dunqul
Abū Simbel
Ash-Shabb
Bīr Dibs
Bīr Misāḥah
Bīr Kusaybah
Bīr Ṭarfāwī
Bīr Shalaṭeyn
Bīr Jifjafah
Bīr Zaʿfarānah
Bīr Abū Minqār
Ayn Dallah
Qaṣr al-Farāfirah
Al-Wāḥāt al-Farāfirah
Al-Wāḥāt al-Baḥrīyah
Al-Wāḥāt ad-Dākhilah
Al-Wāḥāt al-Khārjah
Siwa
Wāḥat Siwah
Wāḥat Jarbūb
Al-ʿAraj
Sitrah
Al-Bawīṭī
Al-Ḥeyz
Al-Qaṣr
Ar-Rāshidah
Balāt
Al-Jadīdah
Mūt
Tūnaydah
Al-Khārijah
Jīnāh
Būlāq
Bāris
Abū Ballās
Al-Maḥāriq
Aṣfūn al-Mataʿinah
Dishnā
Al-Balyanā
Al-Qūṣīyah
Mantalūt
Al-Badārī
Al-Maraghah
Al-Ghurdaqah
Al-Qusayr
Al-Ballās
Nag Hammadi
Naqādah
Thebes
Nekhab
Philae
Aswan High
First Cataract
Bimbān
New Nubia
Salwā Bahrī
Ar-Radisiyah Bahrī
Jabal Shārshar
Jabal Bābayn
Jabal al-ʿUwaynāt
Prince Kemal El-Din's Monument
Ḥaḍbat al-Jilf al-Kabīr
Libyan Desert
Libyan Plateau
Qattara Depression
Ghūrd Abū Muḥarrik
Aṣ-Ṣaḥrāʾ al-Gharbīyah
Western Desert
Al-Wādī al-Jādīd
Aṣ-Ṣaḥrāʾ ash-Sharqīyah
Eastern Desert
Marsā Maṭrūḥ
Lower Egypt
Upper Egypt
Valley of the Kings
Sinai Peninsula
Al-Baḥr al-Aḥmar
Red Sea
Al-Ḥijāz
Arabia
Jordan
Israel
Egypt
Libya
The Sudan
Saudi Arabia
Pyramids of Giza
Heliopolis
Ayyāṭ
Ibshaway
Sinnūris
Al-Wāsiṭah
Biba
Al-Fashn
Al-Minyā
Tallah
Daljā
Dayrūṭ
Abū Qurqās
Mantalūt
Asyūṭ
Tahtā
Sawhāj
Jabal Yu'alliq
Jabal Thamad
Jabal Umm Lasafah
Jabal Ajmah
Jabal Sha'ib al-Banat
Jabal Safājah
Jabal Ḥamāṭah
Jabal Nugrus
Jabal Mishbih
Jabal Shārshar
Jabal al-Fārāʿid
Jabal Bābayn
Jabal al-ʿUwaynāt
Tropic of Cancer
Area occupied by Israel since June 1967.
Key to Governorates in Nile Delta:
(Shown by number on map)
1 AL-ISKANDARĪYAH (ALEXANDRIA)
2 AL-BUḤAYRAH
3 KAFR ASH-SHAYKH
4 AL-GHARBĪYAH
5 AL-MINŪFĪYAH
6 AL-QALYŪBĪYAH
7 AL-QĀHIRAH (CAIRO)
8 AL-ISMĀʿĪLĪYAH
9 ASH-SHARQĪYAH
10 AD-DAQAHLĪYAH
11 DUMYĀṬ
12 BŪR SAʿĪD (PORT SAID)
Area administered by Egypt
Area administered by Sudan
Indefinite Boundary
Administrative Boundary
© Rand McNally & Co.
A-585500-257
EGYPT
Size of symbol indicates relative size of town
Elevations in metres
0 50 100 200 300 km
0 50 100 200 mi

ern part of the Western Desert consists of Miocene limestone (from 7,000,000 to 26,000,000 years old). From 10 to 12 percent of the total area, notably the mountains of Sinai, the Red Sea, and the southwest part of the Western Desert, consists of igneous metamorphic rocks (rocks formed by the solidification of molten magma and structurally changed by heat and pressure).

The Nile silt

The silt, which constitutes the present-day cultivated land in the delta and the Nile Valley, has been carried down from the Ethiopian Highlands by the Nile's upper tributary system, consisting of the Blue Nile and the ʿAṭbarah rivers. The depth of the deposits ranges from over 30 feet in the northern delta to about 22 feet at Aswān. The White Nile, which is joined by the Blue Nile at Khartoum, in The Sudan, supplies important chemical constituents. The composition of the soil varies and is generally more sandy toward the edges of the cultivated area. A high clay content makes it difficult to work, and a concentration of sodium carbonate sometimes produces infertile black-alkali soils. In the north of the delta, salinization has produced the sterile soils of the so-called *barārī* (barren) regions.

Climate. Egypt lies within the North African desert belt; its general climatic characteristics, therefore, are low annual rainfall and a considerable seasonal and diurnal (daily) temperature range, with sunshine occurring throughout the year. In the desert, cyclones stir up sand or dust storms, called khamsin, which can occur from March to June; these are caused by tropical air from the south that moves northward as a result of the extension northeastward of the low-pressure system of the Sudan. The khamsin are accompanied by a sharp increase in temperature of from 14° to 20° F (8° to 11° C), a drop in relative humidity (often to 10 percent), and thick dust; they can reach gale force.

The climate is basically biseasonal, with winter lasting from November to March and summer from May to September, with short transitional periods intervening. The winters are cool and mild, and the summers are hot. Mean January minimum and maximum temperatures show a variation of between 51° and 65° F in Alexandria and 50° and 74° F at Aswān. Extremes of temperature can occur, and prolonged winter cold spells or summer heat waves are not uncommon.

Humidity diminishes noticeably from north to south and on the desert fringes. Along the Mediterranean coast the humidity is high throughout the year but is highest in summer. When high humidity levels coincide with high temperatures, oppressive conditions result.

The rainfall in Egypt occurs largely in the winter months; it is meagre on average but highly variable. The amount diminishes sharply southward; the annual average at Alexandria is about seven inches; Cairo has about an inch; Aswān receives only about a tenth of an inch. The Red Sea coastal plain is almost rainless.

Vegetation and animal life. In spite of the lack of rainfall, the natural vegetation of Egypt is varied. Much of the Western Desert is totally devoid of plant life of any kind, but where some form of water exists the usual desert growth of perennials and grasses is found; the coastal strip has a rich plant life in spring. The Eastern Desert receives sparse rainfall; it supports a varied vegetation that includes tamarisk, acacia, and markh (a leafless, thornless tree with bare branches and slender twigs), in addition to a great variety of thorny shrubs, small succulents, and aromatic herbs. This growth is even more striking in the wadis of the Red Sea Hills and of Sinai and in the Elba Mountains in the southeast.

Water plants

The Nile and irrigation canals and ditches support many varieties of water plants; the lotus of antiquity is to be found in drainage channels in the delta. There are more than 100 kinds of grasses, among them bamboo and *ḥalfā'* (a coarse, long grass growing near water). Robust perennial reeds such as the Spanish reed and the common reed are widely distributed in Lower Egypt, but the papyrus is now found only in botanical gardens.

The date palm, both cultivated and subspontaneous, is found throughout the delta, in the Nile Valley, and in the oases. The doum palm (an African fan palm) is identified particularly with Upper Egypt and the oases, although there are scattered examples elsewhere.

There are very few native trees. The Phoenician juniper is the only native conifer, although there are several cultivated conifer species. The acacia is widely distributed, as are eucalyptus and sycamore. The casuarina, one of the most important timber trees in the country, was introduced in the 19th century. Other foreign importations, such as jacaranda, poinciana (a tree with orange or scarlet flowers), and lebbek (a leguminous tree), have become a characteristic feature of the Egyptian landscape.

Domestic animals include buffalo, camels, donkeys, sheep, and goats, the last of which are particularly noticeable in the Egyptian countryside. The animals that figure so prominently on the ancient Egyptian friezes—hippopotamuses, giraffes, and ostriches—no longer exist; crocodiles occur only south of the Aswān High Dam. The largest wild animal is the mountain sheep, which survives in the southern fastnesses of the Western Desert. Other desert animals are the dorcas gazelle, the miniature desert fox, the mountain goat, the Egyptian hare, and two kinds of jerboa (a mouselike rodent with long hindlegs for jumping). The Egyptian jackal still exists, and the cony (a small rodent) is found in the Sinai mountains. There is a carnivorous wildcat; another carnivore is the striped Egyptian mongoose. Several varieties of lizard are found, including the large monitor. Poisonous snakes include more than one species of viper; the speckled snake is found throughout the Nile Valley and the Egyptian cobra in agricultural areas. Scorpions are common in desert regions. There are numerous species of rodents, among which can be found the powerfully built Pharaoh's rat. Many varieties of insects are to be found, including the Egyptian locust.

Bird life

Egypt is rich in bird life. Many birds pass through in large numbers on their spring and autumn migrations; in all, there are more than 200 migrating types to be seen, as well as more than 150 resident birds. The hooded crow is a familiar resident, and the black kite is a characteristic resident along the Nile Valley and in al-Fayyūm. Among the birds of prey are the lanner falcon and the kestrel. Lammergeier and golden eagles are resident in the Eastern Desert and Sinai. The sacred ibis (a long-billed wading bird) is no longer found, but the great egret and buffbacked heron are resident in the Nile Valley and al-Fayyūm, as is the hoopoe (a bird with an erectile, fanlike crest). Resident desert birds are a distinct category, numbering about 24 kinds.

The Nile contains about 190 varieties of fish, the most common being *bulṭī* (a coarse-scaled, spiny-finned fish) and the Nile perch. The lakes on the Delta coast contain mainly *būrī* (gray mullet). Lake Qārūn in the Fayyūm has been stocked with *būrī*, and Lake Nasser with *bulṭī*.

TRADITIONAL REGIONS

Physiographically, Egypt is usually divided into four major regions—the Nile Valley and Delta, the Eastern Desert, the Western Desert, and Sinai. When both physical and cultural characteristics are considered together, however, the country may be divided into six subregions—the Nile Delta; the Nile Valley from Cairo to south of Aswān; the Nubian Valley (now Lake Nasser); the Eastern Desert and the Red Sea coast; Sinai; and the Western Desert and its oases.

The Delta. The Delta, or Lower Egypt, covers an area of 9,650 square miles. It is 100 miles long from Cairo to the Mediterranean, with a coastline stretching 150 miles from Alexandria to Port Said. Much of the delta coast is taken up by the brackish lagoons of Lakes Maryūṭ, Idkū, Burullus, and Manzilah. The conversion of the Delta to perennial irrigation has made possible the raising of two or three crops a year, instead of one, over an area of 3,500,000 acres.

The Delta population

The majority of the Delta population are peasants—either landowners or labourers—living on the produce of the land. The remainder live in towns or cities, the largest of which is Cairo. As a whole, they have had greater contact with the outside world, particularly with the rest of the Middle East and Europe, than the inhabitants of the

more remote southern valley and are generally less traditional and conservative.

The Valley. The cultivated portion of the Valley between Cairo and Aswān varies from five to ten miles in width, although there are places where it narrows to a few hundred yards and others where it broadens to 14 miles. Since the completion of the Aswān High Dam in 1970, the 2,500,000-acre valley has been under perennial irrigation. The inhabitants of the Valley up to Aswān *muḥāfaẓah* (governorate) are referred to as Ṣaʿīdīs (Upper Egyptians) and are more conservative than the Delta people. In some areas women still do not appear in public without a veil, family honour is very important, and vendetta laws apply.

Until the building of the High Dam, the Aswān *muḥāfaẓah* was one of the poorest in the valley and the most remote from outside influences. There are some Ṣaʿīdīs living in the towns, but the majority of the population belong to other ethnic groups, such as the Jaʿāfirah, ʿAbābdah, or Nubians.

The Nubian Valley, or Lake Nasser. Until flooded by the waters of the High Dam to form Lake Nasser, the Nubian Valley of the Nile extended for 195 miles between the town of Aswān and The Sudan border—a narrow and picturesque gorge with a limited cultivable area. The 100,000 inhabitants were resettled, mainly in the government-built villages of New Nubia, at Kawm Umbū (Kom Ombo), north of Aswān.

Lake Nasser is being developed for its fishing and as a tourist area, and it is expected that new settlements will grow up around it.

The Eastern Desert. The Eastern Desert comprises 22 percent of the land surface of Egypt and covers an area of about 85,690 square miles. One-third of it consists of a limestone plateau stretching from the Mediterranean capital plain to a point roughly opposite Qinā on the Nile. Another third includes the sandstone plateau from Qinā southward; the remainder consists of the Red Sea Hills and the coastal plain. Moving southward from the Mediterranean, the limestone plateau consists at first of rolling hills, but approaching Qinā it breaks up into cliffs about 1,600 feet high and is deeply scored by wadis, which make the terrain very difficult to traverse. The outlets of some of the main wadis form deep bays, which contain small settlements of seminomads. The sandstone plateau farther south is also deeply indented by ravines; these, however, are relatively free from obstacles, and some are usable as routes. The Red Sea Hills run from near Suez to The Sudan border; they are not a continuous range but consist of a series of interlocking systems more or less in alignment. At their foot there is a narrow coastal plain widening southward, and, parallel to the shore, there are almost continuous coral reefs. In popular conception and usage, the Red Sea Littoral can be regarded as a subregion in itself.

The majority of the sedentary population of the Eastern Desert live in the few towns and settlements along the coast, the largest being Ras Gharib, with about 10,000 inhabitants (1966). The total population of Muḥāfaẓah al-Baḥr al-Aḥmar (or Red Sea Governorate, in the Eastern Desert) was estimated in the 1966 sample census at about 38,000. No accurate census figures are available for the nomadic population, but at that time they were believed to number around 6,000. They belong to various tribal groups, the most important being—from north to south—the Ḥuwayṭāt, Maʿāzah, ʿAbābdah, and Bishārīn. There are more true nomads in the Eastern than the Western Desert because of the greater availability of pasture and water. They live either by herding goats, sheep, and camels or by trading—mainly with mining and petroleum camps or with the fishing communities on the coast.

Nomads of the Eastern Desert

The Western Desert. The Western Desert covers an area of about 262,800 square miles. It is a rocky, low-altitude plateau rising, however, to altitudes of more than 3,300 feet (1,000 metres) on the plateau of al-Jilf al-Kabīr. From al-Jilf al-Kabīr, the plateau slopes gradually to the first of the depressions that are a characteristic feature of the Western Desert—that containing the oases of al-Khārijah and ad-Dākhilah. Farther north are the hollows containing the oases of al-Farāfirah and al-Baḥrīyah. Northwestward from the latter the plateau continues to fall toward the Sīwah and Qaṭṭārah depressions. South of the Qaṭṭārah Depression, the Western Desert is composed of great ridges of blown sand, interspersed with stony tracts. Beyond the Qaṭṭārah Depression northward, the edge of the plateau follows the Mediterranean, leaving a narrow coastal plain.

Outside the oases, the habitable areas of the Western Desert are occupied mainly by the Awlād ʿAlī tribe. Apart from small groups of camel herders in the south, the population is no longer totally nomadic. Somewhat less than half are seminomadic herdsmen; the remainder are settled and, in addition to maintaining herds of sheep and goats, pursue such activities as fruit growing, fishing, trading, and handicrafts.

The Western Desert supports a much larger population (estimated at about 124,000 in 1966) than the Eastern Desert. About 73,000 of these are seminomads. Maṭrūḥ, the provincial capital and an important summer resort, is the only real urban centre; it has a population of 11,000. Other scattered communities are found mainly near railway stations and along the northern cultivated strip.

The oases, although geographically a part of the Western Desert, are ethnically and culturally distinct. The southern oases of al-Khārijah and ad-Dākhilah are being developed as part of the New Valley project—reclamation, by exploiting underground water resources, of the five oases of al-Khārijah, ad-Dākhilah, al-Farāfirah, al-Baḥrīyah, and Sīwah.

Sinai. Sinai comprises a wedge-shaped block of territory with its base along the Mediterranean coast and its apex bounded by the Gulfs of Suez and Aqaba; it covers an area of approximately 23,000 square miles. Its southern portion consists of rugged, sharply serrated mountains. The central area of Sinai consists of two plateaus, at-Tīh and al-ʿAjmah, both deeply indented and dipping northward toward Wādī al-ʿArīsh. Toward the Mediterranean, the northward plateau slope is broken by dome-shaped hills; between them and the coast there are long, parallel lines of dunes, some of them more than 300 feet high. The most striking feature of the coast itself is the 60-mile-long salt lagoon, Lake Bardawīl.

The majority of the population are of Arab stock, now sedentary around al-ʿArīsh and in the northern coastal area but for the most part nomadic or semi-nomadic in the central plateau and the Sinai mountains. Another concentration of sedentary population is found at al-Qanṭarah ash-Sharqīyah (on the east side of the Suez Canal). The population of Sinai was estimated in 1966 at about 131,000, of which about 52,000 were nomads or seminomads.

The population of Sinai

THE LANDSCAPE UNDER HUMAN SETTLEMENT

Rural settlement. The settled Egyptian countryside, throughout the Delta and the Nile Valley to the High Dam, exhibits great homogeneity, although minor variations occur from north to south.

The typical rural settlement is a compact village surrounded by intensively cultivated fields. There are about 4,000 villages in Egypt, with populations varying from 500 to more than 10,000. Only about 2 percent of the rural population live in the smallest size villages; more than half the villages have 1,000 to 5,000 inhabitants, representing about 40 percent of the rural population. About a quarter of the villages have 5,000 to 10,000 inhabitants and contain about 40 percent of the rural population; the remainder have 10,000 or more inhabitants, representing 20 percent of the population.

The villages of the Delta and Nile Valley are basically similar in physical appearance and design, except for minor local variations in building materials, design, and decoration. The date palm, sycamore, eucalyptus, and casuerina are common features of the landscape. Until comparatively recently, the only source of drinking water was the Nile; in consequence, many of the villages are built along the banks of its canals. Some of the oldest villages are situated on mounds—a relic of the days of basin irrigation and annual flooding.

Villages of the Delta and Nile Valley

In the Delta the houses, one or two stories high, are built of mud bricks plastered with mud and straw; in the southern parts of the valley more stone is used. The houses are joined to one another in a continuous row. In a typical house, the windows consist of a few small round or square openings, barely permitting enough air or light to enter. The roofs are flat, built of layers of date leaves, with date-palm rafters; they are used to store corn and cotton stalks, as well as dung cakes used for fuel. Roofs are also a favourite sleeping place on hot summer nights. For grain storage, small cone-shaped silos of plastered mud are built on the roof and are then sealed to prevent the ravages of insects and climate.

The houses of the poorer peasants usually consist of a narrow hallway, a bedroom, and a courtyard; part of the courtyard may be used as an enclosure for farm animals. Furniture is sparse. Ovens are made of plastered mud and are built into the wall of the courtyard or inside the house. In the larger and more prosperous villages, houses are built of burnt bricks reinforced with concrete, are more spacious, and often house members of an extended family. Furniture, running water, bathroom installations, and electricity are additional signs of prosperity.

Typical features of the smaller Egyptian village, in both the Delta and the valley, are the mosque or the church, the primary school, the conical decorated pigeon cote, service buildings belonging to the government, and a few shops. Most of the people in the smaller villages are engaged in agriculture. In the larger villages, there may be some professional and semiprofessional inhabitants as well as more artisans, skilled workers, and shopkeepers. Outside the larger settlements, "combined service units" —consisting of modern buildings enclosing the social service unit, health unit, and school—are sometimes to be found, standing in striking contrast to the mud houses of the village itself.

The rural life-style

The population density of the inhabited area is such that the presence of people is obvious everywhere, even in the open countryside. In the early morning and the late afternoon, the peasants can be seen in large numbers on the roads, going to or coming from the fields with their farm animals. During the entire day the men, with their long tunics (*gallābīyahs*) tucked up around the waist, can be seen working the land with age-old implements such as the *fās* ("hoe") and *minjal* ("sickle"); occasionally a modern tractor is to be seen. In the Delta, older women in long, black robes, younger ones in more colourful cottons, and children over six years of age help with the less labourious tasks. In some parts of the valley, however, females over 16 do not work in the field, and their activities are confined to the household. They seldom appear in public except with a black-muslin headdress covering their heads and faces. Young children can be seen everywhere—an omnipresent reminder of the high birth rate.

Unless situated on a highway, villages are reached by unpaved dirt roads. Inside the villages the roads consist mainly of narrow, winding footpaths. All villages, however, have at least one motorable road.

The Western Desert oases are not compact villages but small, dispersed agglomerations surrounded by green patches of cultivation; they are often separated from each other by areas of sand. Al-Khārijah, for example, is the largest of five scattered villages. Traditionally, the houses in the oases were up to six stories high, made of packed mud, and clustered close together for defense. Modern houses are usually two stories high and further apart.

Urban settlement. Egyptian towns range in size between about 36,000 and 340,000 inhabitants. Although for census purposes they are considered urban centres, some of them are overgrown villages, containing large numbers of peasants and persons engaged in work relating to agriculture and rural enterprises. Some of the towns that have recently acquired urban status are still largely rural, although they also contain government officials, people engaged in trade and commerce, industrial workers, technicians, and professional people. One characteristic of towns and, indeed, of the larger cities is their rural fringe. Towns and cities have grown at the expense of agricultural land, with urban dwellings and apartment buildings mushrooming haphazardly among the fields. There is little evidence of town or city planning or of adherence to building regulations; often mud village houses are embraced within the confines of the city.

Urban characteristics

Buildings in towns and smaller cities are usually two-storied houses or apartment blocks four to six stories high. The better ones are lime washed, with flat roofs and numerous balconies; other houses and buildings are often of unpainted red brick and concrete.

Whereas most of the cities of Egypt do not have many distinctive features, some, such as Cairo, Alexandria, and Aswān, have special characteristics of their own. Cairo is a complex and crowded metropolis, with architecture representing its 1,000 years of history.

II. People and population

GROUPS HISTORICALLY ASSOCIATED WITH THE CONTEMPORARY COUNTRY

The use of Arabic

Linguistic groups. For nearly 13 centuries, Arabic has been the written and spoken language of Egypt. Before the Arab invasion in 639 AD, Coptic was the language of both religious and everyday life for the mass of the population; by the 12th century, however, it had been totally replaced by Arabic, continuing only as a liturgical language. Today, Arabic is the language of Egyptian Copt and Muslim alike.

The written form of the Arabic language, in grammar and syntax, has remained substantially unchanged since the 7th century. In other ways, however, the written language has changed—the style, word sequence, and phraseology used today are simpler and more flexible than in classical Arabic and are often directly derivative of English or French.

This neoclassical Arabic, which is developing out of classical or medieval Arabic, is the lingua franca shared by educated persons throughout the Arab world. Alongside it there exist the various regional dialects of Arabic, which differ widely from it as well as from each other. Within the amorphous grouping referred to as "Egyptian" colloquial, a number of separate dialects can be discerned—each fairly homogeneous but with further strata of variation within the group. One such is the dialect of the Bedouin of the Eastern Desert and of Sinai; the Bedouin of the Western Desert constitute a separate dialect group. Upper Egypt (aṣ-Ṣaʿīd, the Upland) has its own vernacular, markedly different from that of Cairo. The Cairo dialect is used, with variations, throughout the towns of the Delta; the rural people have their own vernacular. Cairene colloquial Arabic is extremely virile and expressive. Direct contact with foreigners over a long period has led to the incorporation of many loanwords. The long contact with foreigners and the existence of foreign-language schools also explains the polyglot character of Egyptian society. Most educated Egyptians are fluent in English or French or both, in addition to Arabic.

Non-Arabic languages

There are other minor linguistic groups. The Hamitic Beja (Bujah) of the southern section of the Eastern Desert use To Badawī. In Sīwah, an oasis due east of the Qaṭṭārah Depression, there are groups whose language is related to Berber. The Nubians speak a language containing both Sudanic and Hamitic features. Other minority linguistic groups also exist, notably Greek, Italian, and Armenian, although they are much smaller than they once were.

Ethnic groups. The population of riverain Egypt (comprising about 99 percent of the whole) forms a fairly homogeneous group whose dominant physical characteristics are the result of the admixture of the indigenous pre-Islāmic Hamitic–Armenoid population with Arab stock. The peasant, or fellah, is less racially mixed than the townsman. He is generally of middle stature, sturdy, dark haired, and dark eyed, with a wide-nostrilled nose and full lips. In the towns—the northern Delta towns especially—the foreign invader, Roman, Greek, Persian, crusader, and Turk, has left behind a more heterogeneous mixture. The inhabitants of the middle Nile Valley up to Aswān, the Ṣaʿīdīs, are of the same racial stock as the

inhabitants of the Delta but have darker skin and are slightly taller and sturdier. Settled communities in Aswān *muḥafaẓah* tend to be a mixture of Ṣaʿīdīs, long-settled nomads such as the Jaʿāfirah and ʿAbābdah, and Nubians.

The Nubians, although having Arab blood, have preserved racial characteristics that are non-Arab. They are tall and thin, with Caucasoid features, and are of a much darker colouring than the average Egyptian. They also differ from the Lower Egyptians in that their kinship structure goes beyond the lineage; they are divided into clans and broader segments, whereas among other Egyptians of the valley known members of the lineage are the only ones recognized as kin.

The deserts of Egypt contain nomadic, seminomadic, or sedentary nomadic groups, with distinct ethnic characteristics. Apart from a few tribal groups of non-Arab stock and the mixed urban population, the inhabitants of Sinai and the northern section of the Eastern Desert are all fairly recent immigrants from Arabia. Like the Arabian Bedouin, they are usually slightly built, hooknosed, and brown skinned. Their social organization is tribal, each group conceiving of itself as being united by a bond of blood and descended from a common ancestor. Originally tent dwellers and nomadic herdsmen, many have become seminomads or totally sedentary, as in northern Sinai.

The southern section of the Eastern Desert is inhabited by the Hamitic Beja. Although claiming Arab descent, they are of different racial stock, with oval faces, straight noses, and large eyes; they bear a distinct resemblance to the predynastic Egyptians. The Egyptian Beja are divided into two tribes—the ʿAbābdah and the Bishārīn. The ʿAbābdah occupy the Eastern Desert south of a line between Qinā and Ghurdaqah; there are also several groups settled along the Nile between Aswān and Qinā. The Bishārīn are mainly in The Sudan, although some of them are to be found in the Elba Mountain region, their traditional place of origin. Both the ʿAbābdah and Bishārīn are nomadic pastoralists, living on herds of camels, goats, and sheep.

The Saʿādī and the Mūrābiṭīn

The inhabitants of the Western Desert, outside the oases, although claiming Arab descent, are a mixture of Arab and Berber stock. They are divided into two groups, the Saʿādī and the Mūrābiṭīn. The Saʿādī regard themselves as descended from Banū Hilāl and Banū Sulaymān, the great Arab tribes that immigrated into North Africa in the 11th century. The most important and numerous of the Saʿādī group are the Awlād ʿAlī. The Mūrābiṭīn clans occupy a client status in relation to the Saʿādī and may be descendants of the original Berber inhabitants of the region. Originally herdsmen and tent dwellers, the Bedouin of the Western Desert are now either seminomadic or totally sedentary. They are not localized by clan, and members of a single group may be widely dispersed.

The original inhabitants of the oases of the Western Desert were of Berber stock, and the strain still survives in the lighter colouring of many of them. There has, however, been a considerable admixture of blood—Egyptian from the Nile Valley, Arab, Sudanese, Turkish, and, particularly in the case of al-Khārijah, black African—for this was the point of entry into Egypt of the famous caravan route from Darfur, the Darb al-Arbaʿīn.

In addition to the indigenous groups, there are in Egypt a number of small foreign ethnic groups. The 19th century saw the rapid growth of communities of unassimilated foreigners, mainly European, living in Egypt; these acquired a dominating influence over finance, industry, and government. In the 1920s, which was a peak period, the number of foreigners in Egypt was in excess of 200,000, the largest community being the Greeks, followed by the Italians, British, and French. Since Egypt's independence, the size of the foreign communities has been greatly reduced. The total number of resident foreigners, of all nationalities, is about 90,000.

Religious groups. Islām is the official religion of Egypt; about 93 percent of the population are Sunnī, or orthodox Muslims. A strong sense of piety is a characteristic of the Egyptian Muslim. Prayer is observed punctiliously, particularly public prayer in the mosques, and fasting during the month of Ramaḍān (the ninth month of the Islāmic calendar) is strictly observed. Almsgiving and pilgrimage to Mecca are, if possible, also enjoined.

The Christian population of Egypt numbers about 2,000,000; of these the majority are Copts. In language, dress, and way of life they are indistinguishable from the Muslims. Ever since it broke with the Eastern Church in the 5th century, the Coptic Church has maintained its autonomy, and its beliefs and ritual have remained basically unchanged. The Copts have been traditionally associated with certain handicrafts and trades and, above all, with accountancy, banking, commerce, and the civil service; there are, however, rural communities that are wholly Coptic, as well as mixed Coptic–Muslim villages. The Copts are most numerous in the middle Nile Valley provinces of Asyūṭ, al-Minyā, and Qinā. About one-quarter of the total Coptic population lives in Cairo.

The Copts

Other religious groups exist—Greek Orthodox, Greek Catholic, Armenian Orthodox and Catholic, Maronite, Syrian, Anglican, and Protestant. There is also a small Jewish community, which in the early 1970s numbered only several hundreds.

DEMOGRAPHY

Birth rates and death rates. The population of Egypt was estimated in 1971 at 34,130,000 by the Egyptian Census Department. In 1968, the crude birth rate was about 38 per thousand, while the death rate was about 16 per thousand, resulting in an annual rate of increase of 2.5 percent. The official figures for 1970 showed a decrease in the birth rate to 35 per 1,000 and in the death rate to about 15 per 1,000, giving a rate of increase of 2 percent.

Immigration and emigration. Immigration into Egypt is very slight and has little effect on population trends. Emigration has traditionally been slight also, but in more recent years there has been an increase in the number of persons emigrating or seeking temporary employment abroad; these have usually been from among the educated groups. The numbers, however, have never been large enough to affect population trends.

The distribution of population. The density of the population, in relation to the entire country, is 88 per square mile. When the inhabited area is considered separately, however, the density becomes very high—about 2,300 per square mile. In 1966, the Cairo density was as high as 51,028 per square mile. The population is concentrated largely in the highly fertile Nile Delta (Lower Egypt), where about 64 percent of the total population lives. Upper Egypt and the Valley, where the arable area is smaller and there are fewer large urban centres, has only 34 percent of the total population.

Population density

The majority of the population live in rural areas (about 58 percent in 1970). There has been a steady increase over the years in the proportion of population living in urban areas. The more recent increase is due, in part, to administrative changes, whereby some of the villages were converted into urban units or towns, even though they retained much of their rural character.

The Delta as a whole is more urbanized than the Valley. In 1970 it was estimated that more than 50 percent of the Delta population was urban, whereas about 73 percent of the Valley was rural. If, however, the cities of Cairo, Alexandria, and Port Said—which have *muḥafaẓah* status—are excluded from the Delta population, it can be seen that the other eight *muḥafaẓāt* are overwhelmingly rural (about 75 percent). And if the city of Giza (population over 700,000), which is an extension of the Cairo metropolis, is excluded from the valley population the remainder is almost 80 percent rural. The above figures reflect a striking feature of population distribution in Egypt—the concentration of a large proportion of the urban population in a few centres and the overwhelming dominance of the capital. Cairo alone, excluding its suburbs, contains more than 4,960,000 persons, representing over 14 percent of the total population of Egypt. Alexandria, the second-largest city, has more than 2,000,000 inhabitants.

Arab Republic of Egypt, Area and Population

	area*		population	
	sq mi	sq km	1966 census	1970 estimate
Regions				
Desert				
Governorates (*muḥāfazāt*)				
al-Baḥr al-Aḥmar	78,643†	203,685†	38,000	40,000
al-Wadi al-Jadīd	145,369†	376,505†	59,000	63,000
Maṭrūḥ	1,144	2,962	124,000	132,000
Sīnā' (Sinai)	23,442†	60,714†	131,000	140,000
Lower Egypt				
Governorates				
ad-Daqahlīyah	1,340	3,471	2,285,000	2,492,000
al-Buḥayrah	1,772	4,589	1,979,000	2,215,000
al-Gharbīyah	750	1,942	1,901,000	2,080,000
al-Minūfīyah	622	1,612	1,458,000	1,529,000
al-Qalyūbīyah	375	971	1,212,000	1,379,000
ash-Sharqīyah	1,614	4,180	2,108,000	2,344,000
Dumyāt (Damietta)	227	589	432,000	472,000
Kafr ash-Shaykh	1,327	3,437	1,118,000	1,234,000
Upper Egypt				
Governorates				
al-Fayyūm	705	1,827	935,000	1,008,000
al-Jīzah	390	1,010	1,650,000	1,934,000
al-Minyā	873	2,262	1,706,000	1,813,000
Aswān	262	679	521,000	651,000
Asyūt	591	1,530	1,418,000	1,487,000
Banī Suwayf	510	1,322	928,000	979,000
Qinā	715	1,851	1,471,000	1,559,000
Sawhāj	597	1,547	1,689,000	1,764,000
Urban				
Governorates				
al-Iskandarīyah (Alexandria)	1,035	2,679	1,801,000	2,032,000
al-Ismā'īlīyah (Ismailia)	557	1,442	345,000	395,000
al-Qāhirah (Cairo)	83	214	4,220,000	4,961,000
as-Suways (Suez)	6,888	17,840	264,000	315,000
Būr Sa'īd (Port Said)	28	72	283,000	313,000
Total Arab Republic of Egypt	22,405‡§ 386,900†	58,028‡ 1,002,000†	30,076,000	33,329,000‖

*Inhabited areas only except where footnoted †. †Total area. ‡Total of only the inhabited areas of governorates; the areas of the governorates al-Baḥr al-Aḥmar, al-Wadi al-Jadīd, and Sīnā' are excluded. §Converted area figures do not add to total given because of rounding. ‖Figures do not add to total given because of rounding.
Source: Official government figures.

Demographic trends. Until 1965 the crude birth rate in Egypt was among the highest in the world. Between 1965 and 1970, however, the rate declined somewhat. The mortality rate, on the other hand, has dropped dramatically over the years. At the turn of the century, it was estimated at 35 to 40 per 1,000, with an infant mortality rate of about 400 per 1,000. In 1945 it dropped to 27.7 per 1,000, and by 1970 it had fallen to 15 per 1,000 on average; infant mortality had fallen to 110 per 1,000. This reflects expanded medical and public-health services, the use of modern drugs and pesticides, and epidemic-control measures.

Population growth

Consistently high fertility and rapidly falling mortality rates have resulted in accelerated population growth. In 1907 the population of Egypt was a little less than 11,200,000, and it took 30 years to reach about 15,900,000 in 1937. In the following 30 years it nearly doubled, to about 30,900,000 in 1967. United Nations population projections estimate an average increase of more than 1,000,000 persons per year between 1970 and 1975. Within the decade 1970 to 1980, it is estimated that another 11,500,000 persons will be added. Clearly, such a rapid increase of population places a heavy burden on the country and impedes its efforts to promote social and economic development. (L.S.el-H./Ma.J.)

III. The national economy

A COMPARATIVE VIEW OF THE COUNTRY'S REGIONAL AND WORLDWIDE STANDING

Egypt is an overpopulated developing country. About 96 percent of its total area is desert. Lack of forests, permanent meadows, or pastures throws a heavy burden on the available arable land, which totals over 6,900,000 acres. This limited area, which sustains on average almost five persons per acre, is, however, highly fertile and is cropped more than once a year. Although at least half the population derive their livelihood from agriculture, a growing proportion of the labour force is engaged in manufacturing, and the contribution of the manufacturing sector to the domestic product is only a little less than that of agriculture—with service activities contributing the remainder. Owing to the shortage of land, underemployment of labour began to be manifest in agriculture two generations ago, and the development of nonagricultural production since then has failed to provide full employment to the increasing labour force.

The gross national product of Egypt was estimated at E£2,649,000,000 in 1968 to 1969 (E£0.43 = $1 U.S.; E£0.96 = £1 sterling, Dec. 1, 1970), representing an average per capita income of E£79.48. By the early 1970s national income had slowed down its growth from the fairly rapid development achieved in the early 1960s, largely under the impact of the increased expenditure on defense, so that per capita income was growing imperceptibly, if at all. Despite this, many aspects of the economy exhibited symptoms of relative development unusual in the developing world, such as the high yields per unit of agricultural land, the scale and progress of higher education, the rapid development of heavy industry, and the large contribution made by the manufacturing industry. A redundancy of labour remained, however, a fundamental problem.

Per capita income

THE EXTENT AND DISTRIBUTION OF RESOURCES

Mineral resources. Compared with the physical size of the country and the level of its population, the mineral resources of Egypt are scanty. The search for petroleum began earlier than elsewhere in the Middle East, but the results have been modest. In the early 1970s, discoveries of oil and gas, however, raised hopes that Egypt might become a major producer. Phosphate deposits were estimated in 1971 at 180,000,000 tons, all located in the Eastern Desert near the Red Sea, but it is anticipated that further investigation may reveal total reserves amounting to 400,000,000 tons or more. Coal deposits, amounting to some 80,000,000 tons, have been located near Suez, and the search for more coal continues, despite the low quality of the finds. Manganese deposits also are found in the Eastern Desert. Low-quality iron-ore deposits have been located near Aswān, as well as deposits of better quality in the proximity of al-Baḥrīyah oases in the Western Desert. It is anticipated that 4,000,000 tons of iron ore a year will be supplied to the Ḥulwān iron and steel complex south of Cairo.

Biological resources. The main natural resources with which Egypt is endowed are biological and centre on the Nile. There are no forests or any permanent vegetation of economic significance apart from the land under cultivation. In 1969–70 the animal population of Egypt included nearly 1,800,000 cows and oxen, about the same number of water buffalo, about 2,240,000 sheep, over 800,000 goats, and about 190,000 camels. Animal husbandry and poultry production were increasing in the early 1970s, but the total contribution of livestock and dairy production was only about one-third of all agricultural output.

Hydroelectric and other power resources. The Nile constitutes an incomparable source of energy; further sources are represented by coal, oil, and gas deposits.

SOURCES OF NATIONAL INCOME

Agriculture and fishing. Agriculture is the most important sector of the Egyptian economy, contributing 30 percent of its product, employing more than half of the labour force, and providing the country, through agricultural exports, with much of its foreign exchange. Increased pressure of population has led to an intensification of cultivation almost without parallel elsewhere. Heavy capital is invested in the form of canals, drains, dams, water pumps, and barrages; the investment of skilled labour, commercial fertilizers, and pesticides is also great. Thus, despite multiple cropping, the yields of the land are exceptionally high. Strict crop rotation—in

Investments in agriculture

addition to government controls on the allocation of area to crops, on varieties planted, on the distribution of fertilizers and pesticides, and on marketing—contributes to the high productivity of agriculture.

Since yields are already very high, increased use of manual labour meets with progressively diminishing returns. Apart from agriculture work, opportunities are scarce, and there is a constant stream of migrants from the rural areas to the cities. In 1960 it was estimated that nearly one agricultural worker out of four was redundant; it is thought that the ratio of redundancy may have increased since then, together with the increase in population.

Unlike the situation in comparable developing countries, Egyptian agriculture has an overwhelmingly commercial rather than a subsistence basis. Field crops contribute nearly three-quarters of the total value agricultural production, while the rest is made up by livestock products, fruits and vegetables, and other rural commodities. The main field crop is cotton, which occupies about a fifth of the arable land, absorbs much of the available labour, contributes about 40 percent of the value of field crops, and (in both raw and processed forms) represents up to 60 percent of the value of merchandise exports. Egypt is the world's principal producer of long-staple cotton, normally producing about 40 percent of the world crop of cotton that is one and one-eighth of an inch and longer, although total Egyptian production is only about 4 percent of all cotton produced in the world.

After cotton, which occupied 1,690,000 acres in 1970, the principal field crops are maize (corn; 1,690,000 acres), rice (about 1,200,000 acres), millet (about 520,000 acres), broad beans (314,000 acres), and wheat (1,350,000 acres). In the late 1960s, when Egypt became the world's fourth-largest rice exporter, about 600,000 tons, out of a total production of about 2,700,000 tons, were exported. Despite a considerable output, cereal production in Egypt falls short of total consumption; a substantial proportion of foreign exchange is spent annually on the import of cereals and milling products. Other important crops include sugarcane, lucerne, potatoes, and onions—the latter being normally a main export item. Many fruits are grown, and some, such as citrus, are also exported.

Reclaimed land

Between 1960–61 and 1968–69 about 896,100 acres (863,300 feddans) were reclaimed. Total land to be reclaimed as a result of the Aswān High Dam project should reach 1,300,000 acres by 1975, in addition to 700,000 acres to be converted from basin (*i.e.*, one crop a year) irrigation to perennial irrigation.

Egypt has been the scene of one of the most successful attempts at land reform. In 1952 a limit of 200 acres was imposed on individual ownership of land, and this was lowered to 100 acres in 1961 and to 50 acres in 1969. As a result, about 40 percent of land ownership has been redistributed, largely to landless peasants. The average area per owner is just under two acres.

The success of Egyptian land reform is indicated by the substantial rise of land yields between 1952 and the early 1970s. This was partly the result of several complementary measures of agrarian reform that accompanied the redistribution of the land.

The fishing industry is of relatively minor importance. The total catch reached a peak of 149,000 tons in 1964, but production has declined since the building of the High Dam. The containment of the freshwater that formerly flowed into the Mediterranean has resulted in lower production. Although some fish are being caught in Lake Nasser behind the High Dam, the catch is still small, although slowly increasing. Production of crustaceans, the bulk of which is exported, also declined in the late 1960s.

Manufacturing. The development of the manufacturing industry was handicapped by the policy of free trade imposed on Egypt from the middle of the 19th century until about 1930. Nationalism and World War II gave great impetus to the foundation of industrial projects that are largely agriculturally based and oriented toward import substitution. During the 1950s, manufacturing industry grew at the rate of about 7 percent a year, and the rate of growth was increased in the 1960s to about 10 percent a year. By 1969 manufacturing and mining amounted to over 21 percent of the gross domestic product. Large-scale manufacturing establishments, employing 25 workers or more, employed about 500,000 workers. In the late 1960s, industry employed over 11 percent of the total labour force. The largest section of Egypt's manufacturing industry in the late 1960s was represented by spinning and weaving (mainly cotton). This was followed by food processing, engineering industries, chemicals, and building-material industries. In 1969 the manufactured products included cotton yarn, cotton textiles, edible oils, refined sugar, cheese, bleached rice, cigarettes, vehicle tires and pneumatic tubes, paper, soap, commercial fertilizer, and glass products.

The development of manufacturing and industry

Emphasis has been placed on the development of heavy industry following a long-term agreement signed with the Soviet Union in 1964. Another agreement with the Soviet Union, signed in 1970, provided aid for the expansion of the iron and steel complex at Ḥulwān, the establishment of a number of power-based industries (including an aluminum plant to utilize the power generated by the High Dam), and the electrification of the countryside. An ammonium nitrate fertilizer plant was opened in 1971, based on the gases generated in the coking unit of the steel mill.

Energy. Before the completion of the Aswān High Dam power station in 1970, the bulk of Egypt's electricity was generated in thermal stations using coal or diesel fuel, but some hydroelectric power was also generated by the old Aswān Dam. A rapid rise in electricity production took place under the first five-year plan (1960 to 1965), and output reached more than 7,592,000,000 kilowatt hours in 1970, compared with only about 2,000,000 kilowatt hours in 1959. The 12 turbines of the High Dam power station have a capacity of about 2,000,000 kilowatts and are capable of producing 10,000,000,000 kilowatt hours a year; the present capacity of the thermal stations is about 40 percent of that of the High Dam. Transmission lines carry the current from Aswān to Cairo and farther north for use in urban centres and in manufacturing. A scheme, with Soviet assistance, for the electrification of all Egypt's 4,000 villages by connecting them to the High Dam power network was being implemented in the early 1970s.

Sinai contributed more than 60 percent of all the oil produced in Egypt before the 1967 Arab–Israeli War. Total Egyptian production in 1966 was about 130,000 barrels a day. By early 1971, production outside Sinai had reached 420,000 barrels a day and was being expected to reach 500,000 barrels by the end of 1972. The bulk of this oil comes from the rich Morgan fields (both onshore and offshore) in the Gulf of Suez, which are operated by a partnership of the Egyptian General Petroleum Corporation and a subsidiary of Standard Oil of Indiana. In cooperation with Phillips Petroleum Company, the Egyptians are extracting some 30,000 barrels a day from the al-ʿAlamayn field, discovered in the Western Desert in 1966. Active drilling for oil continues in both the Eastern and Western deserts, with United States, Japanese, and Soviet interests.

Natural gas

In the process of searching for oil some significant gas finds have been located. In the early 1970s a plan was put into effect using as fuel the 300,000,000 cubic feet of gas produced each day from the Morgan fields. Phillips has located wells producing some 35,000,000 cubic feet a day in the Abū Qīr area, northeast of Alexandria. A joint Egyptian–Italian gas discovery was made in the north delta near Abū Māḍī in 1970; this was being developed partly to supply a 750,000-ton-a-year fertilizer plant and partly to fuel the industrial centres in the north and northwest delta.

After the destruction by Israel of the two major refineries near Suez, Egypt has one refinery left; this is at al-Maks (Mex), near Alexandria, and it has a capacity of 65,000 barrels a day. Egypt still exports crude petroleum but has to import petroleum products. Another compli-

cating fact is that, with the Suez Canal closed, the Morgan crude oil cannot economically be shipped to Alexandria for refining, so that Egypt also has to import some crude oil. The continued closure of the Suez Canal and the development of large oil tankers have induced Egypt to plan a pipeline linking Suez to the Mediterranean near Alexandria. The proposed "Sumed" line has had the financial and technical support of a consortium of European companies led by the French. The pipeline is expected to transmit 120,000,000 tons a year; it is scheduled to be completed in 1974.

Financial services. The banking system of Egypt is centred on the Central Bank of Egypt, created in 1960 from the issue department of the National Bank of Egypt. In 1961 all banks operating in Egypt were nationalized, and their operations were concentrated in five commercial banks, in addition to the Central Bank, the government-sponsored Agricultural Credits and Co-operatives, and Industrial Bank, and three mortgage banks. Shortly after the nationalization measures of 1961, both the stock exchanges in Cairo and Alexandria and the cotton exchange in Alexandria closed down.

There are about 500 bank branches throughout Egypt; these include some 120 branches of the Agricultural Credits and Co-operatives, which is the main instrument for agricultural credit and marketing of crops, as well as for the distribution of seed, fertilizer, and pesticides.

The supply of money has, in general, followed the development of the economy; the authorities have aimed at tolerable increases in the price level, although since the 1967 war some prices have soared.

Egypt is a member of the International Monetary Fund. For a long time the international liquidity of the Egyptian economy, including the recently added Special Drawing Rights, has been depressed. In the early 1970s both internal and external debts were rising.

Imports and exports

Foreign trade. From 1965 to 1969, imports into Egypt averaged about 16 percent and exports 19 percent of the gross domestic product. Since World War II, exports have tended to fall short of imports. The trade deficit reached a peak in 1966 and was particularly sizable from 1960 to 1965 as expenditure on development rose. After the 1967 war there was a decided effort to restrict imports and stimulate exports.

Two-thirds of imports consist of raw materials and capital goods (mineral and chemical products, machinery, and transport equipment), a quarter of foodstuffs, and the remainder of other consumer goods. More than half of the exports by value consist of raw cotton, cotton yarn, and fabrics. Rice exports are also significant, and mineral and chemical products (crude oil, cement, petroleum products, phosphates) are also exported. Other exports include agricultural produce, notably onions, garlic, and citrus fruit.

Egypt's largest trading partner is the Soviet Union, followed by France, the United States, West Germany, Italy, India, Spain, East Germany, the United Kingdom, Poland, Japan, Czechoslovakia, and Yugoslavia.

With the closure of the Suez Canal and dwindling revenue from incoming tourists since the 1967 war, Egypt's foreign-trade balance in services is also in deficit, despite much reduced government expenditure abroad. The overall balance-of-payments deficit is financed by international borrowing (largely from Communist countries), transfers from Arab oil-producing countries, and changes in foreign assets and liabilities.

MANAGEMENT OF THE ECONOMY

The private sector. The economy of the country, according to the provisional constitution in force in 1971, is one based on Socialism, with the people controlling all means of production. The progress of Socialism after 1952 was initially hesitant, despite the land-reform measures, but gathered momentum after 1961, when major nationalization steps were taken in an attempt to curb the private sector and destroy the political power of capitalists. Practically all important sectors of the economy now either are public or are strictly controlled by the government. This includes large-scale industry, communications, banking and finance, the cotton trade, foreign trade as a whole, and many other sectors. Private enterprise came gradually to find its scope restricted, but some room for manoeuvre was still left in real estate and in agriculture and more recently in the export trade. Personal income is virtually limited to about E£4,800 a year and land ownership to 50 acres.

The public sector and the role of government. As the role of the private sector lessened in the 1950s and 1960s, that of the government continuously expanded. The government, when not actually in possession, regulates all important aspects of production and distribution. It imposes controls on agricultural prices, controls rent, runs the internal trade, regulates foreign travel and the use of foreign exchange, and appoints and supervises the boards of directors of corporations. The government initiates projects and allocates investment. Although the everyday running of corporations is left to the boards of directors, these receive instructions from public boards, and the chairmen of boards receive their instructions from the appropriate minister. Between 1964 and 1970, the bulk of all credit created by the banking system was for the use of the public sector.

The five-year plans

The government formulated a five-year development plan for the period 1960 to 1965, involving total investments of about E£1,513,000,000. The plan was highly successful; growth averaged 7 percent in the period, even though the economy was badly stretched and the trade deficit worsened. Annual investment allocations during the five-year period ending in the year 1969 to 1970 totalled E£1,513,000,000; while no plan was being implemented, this period was often referred to as the second five-year plan. A plan has been drawn up for the period ending 1974 to 1975, and annual investments in the plan form part of the annual budgets.

Taxation. With the majority of the population earning very low incomes, direct taxation falls on the few rich; income-tax rates are made sharply progressive in an attempt to achieve a degree of equality in the income distribution. Direct taxes on income, mostly levied on businesses, accounted for about one-fifth of governmental revenue; indirect taxes and duties provided about one-third, and the remainder came from various duties.

Trade unions and employer associations. Trade unions are closely controlled by the government. Workers obtain a share of the profits earned by corporations and elect their representatives to boards of directors; they are also heavily represented in the National Assembly. In all these activities, however, official selection works side by side with free elections. Trade unions are often vocally active in national policies but are seldom the instrument of negotiating higher wages or better work conditions. A few employers' associations are in existence, but these have little industrial power.

Economic adjustments

Contemporary economic policies. Much of the effort exerted by the government in the early 1970s was devoted to adjusting the economy to the situation resulting from the 1967 hostilities. With increased expenditure on defense, reduced allocations were made available for development; the country has also had to cope with the massive exodus of the urban population of the Suez Canal region, which moved toward the interior and away from Israeli guns. Israeli fire had destroyed vital economic targets such as the oil refineries and the industrial complex at Suez, and a metal factory near Cairo. Economic growth was slowed down roughly to the level of population growth. If a Middle East settlement were to be reached and resources freed for use in economic development, it was hoped that the policies of stimulating heavy industry, developing the oil and gas industry, and making use of the power generated by the High Dam would come to fruition.

The problem of population pressure on resources, however, was considered likely to remain for a long time, and underemployment had become endemic. (E.I.U.)

TRANSPORTATION AND COMMUNICATIONS

Almost the entire communication system is state owned. It is adequate in terms of coverage, but stresses arise

from excessive usage. The main patterns of transport flow reflect the topographical configuration of the country—that is to say, they follow the north–south course of the Nile, run along the narrow coastal plain of the Mediterranean Sea, and expand into a more complex system in the Delta.

Road network. There are about 29,400 miles of roads, of which about 5,900 miles (20 percent) are paved. Rural roads are of dried mud, usually following the lines of the irrigation canals; many of the desert roads are little more than tracks. The Cairo-to-Alexandria highway runs via Banhā, Ṭanṭā, and Damanhūr. The alternate desert road to Cairo from Alexandria has been extensively improved, and a good road links Alexandria with Libya by way of Maṭrūḥ. Paved roads exist between Cairo and al-Fayyūm and connect the various Delta and Suez Canal towns. There is a paved road from Asyūṭ to al-Khārijah and ad-Dākhilah. The coastal Red Sea route to Marsā 'Allām is poorly paved, as are the connecting sections inland.

Railways. Egypt has 2,600 miles of railway tracks, which connect Cairo with Alexandria and with the Delta and canal towns and which also run southward to Aswān and the High Dam. Branch lines connect Cairo with al-Fayyūm and Alexandria with Maṭrūḥ. A network of light railways connects the Fayyūm area and the Delta villages with the main lines; a narrow-gauge railway links al-Khārijah with the main north–south line near Naj' Ḥammādī. The main lines are used by diesel-driven trains; electric lines connect Cairo with Ḥulwān and Heliopolis.

Rivers and canals

Navigable waterways. The Suez Canal has been closed since June 1967. The Nile and its associated navigable canals provide an important means of transportation, primarily for heavy goods. There are approximately 2,000 miles of navigable waterways—about half of this total on the Nile itself, which is navigable throughout its length. The inland-waterway freight fleet consists of tugs, motorized barges, towed barges, and flat-bottomed feluccas (two- or three-masted lateen-rigged sailing ships).

Ports and shipping. In spite of its long coastline, the Arab Republic of Egypt has only three ports of any significance—Alexandria, Port Said, and Suez. Alexandria handles about four-fifths of both imports and exports, as well as the bulk of passenger traffic. It has a fine natural harbour. Port Said, at the northern entrance to the Suez Canal, lacks the berthing and loading facilities of Alexandria. Suez's main function is that of an entry port for petroleum and minerals from the Egyptian Red Sea coast and for goods from the Far East.

Air transport. Cairo is an important communication centre for world air routes. The enlarged airport at Heliopolis, with its modern terminal building, is used by major international airlines.

The national airline, Egyptian, runs external services throughout the Middle East, as well as to Europe and the Far East; it also operates a domestic air service linking Alexandria, Port Said, Luxor, Aswān, Asyūt, Maṭrūḥ, al-Minyā, the New Valley, and Ghurdaqah.

Postal, telegraph, and telephone services. There are about 7,000 post offices and postal agencies in Egypt. Both telegraph and telephone services are state owned. The number of telephones in use in 1969 totalled 365,000 —slightly fewer than one telephone for every 90 people.

IV. Administration and social conditions

THE STRUCTURE OF GOVERNMENT

The constitutional framework. Before the 1952 revolution, Egypt was a constitutional monarchy; the 1923 constitution, which followed the declaration of the end of the British protectorate, stated that Egypt was an independent sovereign Islāmic state with Arabic as its language and provided for a representative parliament. In 1953 this constitution was abolished by the new revolutionary regime and a temporary constitution substituted; political parties were dissolved. A Liberation Rally was organized in place of the parties—eventually to become, in 1962, the Arab Socialist Union (ASU). In 1953 Egypt was declared a republic.

In 1971, Egypt, Libya, and Syria agreed to the establishment of the Confederation of Arab Republics. A draft constitution was agreed to by the heads of state of each country, and was approved by referendum in each of the three member states. The capital of the confederation is Cairo.

On September 11, 1971, a new constitution for Egypt was approved by referendum. It declared the Arab Republic of Egypt to be "a democratic, socialist state" with Islām as its state religion and Arabic as its national language. It recognized three types of ownership—public, cooperative, and private. It guaranteed the equality of all Egyptians before the law and their protection against arbitrary intervention in the processes of law. It also affirmed the rights to peaceful assembly, education, health and social security, and the right to organize into associations or unions and to vote.

According to the constitution, the president of the republic is the head of state; together with the Cabinet, he constitutes the executive authority. The president must be Egyptian, born of Egyptian parents, and not less than 40 years old. The presidential term is six years, and may be extended to a second term. The president has the power to appoint and dismiss one or more vice presidents, the prime minister, ministers, and deputy ministers. The legislative body is comprised of the People's Assembly, which nominates the presidential candidate by a two-thirds majority. The candidate is then confirmed by national plebiscite.

The president is the supreme commander of the armed forces. He has the right to grant amnesty and reduce sentence. He appoints civil and military officials and can dismiss them in a manner prescribed by the law. He may call a referendum of the people on matters of supreme importance. He can, in exceptional cases and by investiture of the Assembly, issue decrees having the force of law—but only for a defined period of time.

The People's Assembly

Legislative power resides in the People's Assembly, which is composed of at least 350 elected members, representing 175 constituencies (two members per constituency). Fifty percent of the members of the Assembly must be workers and farmers. The president may appoint up to ten members. To be eligible for membership a candidate must be not less than 30 years old and must be a member of the ASU. The Assembly is elected for a five-year term. All males of 18 years of age and over are required to vote, as well as all women on the register of voters. The president of the republic convenes and closes the sessions of the People's Assembly.

The People's Assembly must ratify all laws and examine and approve the national budget. It also approves the program of each newly appointed Cabinet. Should it withdraw its confidence from the Cabinet or any of its members, they would then be required to resign. The president cannot dissolve the Assembly except under special circumstances and after a vote of approval by a people's referendum. Elections for a new Assembly must be held within 30 days.

The constitution also provides for a judiciary, independent of other authorities whose functions and authority are governed by special legislation. A Council of National Defence, presided over by the president of the republic, is responsible for matters relating to security and defense.

Local government and administration. Until 1960, government administration was highly centralized; in that year, however, the present local-government administrative system was established with the aim of promoting decentralization and greater citizen participation in local government.

Muḥāfaẓāt, districts, and villages

The 1960 Local Administration Law provides for three levels of local administration—the *muḥāfaẓāt* (governorates), the districts (or counties), and the villages. The structure combines features of both local administration and local self-government. The chief executives, for example, are appointed by the national authorities, but there are also councils at each administrative level composed of elected, appointed, and ex officio members. Although these councils exercise broad legislative powers, they are closely controlled by the central government.

The country is divided into 25 *muḥāfaẓāt;* all except Sinai apply the local-administration system. Four cities—Cairo, Alexandria, Port Said, and Suez—have *muḥāfaẓah* status. The governor is appointed and can be dismissed by the president of the republic. He is the highest executive authority in the *muḥāfaẓah*, has administrative authority over all government personnel (except judges) in his *muḥāfaẓah*, and is responsible for policy implementation and the supervision of the local branches of the various government ministries. He reports to the minister of interior on matters of security and to the minister of local administration on matters relating to local government.

The *muḥāfaẓah* council is composed of a majority of elected members—two representatives from each district or township (these must be from among the members of executive committees of the ASU), not more than 19 appointed members selected by the minister of local administration in consultation with the governor from among the active members of the ASU, and ex officio members representing various ministries such as education, health, agriculture, or social affairs. Although this has not been possible in practice, at least half the members of the council should be farmers and workers.

The town or district councils and the village councils are established on the same principles as those underlying the *muḥāfaẓah* councils. Apart from the ex officio members, there are 20 elected members on the town councils and up to five appointed members; there are 12 elected members on the village councils and two appointed members. Village councils may serve more than one village, depending on size.

The local councils have broad powers, even though limited by budget restrictions and central-government controls. They perform a wide variety of functions in education, health, public utilities, housing, agriculture, and communications; they are also responsible for promoting the cooperative movement and for implementing parts of the national plan. Local councils obtain their funds from national revenue, a tax on buildings and lands within the *muḥāfaẓah*, miscellaneous local taxes or fees, profits from public utilities and commercial enterprises, and national subsidies, grants, and loans.

The political process. The source of popular participation and representation in the political process is through the Arab Socialist Union. All Egyptians over 18 are eligible for membership, except those with criminal records or those who have been politically disfranchised. The ASU is viewed as a union representing all working classes—a category that includes the farmers, workers, soldiers, "intellectuals" (members of organized professions), and "national capitalists" (private businessmen or shopkeepers). Fifty percent of all the membership of ASU committees must consist of workers and farmers—a farmer being a person engaged in agriculture and owning ten acres or less and a worker being any person working for a living but not belonging to one of the four previously designated categories. The ASU has a pyramidal structure, with executive committees and congresses at each level of the hierarchy. At the grass-roots level, the basic units are the villages, the *shiyākhas* (subdistricts in the cities), or organizations with 200 or more employees (such as factories, universities, business firms, or ministries). Each basic unit has a congress made up of all active members, as well as an executive committee composed of ten elected members. There were about 6,900 basic units when the ASU was first formed; in the early 1970s the active members were estimated to number 5,000,000.

The Arab Socialist Union

The next level in the hierarchal organization of the ASU is the *markaz* (township) in the rural areas, the district in the cities, or the public organization consisting of more than one basic unit. The congress at this level consists of members elected from the committees of basic units and other members selected according to the regulations of the higher executive committee. The *markaz* elects an executive committee for a two-year term. At the *muḥāfaẓah* level, the congress, which is elected for a four-year term, is made up of delegates elected from the members of the *markaz*. The members of the *muḥāfaẓah* executive committees are elected for a four-year term from among the delegates to the *muḥāfaẓah* congress. The different governorate congresses form the general national congress, which elects a general committee from among its members. The latter elects, from its members, the higher executive committee.

The ASU organization is closely interwoven with that of the local administration, since members of local councils, at all levels, must be members of the ASU executive committees.

JUSTICE

The Egyptian constitution emphasizes the independent nature of the judiciary. There must be no external interference with the due processes of justice. Judges are subject to no authority other than the law; they cannot be dismissed and are disciplined in the manner prescribed by law. Judges are appointed by the state, with the prior approval of the Supreme Judicial Council under the chairmanship of the president. The Council is also responsible for the affairs of all judicial bodies; its composition and special functions are specified by law.

The courts

The court structure can be regarded as falling into two broad categories—the civil and criminal courts. The civil courts comprise summary courts, primary courts, courts of appeal, and the Court of Cassation. The criminal courts consist of petty sessional courts, courts of appeal, criminal courts, and the Court of Cassation (criminal division). Court sessions are public, except where consideration of matters of public order or decency decides otherwise. Sentence is passed in open session.

In addition, there are special courts, such as military courts and courts of public security—the latter dealing with crimes against the well-being or security of the state. The Council of State is a separate judicial body, dealing especially with administrative disputes and disciplinary actions. The Supreme Constitutional Court, held in Cairo, is responsible for judicial surveillance over the constitutionality of laws and regulations, and for the explanation of legal ordinances. Under the constitution, the Socialist Public Prosecutor is responsible for ensuring the rights of the people, the security of society and its political system, the guarding of socialist gains, and holding to the socialist way. The prosecutor is controlled by the People's Assembly.

ADMINISTRATIVE SERVICES

Educational services. At the end of the 19th century there were only three secondary and nine "higher" schools in Egypt; the educational structure continued to be based on the *kuttābs*, or Qur'ān schools. In 1916 the latter were turned into elementary schools, and in 1923 a law was passed providing free compulsory education between the ages of seven and 12. Following the revolution of 1952, educational progress already achieved was accelerated and was accompanied by both the Egyptianization and Arabicization of the educational system.

A sharp increase in the annual budgetary allocation devoted to education occurred between the end of World War II and 1970. By the later 1960s educational expenditure amounted to 17 percent of the general budget.

There are three stages of state general education—primary (six years), preparatory (three years), and secondary (three years). Primary education between the ages of six and 12 is compulsory. In 1969–70, 75 percent of those eligible entered the state primary system. Pupils who are successful in examinations have the opportunity to continue their education first at the preparatory and then at the secondary level. There are two types of secondary school, general and technical; most technical schools are either commercial, agricultural, or industrial and admit only boys. Technical education for girls takes the form of training for the electrical and electronics industries and in the manufacture of cosmetic and pharmaceutical products.

Alongside the Ministry of Education's system of general education, there is that provided by the institutes connected with al-Azhar. Instruction is given at levels

equivalent to those of the ministry, but in order to allow for greater emphasis on traditional Islāmic subjects, the duration of training is lengthened by one year at the preparatory stage and two at the secondary.

In the 1950s there were almost 300 foreign schools in Egypt, the majority of them French; many of these have since become, to varying degrees, Egyptianized. Pupils of these schools, at all levels, sit for the same state certificate examinations as those in the normal state system.

The universities

Apart from five universities—al-Azhar, Cairo, Alexandria, 'Ayn Shams, and Asyūṭ—there are also a number of other institutions for higher education.

Al-Azhar has been a teaching centre for the entire Muslim world for nearly a millennium. A large-scale modernization of the curriculum, making it comparable to those of other state universities, has been carried out since 1961. In 1970 to 1971, out of an enrollment of almost 25,000 students, al-Azhar had more than 22,000 males and 2,500 females. The student enrollment at Cairo University for the same year was 50,000 (36,000 males and 14,000 females); at Alexandria it was about 40,000 (30,000 males and 10,000 females); and at Asyūṭ it was 14,000 (11,000 males and 3,000 females).

In addition to the state university system, there is one private university—the American University in Cairo; enrollment is about 1,500, and males and females attend in approximately equal numbers.

There are about 40 institutes of higher learning, excluding institutes attached to universities or affiliated to the Ministry of Culture—such as the Institute of Dramatic Arts, the Cinema Institute, and the Institute of Ballet. These institutes specialize in commerce, industry, agriculture, the arts, physical culture, social service, domestic economy, and languages. Courses of study lead to a degree.

In spite of the rapid advance in the provision of education services, illiteracy remains a problem. Since 1964 funds have been available for a 15-year campaign to combat illiteracy.

Health and welfare services. In the early 1970s the budget of the Ministry of Health reflected a steadily increasing expenditure on public-health programs. The budget for 1968–69 alone exceeded the sum spent on health services during the 25 years prior to 1950.

The number of government health units increased from about 1,500 in 1960 to about 4,000 in 1969, and the number of beds in public hospitals almost doubled between 1952 and 1969, while the number of doctors quadrupled and the number of dentists increased 20 times over.

An important aspect of recent development has been the expansion of health facilities in the rural areas. In 1953 the government introduced what are termed combined service units; these differ from health centres in that they combine the functions of health centre, school, social-welfare unit, and agricultural extension services. In 1969 there were more than 300 of these units in operation. In addition, rural health units, of which by 1970 there were more than 1,100, further extended the health services available in rural communities; the eventual aim is to establish a total of 2,500, covering the entire countryside. Each unit is manned by a team of seven or eight persons, including one physician.

Well-trained physicians and specialists are available in large numbers in the cities and larger towns. The medical profession has prestige, and only the better qualified high school graduates are accepted in medical schools. In the decade 1960–70 more than 10,000 physicians, 1,300 dental surgeons, and 3,000 pharmacists were graduated. The paramedical level is also being given increasing attention. In 1970 almost 90 paramedical institutes provided training for medical visitors. There is also a university-level nursing school in Alexandria.

Public health campaigns

Significant efforts are being made to promote preventive medicine. Compulsory vaccination against smallpox, diphtheria, tuberculosis, and poliomyelitis is enforced for all infants during their first two years. Bilharzia (a parasitic disease) presents a most vexing health problem. It is curable, and all health centres offer treatment against it, but peasants become easily reinfected. Epidemics of malaria have been eliminated, but the disease still exists in endemic form, mainly in southern Egypt. Treatment for malaria is provided at all health centres, and the spraying of houses in mosquito-breeding areas is carried out regularly. Attention has also been given to the problem of tuberculosis; centres have been established in every *muḥāfaẓah*, and mass X-ray and immunization campaigns have been carried out.

The government has attempted to socialize medicine through such measures as the nationalization and control of pharmaceutical industries, the nationalization of hospitals run by private organizations and associations, and expanded health insurance. A health insurance law was passed in 1964; it provides for compulsory health insurance for workers in firms employing over 100 persons, as well as for all governmental and public employees.

Housing. Egypt faced a serious urban housing shortage after World War II. The situation subsequently became aggravated by increased immigration from rural to urban areas, resulting in extreme urban overcrowding. A Ministry of Housing survey (1970) showed that 55 percent of one-bedroom flats were occupied by four or more persons, 40 percent of two-bedroom flats were occupied by seven or more, and 36 percent of three-bedroom flats were occupied by ten or more persons.

Although there is considerable concern over the housing problem, the combined efforts of both public and private sectors have been unable to meet the demand. Between 1960 and 1970 about 300,000 housing units were built, representing an increase of about 18 percent in the housing available. Investment on housing in the 1960s represented between 10 and 15 percent of total national expenditure. The increase in the urban population, however, was estimated at more than 4,000,000 (over 40 percent) during the same period; *i.e.*, for every new housing unit built, 13 persons were added to the urban population. It is estimated that housing needs during the 1970s will amount to more than 1,200,000 units.

Housing construction

Between 1960 and 1970 about one-third of the units built were for those in low-income groups. The *muḥāfaẓāt* and the nine public-housing-development companies are responsible for public housing, but members of housing cooperatives can also apply for loans from the Co-operative Organization for Building Private Houses. These loans are made at a low rate of interest and are repayable within a period of five to 25 years. Private individuals can obtain loans for building at higher interest rates. Apart from public housing, large industrial firms, such as textile factories, steelworks, petroleum refineries, and chemical industries, have special housing projects for their workers.

In the rural areas, villagers build their own houses at little cost with the materials available. The government has experimented in aiding self-help projects with state loans. Ambitious rural housing projects have been carried out on newly reclaimed land; whole villages have been built with all the necessary utilities, and about 30,000 farming families from congested areas had been resettled in them by the early 1970s. It is estimated that 1,500,000 persons will eventually be resettled on the 1,250,000 acres of land reclaimed by irrigation as part of the Aswān High Dam project.

Police services. The Ministry of the Interior has direct control and supervision over all police and security functions at the *muḥāfaẓah*, district, and village levels. At the central level, the ministry is responsible for general security, immigration, naturalization, passports, civil defense, the traffic police, police training, and prisons. The General Security Department is responsible for police services, including criminal investigation, narcotics control, delinquency, smuggling, and emergency and tourist police. Every *muḥāfaẓah* has a motorized emergency police service, which is always on call and often renders services beyond normal police functions.

In the early 1970s efforts were being made to improve police services by raising the educational level of the police force and improving its technical facilities. Policemen had previously been recruited without any educational requirements other than literacy. A program

launched in 1969 aimed at the eventual replacement of the existing police force below the officer level with recruits who had received a high school education, as well as a full year of police training. Officers receive four years of university training.

SOCIAL CONDITIONS

Living standards

Wages and cost of living. The general standard of living in Egypt is rather low; the country is also subject to many of the ills common to other developing nations. In relation to the size of its population, its present economic resources are limited. Land remains its main source of natural wealth, but the amount of land is insufficient to support the population adequately.

Before 1952 the distribution of land was so disparate that less than 3 percent of the population owned more than 55 percent of the total, and only 13 percent owned any land at all. Subsequent legislation has limited land ownership, as mentioned above, to a maximum of 50 acres per person, with a 100-acre maximum for each family. Much of the land expropriated from the big landowners has been distributed to landless peasants. But the realization of the need to curb the rate of population increase led to the launching, in 1964, of a national family-planning program, to which high priority has been given.

The rural population, especially the portion represented by the landless agricultural labourer, has the lowest standard of living in the country. Industrial and urban workers enjoy, on the whole, a higher standard. The highest wages are earned in such industries as the petroleum and wine industries; many workers in industry receive additional benefits by way of social insurance and extra health and housing facilities. The salaries of professional groups are also low. Low wage levels are to some extent offset by the low cost of living, although by the early 1970s this was rising; bread, however, could still be bought cheaply, and the price of a basic staple, such as beans, was low enough to be within the reach of the poorest. Rent control also helped to keep housing costs at a reasonable level.

Health conditions. A generally low standard of health continues to be a major social problem. In spite of the great measure of progress achieved during the last few decades and the conducting of successful public-health campaigns against specific diseases, the endemic diseases that remain are a continuing threat to national well-being and progress.

The literacy rate

Educational conditions. Illiteracy—estimated at about 70 percent—also remains a serious social problem. The percentage is much higher among females than among males—about 87 percent as compared to 64 percent, respectively, in 1960. Even so, one of the most encouraging and significant features of the 1950s and 1960s has been the spread of education, particularly of women's education. In 1970 to 1971 about 30 percent of the students attending university were women. Women are no longer confined to the home; many fields of employment, including the professions and even politics, are now open to them. A further result of the expansion of education has been the emergence of an intellectual elite and the growth of a middle class, consisting of members of the professions, government officials, and businessmen.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. In spite of the many ancient civilizations with which it has come into contact, Egypt unquestionably belongs to a sociocultural tradition that is Arabic and Islāmic. This tradition remains a constantly operative factor in determining Egyptian views both of self and of the world.

The story of the cultural development of modern Egypt is, in essence, that of the response of this traditional belief system to the intrusion into it, at first by conquest and later by the penetration of ideas, of the alien and materially superior civilization of the West. The response covered a broad spectrum—from the rejection of new ideas and reversion to traditionalism, through self-examination and reform, to an uncritical acceptance of new concepts and the values that went with them. The end result has been the emergence of an unselfconscious cultural identity, which has assimilated much that is new while remaining distinctively Egyptian. The process is to be seen at work in all branches of contemporary culture.

The state of the arts. The impact of the West is one of the recurring themes in the modern Egyptian novel, as in Tawfīq al-Ḥakīm's *ʿUsfūr min ash-Sharq* ("The Bird from the East") and Yaḥyā Ḥaqqī's novella *Qindīl Umm Hāshim* ("The Lamp of Umm Hāshim"). A further theme is that of the Egyptian countryside—romantically handled at first, as in Muḥammad Ḥusayn Haykal's *Zaynab*, and later realistically, as in ʿAbd al-Raḥmān ash-Sharqāwī's *al-Arḍ* ("The Land") and *al-Fallāḥ* ("The Peasant") and in Yūsuf Idrīs's *al-Ḥarām* ("The Forbidden"). A Dickensian capacity to catch the colour of life among the urban poor is a characteristic quality of the early and middle work of Egypt's greatest contemporary novelist, Najīb Maḥfūẓ, notably in *Zuqqāq al-Midaqq* ("The Alley of al-Midaqq").

The modern theatre in Egypt is a European importation—the first Arabic-speaking plays were performed in 1870. Two dramatists, both born at the turn of the century, have dominated its development—Maḥmūd Taymūr and Tawfīq al-Ḥakīm. The latter, a versatile and cerebral playwright, reflects in his themes not only the development of the modern theatre but also, in embryo, the cultural and social history of modern Egypt. The changes in Egyptian society during the 1950s and 1960s are also reflected in the themes adopted by younger contemporary dramatists.

Film making in Egypt

There is a relatively long tradition of film making in Egypt going back to World War I, but it was the founding of Miṣr Studios in 1934 that stimulated the growth of the Arabic-speaking cinema. Nowadays, Egyptian films are shown to audiences throughout the Arab world and are also distributed in Asian and African countries. The industry is both privately and state owned—there are almost a score of private film-production companies, as well as the Ministry of Culture's Egyptian General Cinema Corporation.

Contemporary Egyptian music embraces indigenous folk music, traditional Arabic music, and Western-style music. The revival of traditional Arabic music, both vocal and instrumental, owes much to state sponsorship. Popular Arabic music consists of a blend of classical Arabic music, folk songs, and Western music. Muḥammad ʿAbd al-Wahhāb has been one of the leading figures in the development of this genre, as both composer and singer. Of the same generation is Umm Kulthūm, the leading vocalist not only of Egypt but also of the whole Arab world for nearly half a century. Western-style music has been a familiar component in Egyptian musical culture since the 19th century. Pioneers such as Yūsuf Greiss (1905–61) and Abū Bakr Khayrat (1910–63) succeeded in giving a national colouring to their Western-style compositions.

A return to folklore as a source of inspiration for the arts is a generalized phenomenon in modern Egyptian culture. It has resulted in a revived interest in traditional crafts, in the collection of folk music, and the maintaining, with government sponsorship, of two folk-dance ensembles—the Riḍa Troupe and the National Folk Dance Ensemble. In the plastic arts, the highly original use of local themes is particularly striking. An active school of Egyptian painting and sculpture has emerged.

Cultural institutions. The oldest learned academy in Egypt, the Institut d'Égypte, was founded in 1859, but its antecedents go back to the institute established by Napoleon in 1798. The Academy of the Arabic Language, founded in 1932 and presided over by the veteran man of letters Taha Husein, is, in terms of prestige and influence, one of the most important cultural institutions in Egypt. Linked to the Ministry of Culture, it enjoys a large measure of autonomy, guaranteed by its own charter. Also attached to the Ministry of Culture is the Higher Council for Arts, Letters, and the Social Sciences. Intended as a consultative body on cultural matters, the Higher Council is also a means of channelling state patronage.

The learned societies

There are some 30 learned societies catering to a wide variety of subject interests—such as the physical and natural sciences, medicine, agriculture, the humanities, and social sciences. Increased government concern with research, especially in science and technology, was reflected in the establishment of the Atomic Energy Commission and the National Research Centre in 1956. In addition to the latter, there are almost 20 specialized research institutes in the country.

Most of the learned societies and research institutes have library collections of their own. In addition to large collections at the universities, the municipalities of Alexandria, al-Manṣūrah, and Ṭanṭā maintain libraries. There is also a central public library in each *muḥāfaẓah*, with branches in small towns, and service points in the villages. The Ministry of Culture is responsible for the Egyptian National Library (Dār al-Kutub), the National Archives, and the Public Libraries Administration. The Egyptian National Library, which has a large collection, is also a centre for the collection and preservation of manuscripts.

The Ministry of Culture is also responsible, through its department of antiquities, for the Egyptian Museum, the Coptic Museum, the Museum of Islāmic Art, and other institutions, including such fine-arts museums as the Mukhtār Museum, the Nājī Museum, and the Museum of Modern Art.

Press and broadcasting. Ownership of major newspapers is vested in the ASU. Control is exercised through four large publishing houses—al-Ahrām, Akhbār al-Yawn, al-Jumhūrīyah (al-Gumhouriya), and al-Hilāl; among them they publish all the large-circulation dailies, weeklies, and periodicals.

Among the daily newspapers, *al-Ahrām* remains the most authoritative. Other dailies (with estimated circulations of about 250,000) are *Akhbār al-Yawm* and *al-Jumhūrīyah*. The Jumhūrīyah group also publishes the evening paper *al-Masā'*, as well as the foreign-language dailies printed in English, French, Greek, and Armenian. In addition to widely read weeklies and cultural monthlies, there are many publications catering to particular professional interests.

Radio and television

Radio and television services are grouped together to form one corporation under the aegis of the Ministry of Information. In 1971, radio broadcasting consisted of 11 programs, with a total average transmission time of more than 180 hours per day. In addition to programs intended for other Arabic-speaking countries, such as *The Voice of the Arabs*, the Overseas Programme transmits in 32 foreign languages—representing, next to the United States, the most extensive range of foreign-language broadcasting in the world.

There is one television station, and television plays a role in relation to national policies. It has been used, for example, in the literacy campaign, as well as to advocate agricultural improvements. In 1970 there were about 560,000 television sets in use.

Prospects for the future. In the early 1970s Egypt found itself playing an increasingly important role in Africa and in the developing world. It belongs, however, both culturally and politically, in the Arab Middle East. As the most populous and technologically advanced of the Arab states, it has inevitably played a major role in Arab cultural and political affairs.

Like other Asian and African countries, Egypt has faced many problems of development since World War II, such as the limitations imposed by an agriculture-based economy, a high rate of population growth, a low per capita income, and a lack of resources. Cheap electric power from the Aswān High Dam and recent discoveries of petroleum deposits will undoubtedly facilitate economic growth and the process of industrialization. The nature and pace of Egypt's future development will be determined to a large extent, however, by the course of events in the Middle East region as a whole.

BIBLIOGRAPHY

General works: *Egypte*, new ed., "Les Guides Bleus Series" (1971), the best of the popular guides; GEORGE L. HARRIS (ed.), *Egypt* (1957), provides general coverage, also important as a source of information on agriculture and social change; RAYE R. PLATT and BAHY MOHAMMED HEFNY, *Egypt: A Compendium* (1958), provides general background and information on detailed aspects of contemporary Egypt.

The landscape: W.B. FISCHER, *The Middle East: A Physical, Social and Regional Geography*, 4th ed. (1961), contains basic geographical information; RUSHDI SAID, *The Geology of Egypt* (1962), an authoritative specialized study; CAIRO UNIVERSITY, *Bulletin of the Faculty of Science*, no. 17, 28, 30, 36, *The Flora of Egypt*, as yet incomplete, is the standard work on the subject; RICHARD MEINERTZHAGEN, *Nicoll's Birds of Egypt*, 2 vol. (1930), a primary source copiously illustrated; G.A. BOULENGER, *Zoology of Egypt: The Fishes of the Nile* (1907, reprinted 1965), the authoritative and amply illustrated standard work; HENRY HABIB AYROUT, *Fellahs* (1942; Eng. trans., *The Fallaheen*, 1945), although written before 1952, contains observations on the customs, dress, and psychology of the Egyptian peasant that are still valid; HAMED AMMAR, *Growing Up in an Egyptian Village* (1954), the best and fullest account of village life in Egypt; SAYED MAREI, *Agrarian Reforms in Egypt* (1957), general work on agricultural reform; TOM LITTLE, *High Dam at Aswan* (1965), on the building of the High Dam and its effect on landscape and people.

Peoples and population: ABBAS M. AMMAR, *The Peoples of Sharqiya*, 2 vol. (1944), a physical anthropologist's description of the inhabitants of the eastern Delta; WILLIAM H. WORRELL, *A Short Account of the Copts* (1945), a concise study of the indigenous Christian population of Egypt; ANWAR G. CHEJNE, *The Arabic Language: Its Role in History* (1969), background on the role of classical Arabic and the dichotomy between it and the various dialects. For information on European and other ethnic minorities, see Platt and Hefni, already cited, and the works of Issawi and Little listed below.

Administration and social conditions: LUTHER GULICK and JAMES K. POLLOCK, *Government Reorganization in the U.A.R.* (1962), a report on the reorganization of the machinery of government; HAROLD F. ALDERFER and M. FATHALLA EL KHATIB, *Local Government in the United Arab Republic*, 2nd ed. (1964), a descriptive outline; MORROE BERGER, *Bureaucracy and Society in Modern Egypt* (1957), one of the few English-language statements on the subject; P.J. VATIKIOTIS, *The Egyptian Army in Politics: Pattern for New Nations?* (1961), a provocative interpretation of the role of the army in politics and its transformation into a political elite; AMIR BOKTOR, *The Development and Expansion of Education in the United Arab Republic* (1963), an important general survey; BAYARD DODGE, *Al-Azhar: A Millennium of Muslim Learning* (1961), deals with the special role of al-Azhar in Islamic education; HARRY HOPKINS, *Egypt, the Crucible* (1969), a graphic, if personalized, description of contemporary Egyptian politics and society; TOM LITTLE, *Modern Egypt* (1967), a fuller and improved version of the author's earlier work (*Egypt*), with greater emphasis on social and political structures; PETER MANSFIELD, *Nasser's Egypt*, 2nd ed. (1969), a clear and orderly description of political, economic, and social changes in Egypt since 1952; CHARLES ISSAWI, *Egypt in Revolution* (1963), an informative study of economic and social conditions in modern Egypt, replacing the author's earlier books—*Egypt: An Economic and Social Analysis* and *Egypt at Mid-Century*.

The cultural scene: MUSTAFA HABIB (ed.), *Cultural Life in the United Arab Republic* (1968), a collection of articles by Egyptian specialists; MAGDI WAHBA, *Cultural Policy in the United Arab Republic* (in press), an authoritative statement on inter-ministerial responsibility for cultural affairs, with particular emphasis on the Ministry of Culture; ALBERT HOURANI, *Arabic Thought in the Liberal Age* (1962), deals with the interaction of Western and indigenous culture in its historical context; JACOB M. LANDAU, *Studies in the Arab Theatre and Cinema* (1958), on the origins of the theatre and cinema; ABD AL-MONEM ISMAIL, *Drama and Society in Contemporary Egypt* (1967), an analysis of political and social themes in contemporary Egyptian drama, with synopses of plots; ABDEL-AZIZ ABDEL-MEGUID, *The Modern Arabic Short Story* (n.d.), deals mainly with the genesis and early development of the modern Arabic short story; HAMDI SAKKUT, *The Egyptian Novel and Its Main Trends from 1913 to 1952* (1971), a much-needed study on the growth of the Egyptian novel.

Egyptian government publications: CENTRAL AGENCY FOR PUBLIC MOBILIZATION AND STATISTICS, *Statistical Abstract of the United Arab Republic 1951/52–1968/1969* (1970); *Statistical Handbook* (1968); *The Increase of Population in the United Arab Republic and Its Impact on Development* (1969); MINISTRY OF CULTURE, *Cinema in the United Arab Republic* (1970); INFORMATION DEPARTMENT, *The Charter* (1962); and *The Constitution, March 25, 1964* (1964).

(L.S.el-H./Ma.J.)

Egypt, History of

Egypt was the home of one of the principal civilizations of the ancient Near East and, like Mesopotamia, of one of the very earliest urban and literate societies. Its culture had an important influence on both ancient Israel and ancient Greece, which in turn helped to form the civilization of the modern West. Egypt also provided Africa with its earliest civilization and may well have had considerable influence on the development of other African cultures. Long known to readers of the Bible and of the Greek historian Herodotus, ancient Egypt first became an object of scholarly research with the development of archaeology in the early 19th century.

This article is divided into the following sections:

I. Egypt to the end of the 17th dynasty
  Introduction to Egyptian civilization
    The nature of ancient Egyptian civilization
    Sources and chronology
    Egyptian archaeology
  The Predynastic and Early Dynastic periods
    The Predynastic Period (to *c.* 3100 BC)
    The Early Dynastic Period (*c.* 3100–*c.* 2686 BC)
  The Old Kingdom and the First Intermediate Period
    The Old Kingdom (*c.* 2686–*c.* 2160 BC)
    The First Intermediate Period (*c.* 2160–*c.* 2040 BC)
  The Middle Kingdom and the Second Intermediate Period
    The Middle Kingdom (*c.* 2040–1786 BC)
    The Second Intermediate Period (1786–1567 BC)
II. From the beginning of the 18th dynasty to *c.* 330 BC
  The New Kingdom
    The 18th dynasty
    The 19th and 20th dynasties
    Society and culture in the New Kingdom
  Egypt from 1085 BC to the Macedonian invasion
    Weakness and foreign domination to the mid-7th century
    Independence and cultural revival under the 26th dynasty
    Egypt under Achaemenid rule
III. Hellenistic and Roman Egypt
  Macedonian and Ptolemaic Egypt (332–30 BC)
    The Macedonian conquest
    The Ptolemaic dynasty
    Government and civilization of Hellenistic Egypt
  Roman Egypt (30 BC–AD 642)
    Roman administration
    The end of Roman rule
IV. Egypt from *c.* 630 to *c.* 1800
  From the Islāmic conquest to 1250
    Period of Arab and Turkish governors (639–868)
    The Ṭūlūnid dynasty (868–905)
    The Ikhshīdid dynasty (935–969)
    The Fāṭimid dynasty (969–1171)
    The Ayyūbid dynasty (1171–1250)
  The Mamlūk and Ottoman periods (1250–1800)
    The Mamlūk dynasty (1250–1517)
    The Ottomans (1517–1798)
V. Egypt since *c.* 1800
  From the French to the British occupation (1798–1882)
    The French occupation and its consequences (1798–1805)
    Muḥammad ʿAlī and his successors (1805–82)
  The period of British domination (1882–1952)
    The British occupation and the Protectorate (1882–1922)
    The Kingdom of Egypt (1922–52)
  The revolution and the republic
    Struggle for power
    Egypt and the U.A.R. (1955–*c.* 1971)

The special character evident in the civilization of ancient Egypt over a period of 3,000 years developed very rapidly at the time when the country first achieved unity. This great event happened in about 3100 BC, and while some of the seeds of Egyptian culture had sprouted before this time, it is proper to regard the start of the 1st dynasty as the virtual beginning of Egypt, as the country and its civilization are now generally envisaged.

Perhaps the first and most important quality that typified this civilization was continuity. In every aspect of Egyptian life, in every manifestation of its culture, a deep conservatism can be observed. This clinging to the traditions and ways of earlier generations was the particular strength of the Egyptians. It can also be regarded as a weakness; but for a relatively primitive culture there was more to be gained than lost in attachment to the past. Regularity was a built-in characteristic of Egypt; life in the Nile Valley was determined to a great extent by the behaviour of the river itself. The pattern of inundation and falling water, of high Nile and low Nile, established the Egyptian year and controlled the lives of the Egyptian farmers—and most Egyptians were tied to a life on the land—from birth to death, from century to century. On the regular behaviour of the Nile rested the prosperity, the very continuity, of the land. The three seasons of the Egyptian year were even named after the land conditions produced by the river; *akhet*, the "inundation"; *peret*, the season when the land emerged from the flood; and *shomu*, the time when water was short. When the Nile behaved as expected, which most commonly was the case, life went on as normal; when the flood failed or was excessive, disaster followed.

Importance of the Nile

I. Egypt to the end of the 17th dynasty

INTRODUCTION TO EGYPTIAN CIVILIZATION

The nature of ancient Egyptian civilization. That Egypt was a land in which civilization was nurtured was the direct result of the Nile; the Nile made Egypt, and the valley of the Nile was Egypt. Only that part of the country susceptible to flooding could provide a livelihood for the population. The vast areas of desert to the east and west of the green, fertile valley were barren, inhospitable wastes; it was possible to find interesting things in these deserts—fine, hard stones, precious metals, and some game—but such places could hardly be considered parts of the country of Egypt. In a sense Egypt was an island, cut off from other habitable lands—to the north by the sea and in all other directions by deserts. Being thus isolated it was both protected from invasion and insulated from external influences.

Isolation and self-sufficiency of ancient Egypt

This isolation contributed in no small degree toward the conservative, inward-looking character of life in Egypt. For the Egyptian there could be no doubt that his country was quite distinct; it was a universe on its own, unrelated to other lands abroad. To be away from Egypt was to be divorced from reality; being an Egyptian meant living in Egypt, worshipping Egyptian gods (who had nothing to do with the world outside), dying and, above all, being buried in Egypt. The special features of Egyptian life reflected this egocentric view, which treated all things Egyptian as being of their kind unique. Egyptian techniques in working materials were in many respects different from those practiced elsewhere in the ancient world; artistic methods and conventions, established at the very beginning of the Dynastic Period, were quite peculiar to Egypt; even Egyptian writing, the hieroglyphic script, was a medium of communication developed specifically for the use of the tongue spoken in Egypt, and it was incapable of adaptation to the requirements of other languages. In techniques, art, and writing, the methods developed in the earliest times remained, in general terms, satisfactory for the needs of the Egyptian people and over the centuries required only the modifications resulting from natural development within a fairly closed culture. This self-sufficiency, which amounted almost to a sort of cultural stagnation, is well demonstrated in the matter of writing. When it became necessary to correspond diplomatically with other countries, the cuneiform script and the Akkadian language were used; this script, which was employed for many different ancient languages, displayed a versatility that Egyptian hieroglyphs never acquired.

One unfortunate result of the isolated but comfortable character of Egyptian life was that people outside regarded Egypt as specially favoured and therefore as a desirable prize to be conquered. For many centuries the greedy aspirations of foreigners amounted to no great threat because of the difficulties of invading Egypt across broad desert tracts or the sea; furthermore, the likely invaders were never of sufficient power and strength to undertake full-scale warfare. There was, however, noth-

ing to prevent small groups of nomads and traders from entering Egypt and settling down, particularly in the sparsely populated region of the Delta. The threat posed by such groups eventually became very real; but at first there seemed to be no cause for alarm, and the rulers of Egypt were rarely obliged to engage in serious defensive warfare. In general Egypt was a peaceful country; its people were pacific—they made poor soldiers and always despised the military life, preferring the peaceful lives of the farmer and the scribe.

From the New Kingdom (1567–1085 BC) onward Egypt became more and more involved in the affairs of its foreign neighbours and, in consequence, in warlike activities; but the peaceful character of its people was not much modified by foreign adventures, nor was the basically pastoral nature of life in Egypt much changed. The stability and regularity of most Egyptians' existence imparted to Egyptian civilization the notable continuity from which, in turn, sprang the best and most enduring achievements of ancient Egypt. In the earliest dynasties, after the period when Upper and Lower Egypt were united into a single realm, the country was organized administratively and agriculturally in such a way that thereafter, even in periods of little central control, the ideal of the well-regulated life remained strong in the memories and desires of men.

Despotism of Pharaohs

The king and his chief ministers controlled life in all its aspects, but there is much to suggest that the pharaonic despotism was inspired by noble ideals of responsibility toward the ordinary Egyptians. Below the central authority, administration of the royal will was exercised provincially through the nomes, or districts, into which Upper and Lower Egypt were divided. These nomes, probably based on primitive tribal areas, were retained for administrative purposes; and at nome level was organized most of the governmental activity that directly affected the life of the ordinary Egyptians: the administration of justice; the control of waterways, canals, and irrigation; the carrying out of local works; the raising of levies for the annual corvée after the inundation or for occasional military service. Inscriptional evidence declares the high moral intention of local administration; literary evidence suggests that on the whole the system operated in a just manner. In a society as hierarchically organized as that of ancient Egypt, the interests of the lowly were subordinated to those of the great, but in this respect Egypt differed little from most ancient societies. Within the limits set by such a system, the people of Egypt lived a reasonably happy and comfortable life.

In matters of agriculture, very high standards were achieved; by careful control of water at all seasons, the best was made of the fertile soil of the country. Good crops were raised, often as many as three in succession throughout the year; fine herds of cattle were reared, pastured mostly in the Delta. Splendid results were attained in breeding, and many experiments were carried out in extending the range of domestic animals and fowl. The actual techniques of farming and husbandry were simple throughout the three millennia of dynastic history, but they were finely judged to obtain good results. The peculiar nature of Egypt's cultivable land, with the unusual character of the country's climate, required a form of agriculture that differed in many respects from that practiced in less favoured lands. The Egyptian peasant did not have to toil for long hours, scratching a living from a barren soil; the relative ease of cultivation enabled the cultivator to concentrate his attention on more advanced horticulture. Egyptian domestic gardening was of a high standard and well-adapted for a culture in which much emphasis was laid on the excellence of garden produce and on viticulture and its subsidiary, wine making.

Beyond the fields—in the workshops of the great estates, the temples, and the towns—other crafts were practiced and, for their contemporary purposes, perfected. On the practical level, the Egyptians were outstanding among the peoples of antiquity. The manipulation of difficult materials seemed to present few problems to their craftsmen. In the techniques of stoneworking they particularly excelled. The Nile Valley served as a vast quarry from north to south, providing the building stone that Egyptian architects employed so magnificently; the Eastern Desert was the source of other fine stones used for vessels and statuary, of semiprecious stones, and of metals. The only important material lacking was timber, but adequate supplies could usually be obtained from Lebanon and Syria. Working with simple but efficient tools, the Egyptian craftsmen mastered all available materials, while Egyptian artists wrought masterpieces in many media. Sculpture was undoubtedly the field in which the greatest results were achieved; but Egyptian painters, both when working with carvers in producing low-relief scenes and in the execution of flat wall paintings in gouache, displayed a complete mastery of their art.

The Egyptians were a practical people, and they reveal through the products of their arts and crafts their particular genius. In classical times these early Egyptians were also credited by the Greeks with great knowledge and wisdom; but the evidence provided by Egyptian writings does not support this Greek opinion. It is probable that Greek travellers in Egypt, impressed by the grandeur and antiquity of the monuments of the land and misled by the accounts of past ages given to them by their priestly guides, grossly misinterpreted the evidence and jumped to unwarranted conclusions. Unlike the Greeks, the Egyptians were not philosophically inclined, intellectually inquisitive, or prone to theorizing. They tended to accept the world as they saw it and to make use of its advantages without looking too deeply into the properties of its parts. They were good engineers and builders but not good mathematicians, limiting their interest in calculation to the solution of practical problems. As astronomers they knew the sky well, enjoying exceptionally good conditions for observing; but they did not use the knowledge derived from astronomical observation to correct their civil calendar of 365 days. In medicine, for which they subsequently achieved a great reputation, they were again skilled practitioners, making use of the knowledge derived from the dissection of bodies in the course of mummification; but in the treatment of illnesses they never abandoned the magic of primitive times, and their pharmacopoeia was a strange amalgam of efficacious drugs and magical ingredients of no therapeutic value.

Classical misinterpretations of early Egypt

Nevertheless Egyptian civilization, through its continuity, stability, and age, was properly revered in antiquity; while the country itself was desired, the achievements of its people were greatly honoured. Much of this reputation may have been slenderly based, but its reality cannot be denied.

Sources and chronology. Strictly speaking, Egyptian history can only be written for those periods from which written records survive. Earlier periods, which elsewhere would be called prehistoric, are in Egypt's case termed predynastic. This term derives from the habit of calling the historic period in Egypt the Dynastic Period—the time covered between the 1st dynasty and the Ptolemaic Period. Dynasty was the word used by the annalist Manetho, an Egyptian priestly writer who lived in the 3rd century BC, to designate each of the 30 "houses," or groups, into which he organized the successive kings of Egypt. He began with the union of Egypt, making the uniter, whom he named Menes, the founder of the 1st dynasty. The 30th dynasty ended in 343 BC with the death of the last native Egyptian pharaoh, Nectanebo II. Manetho's original text has not survived; but abstracts, or epitomes, preserved in later writers, give the gist of his account, usually including the lengths of the various reigns. There are serious drawbacks to using these epitomes of Manetho as sources of Egyptian history—they are very summary, imperfect, and in some cases display a corrupt text; the Greek forms of names are not easy to translate into Egyptian forms; the dynasties themselves do not always conform to the sequence of succession and of changes of houses as they can be observed from more ancient sources. Nevertheless, Manetho does provide a serviceable framework that is sufficiently in accord

The annals of Manetho

with Egyptian history as it is now known; and it is more convenient to retain the use of his dynasties than to evolve a new system.

Before the discovery of the Rosetta Stone and the consequent decipherment of hieroglyphs, Manetho's account formed the principal source for the history of Egypt. His framework has subsequently been strengthened and modified by older lists of kings, most of which date from the 19th dynasty, approximately 1,000 years before the time of Manetho. The most important list, called the Turin Canon, is preserved on a papyrus in the Egyptian Museum in Turin. It, like the abstracts of Manetho, contains lists of names with lengths of reigns, the lists being interrupted from time to time by totals of groups of reigns, which correspond in part to the dynasties of Manetho. Unfortunately the document is badly damaged and otherwise rendered obscure by the inclusion of ephemeral rulers whose existence cannot be confirmed from other sources. But for the best known periods of Egyptian history before the 19th dynasty there is considerable agreement, at least in the matter of the number of reigns, between the Turin Canon and Manetho.

Lists of kings were also set up in temples and tombs. Two examples from Abydos and one from Ṣaqqārah were inscribed during the 19th dynasty, and one was inscribed at Karnak in the reign of Thutmose III of the 18th dynasty. The weakness of these lists as historical records is that they include only the names of kings deemed worthy of honour; many modest and certain unpopular rulers are wholly overlooked—expunged from the record.

The above records only provide the barest outline of Egyptian history. For further information the historian has to make use of the inscriptional, documentary, and other remains recovered in the course of excavation or still preserved on standing monuments. Thus the historical inscriptions set up by kings provide much detailed information about campaigns and conquests, but the reliability of any text has to be estimated carefully in the context of its time. Certain kings were particularly boastful in their claims and often incorporated into accounts of their conquests details culled from the inscriptions of their predecessors. To establish a reliable account of any period of Egyptian history requires much patient weighing of evidence. Even so, and in spite of the wealth of material remains, the narrative of Egyptian history is insubstantial and full of gaps. This inadequacy becomes more marked the farther back the story goes. For the earliest dynasties the account would be very slight but for the royal annals preserved on a large stone slab, the Palermo Stone, only a fragment of which has survived. This stone when complete contained a yearly record of the reigns of all Egyptian kings from late predynastic times down to the 5th dynasty. Each year is characterized by one or more significant events, some of which are of little historical interest (*e.g.*, the making of a cult statue) but others of which provide important information such as the dispatch of foreign expeditions.

The Palermo Stone

Because evidence of the sequence of events and reigns in Egyptian history is incomplete, dating can only be approximate. Some control of the chronology derived from the king lists is provided by the records of an annual natural phenomenon known as the heliacal rising of Sirius, the Dog Star (also known as Sothis). The rising of this star on the horizon occurs in the morning just before sunrise on approximately the 19th of July in the Julian calendar; it was observed particularly by the ancient Egyptians, possibly because it happened each year at about the time the Nile began to rise rapidly. The day of the rising was thus held to be the true New Year's Day. The Egyptian civil year of 365 days, being astronomically one-quarter of a day too short, gradually fell out of step with the astronomical year and thus with the rising of Sirius; the two years coincided again only after a period of about 1,460 years. The Roman scholar Censorinus, writing in the 3rd century AD, recorded a coincidence of the heliacal rising of Sirius with New Year's Day in the Egyptian civil calendar in AD 139. By calculation it follows that similar coincidences happened in about 1317 and 2773 BC. More precise calculation corrects these dates to 1314 and 2770 BC. Three records of heliacal risings of Sirius between these two dates provide reasonably fixed dates of 1872 for Year 7 of Sesostris III of the 12th dynasty, 1539 for Year 9 of Amenhotep (Amenophis) I of the 18th dynasty, and 1469 for an unspecified year of Tuthmose III of the same dynasty. A further reasonably well-established date, based on lunar observations and other evidence, is 1304 for the first year of the reign of Ramses II of the 19th dynasty.

Dates earlier than 1872 BC can be calculated with some degree of certainty back to the 11th dynasty, but before that time the paucity of detailed information allows only reasonable estimates to be made. Happily, a fragment of the Turin Canon preserves the total of years from the start of the 1st dynasty to the end of the 8th dynasty. This total, together with a reasonable estimate for the shadowy period before the 11th dynasty, yields a date for the beginning of the 1st dynasty of about 3100 BC. Some confirmation for the dates accepted below is provided by radiocarbon dating. Recent examination of material of the period from the 1st dynasty down to the 12th dynasty has produced carbon-14 dates that in general confirm the commonly agreed chronology. The margin of error inherent in such dates as at present calculated does not permit a precise check to be made on accepted historical dates.

Egyptian archaeology. The modern study of ancient Egypt dates from the French military expedition to Egypt in 1798, led by Napoleon Bonaparte. Before that time, visitors to the Nile Valley were few and opportunities for working there were practically nonexistent. There had been interest in the country and its monuments among European scholars during the Middle Ages and the Renaissance, but the difficulties of access thwarted serious study. During the 18th century it became increasingly easy for travellers to visit Egypt, and the illustrated publications of travels by the English bishop Richard Pococke in 1743 and by the Dane F.L. Norden in 1755 greatly stimulated interest. The French expedition, therefore, came at the right moment for scholarly exploitation, and it was important for another reason: in the course of military operations the Rosetta Stone was discovered. Thus were two keys provided—one opening the door to Egypt, the other the door to the ancient Egyptian language and thence to knowledge of ancient Egyptian history and civilization. Simultaneously, a commission of scholars brought from France by Napoleon was hard at work recording, among the natural and other wonders of Egypt, many of its ancient monuments. The work of these scholars was ultimately published in the *Description de l'Égypte*, itself a monument to scholarly industry.

Beginning of modern study

Fieldwork in Egypt has always taken two forms, excavation and recording. The former has, by the nature of its discoveries, ever been more dramatic; but the latter over the years has provided much of the material from which knowledge of ancient Egypt has been derived. Early excavation tended to be a search for buried treasure or at least for fine museum objects. Visitors to the country and consular officials, obtaining official permission by persuasion or bribery, dug where they wished and destroyed in the process as much as they discovered. The French scholar Auguste Mariette put a stop to this pillaging. In 1858 he founded, and became first director of, the Antiquities Service of Egypt, which thenceforth controlled archaeological work in Egypt. Mariette, himself a great excavator, introduced a new element of responsibility into excavation; but it was an Englishman, Sir Flinders Petrie, who (beginning in 1880) first worked in a truly scientific manner in this field.

Petrie laid great emphasis on control, observation, attention to detail, and, especially, on the publication of results. He often worked too quickly, but his intentions were always scholarly. His working life in Egypt extended from the last years of Mariette until well after World War I. He not only established a school of excavators who carried on and extended his work but he also incidentally created a climate of trust that enabled excavators of other countries to work profitably in Egypt.

20th-century expeditions

The first three decades of the 20th century saw many expeditions from American museums at work, a great increase on the part of French excavators working principally through the French Institute of Oriental Archaeology in Cairo, and missions from many other countries engaged in fieldwork among the seemingly inexhaustible ancient sites of Egypt. George Andrew Reisner, who conducted excavations for a number of American institutions—particularly the Museum of Fine Arts in Boston—established standards of recording and methods of classification that went far beyond what most workers in the field were prepared to emulate; Reisner's excavations at Giza, which continued up to the outbreak of World War II, were particularly important. Of all the discoveries in Egyptian archaeology, that of the tomb of Tutankhamen by Howard Carter in 1922 was undoubtedly the most remarkable. Pertinacity and a shrewd judgment of slender evidence inspired Carter to continue digging when most excavators would have abandoned work; great manual skill, immense patience, and superlative draftsmanship rendered his work on the contents of the tomb beyond praise.

The recording, or epigraphic, side of Egyptian archaeology, after the initial example provided by the Napoleonic Commission, achieved high standards of scholarly accuracy with the work of a Prussian expedition led by Karl Richard Lepsius in 1842–45. The volumes of Lepsius' *Denkmäler aus Ägypten und Äthiopien* contain drawings of reliefs and inscriptions on standing monuments throughout Egypt and the Sudan, which are still much consulted. In view of the date when the expedition worked, the results are remarkably accurate; furthermore, much was recorded that has since perished.

The copying of scenes and texts for accurate reproduction in publication was not, however, pursued as a positive ideal until the late 19th century. Norman de Garis Davies, working for the Egypt Exploration Fund of London and later for the Metropolitan Museum of Art, New York, combined accuracy, artistry, and scholarship in his many publications of tombs. His wife, Nina, has never been surpassed as a facsimile copyist of Egyptian paintings. The finest epigraphic publications since World War I have been produced by the Egyptian Expedition of the Oriental Institute of the University of Chicago, whose great folio volumes of Theban temples represent the ultimate achievement in accuracy.

After World War II, Egyptian archaeology became dominated by the campaign, sponsored by the United Nations, for saving ancient monuments and sites in Egyptian and Sudanese Nubia. The construction of a great new dam a few miles to the south of Aswān threatened with inundation all archaeological sites southward for 300 miles (480 kilometres). From 1960 the archaeologists of many countries cooperated in the vast scheme to record and, where necessary, to excavate every identifiable ancient site threatened with submersion. All monuments still standing in the doomed area were to be recorded and, again where possible, moved to higher ground. For eight years the campaign continued; many dozens of sites were excavated, and many ancient buildings were moved—the most extraordinary operation was the removal and reconstruction of the two great rock-cut temples of Abu Simbel to a site well above the expected level of the new lake. The whole campaign was a remarkable example of international cooperation; Egyptian and Sudanese Nubia, though now lost beneath the waters of Lake Nasser, is probably the best recorded area in the world from an archaeological point of view.

Reconstruction of Abu Simbel

THE PREDYNASTIC AND EARLY DYNASTIC PERIODS

The Predynastic Period (to c. 3100 BC). The favourable environmental conditions existing in the Nile Valley enabled man to achieve advances in his ways of life at a very early date. In remote antiquity the character of the culture enjoyed by man in Egypt, as far as it can be judged from the few surviving tools and weapons, seems to have differed very little from that of other inhabitants of northern Africa. In late Paleolithic times, as a result of climatic changes in this region, vast tracts of land became inhospitable to man and he was obliged to retreat from the open plains, which now became deserts, to the coastal belt along the Mediterranean and to the Nile Valley.

Remains of occupation sites of the Neolithic Period (before *c.* 4000 BC) have been found in a few places in Egypt, but the majority of settlements have probably been lost beneath the deposits of Nile silt in the river valley. On the edge of the valley, however, traces of villages have been found in Middle and Lower Egypt and in the Fayyūm, northwest of Hawāra. The inhabitants of the Fayyūm settlements, which are probably to be dated to the mid-5th millennium BC, were fishers and hunters; they also grew grain and stored it in silos made of woven straw. They made tools from stone, flint, and bone; pottery was crude and very simple in shape; ornaments were made from shells and coloured pebbles, roughly shaped and pierced. Settlers such as those at Merimda Banī Salāmah on the southwestern edge of the Delta enjoyed a more advanced culture than did the Fayyūm people; a varied series of artifacts has been recovered from their villages, including pottery of finer quality than that of the Fayyūm, and many objects modelled in clay. The dwellings at Merimda Banī Salāmah were oval- or horseshoe-shaped light huts of wickerwork posts, probably covered with skins; here the dead were buried among their dwellings. In this respect they reveal similarities to the inhabitants of Neolithic sites in western Asia; in other respects, particularly in their skilled working of stone, they may be compared with their African contemporaries to the west and south.

These early peoples of Middle and Lower Egypt were not metal users. For the first traces of copper implements and ornaments it is necessary to turn to Upper Egypt. The settlements and cemeteries of the earliest identifiable copper-using Egyptians were discovered at al-Badāri, Mostagedda, and Matmar. The Egyptians who lived in this region took their first steps in metalworking early in the 4th millennium BC. A distinct advance can be noticed between the stages of culture found in the remains at Dayr Tāsā and those of al-Badāri. The so-called Tasian culture, primitive and metal-free, is probably to be taken as an early phase of the Badarian culture. Badarian sites are characterized by the high quality of the pottery produced—particularly a very thin, polished red ware, partly blackened and often decorated with a ripple pattern—and also by the presence of simple ornaments and a few small implements made of copper, and of steatite beads covered with a primitive glaze. The glazing of steatite was characteristically Egyptian, and its appearance during Badarian times heralds the existence of a culture that can confidently be termed Egyptian. Badarian is taken to be the earliest stage of Egyptian predynastic civilization and, as such, distinct from earlier cultures in the Nile Valley.

The earliest metal users

In manners of burial the Badarians differed from the Neolithic people of Middle and Lower Egypt because they possessed cemeteries set apart from their dwellings. In this practice they were followed by the later predynastic cultures of the Nile Valley, which are now generally named in sequence Naqādah I and Naqādah II, after the modern village of Naqādah in Upper Egypt, where Petrie first excavated and identified the predynastic cultures of Egypt. (Petrie himself called the first phase Amratian, after the village of al-ʿĀmrah near Abydos, and he called the second phase Gerzean, after al-Girza, a village much farther north.) In temporal terms it is difficult to make a strict division between the two Naqādah cultures; and there is some reason to believe that Naqādah II represents a more highly developed form of civilization that originated in Lower Egypt and gradually impinged upon and ultimately replaced the more simple civilization of Naqādah I, which was basically Upper Egyptian. This does not necessarily imply an actual conquest of south by north, but it perhaps rather indicates a cultural domination of the south by the north. In general terms Naqādah I displays a distinct advance on Badarian; techniques for the working of basic raw materials are much improved, and the products of such simple

The Naqādah cultures

crafts as pottery and stoneworking are better designed and more finely finished. Slate palettes for the grinding of cosmetics, which had first appeared in Badarian times, became more common and took more interesting shapes, although they cannot be compared with those in the form of animals, birds, and fish found in Naqādah II burials. Stone maceheads and well-worked flint javelin heads and arrowheads testify to the development of warlike activities, although it cannot be determined if they were produced for offensive or defensive purposes.

In Upper Egypt the graves of the Naqādah I people were relatively simple, shallow holes, in which the bodies were buried in contracted position with the faces turned to the west. They were commonly wrapped in skins or placed in wicker baskets. In addition to the pottery and palettes, grave goods included bone cosmetic objects like combs and armlets, beads of pottery and stone, and occasionally rough figurines of men and women, the significance of which remains in much doubt. The actual bodies reveal that the predynastic inhabitants of Upper Egypt were short and slimly built and had delicate features.

The inhabitants of Lower Egypt had broader heads, and they were buried in more highly developed graves, often carefully walled with wickerwork or matting and sometimes having more than one chamber. The goods found in the Lower Egyptian tombs, and also in Upper Egyptian graves of somewhat later date, are superior in craftsmanship and design. The pottery, made of a light baking clay, is often painted with representations of boats and human figures; stone vessels are cut from all kinds of hard stone mostly found in the Eastern Desert—*e.g.*, schist, diorite, volcanic ash, porphyritic rock, granite, basalt. A greater use of copper and of glazed steatite suggests a better technology, while the presence of non-Egyptian materials like lapis lazuli postulates contacts with western Asia.

Early political organization

Politically, little is known about Egypt at this remote time. Motifs in pottery decoration and marks on other objects suggest that a form of nome organization already existed and that there were probably two principal centres of power—Naqādah in the south, with the god Seth as tutelary deity, and Behdet in the Delta, with the falcon Horus as god. It is probable that around these two centres confederations of nomes were built up that in time took on the form of "kingdoms," the rulers of which assumed crowns—that of the south being white and that of the north being red (the two crowns that were subsequently joined into the double crown of the king of Upper and Lower Egypt). As time advanced toward the final clash that resulted in the union of these two lands, new capitals were established—at Hierakonpolis in Upper Egypt and at Buto in the Delta. In spite of the apparent cultural supremacy of the north, victory eventually went to the south; and the unification of Upper and Lower Egypt was achieved by the man known to later tradition as King Menes. So far no monument of that time has been discovered certainly bearing the name Menes, although one small table of ivory from Naqādah may bear it. Most scholars now hold that Menes should be identified with Narmer, the king whose great ceremonial palette from Hierakonpolis shows him in an indubitably victorious role, triumphing over northern enemies.

The Early Dynastic Period (c. 3100–c. 2686 BC). Historically speaking, the union of Upper and Lower Egypt by Menes was the most important event in the chronicle of the country's existence; and so it must have been regarded in antiquity. This was the moment when Egypt's history began for the ancient Egyptians, the time from which the succession of dynasties progressed. Unification was the political necessity of the land; unification remained the sought-after political ideal, prompting the actions of the Theban kinglets who much later brought to an end the anarchic First and Second Intermediate Periods. The first unification, therefore, was felt to be so important that it liberated great creative forces in the land. Overnight, as it were, Egyptian civilization achieved maturity; and in its physical manifestations it revealed the characteristics that were to be constantly visible for the next three millennia. It became a literate civilization; from the early 1st dynasty, hieroglyphs were used for secular and religious purposes and for official and private texts. It is true that none of the surviving texts from this period are very long and that only the simplest of ideas and facts are conveyed, but in the use and form of the unique hieroglyphic script they exhibit all the characteristics of long texts of later periods. These same early texts also provide valuable evidence of political organization, of the titles of officials, and of names of people and places. Inasmuch as most of the inscribed material comes from burials, it is not surprising that the resulting picture is somewhat one-sided. Nevertheless, such information as can be retrieved shows a country fully organized, with an efficient bureaucracy, well-established practices in accounting, recording, and general administration, and a social hierarchy similar in kind to that known for the Old Kingdom.

Texts from the early period

The king was, without a doubt, supreme; and his high officials and nobles attended him in death as in life, to judge from the placing of tombs in royal cemeteries. There is a debate about the location of the principal royal cemetery at this time—the choice lying between Abydos in the south and Ṣaqqārah in the north. At both places great tombs have been found that can be assigned to kings of the first two dynasties; but the evidence is at present insufficient to allow a certain decision to be made between the two. These and other tombs of the period, although badly robbed, have yielded enough fragments of funerary equipment to show that the products of Early Dynastic culture in Egypt were of fine quality. Much use was made of rare materials like gold, ivory, and lapis lazuli, some of which must have been imported, while the great timbers used in the tombs provide positive evidence of trading expeditions to Syria. The imposing brick superstructures of the tombs, far superior to anything built in predynastic times, equally reveal the amazing step forward in achievement made during the 1st and 2nd dynasties. The early architects, perhaps inspired by western Asiatic prototypes, were complete masters of building monumentally in mudbrick; and they even began to experiment in the use of stone for such architectural elements as lintels, steps, pavements, and portcullis blocks.

The 1st dynasty. In adopting a panelled design for the outside walls of the earliest tombs, the architects may have been reproducing in smaller scale the great walls that surrounded the new capital built by Menes in the valley near Ṣaqqārah, the city later called Memphis. The capital was placed strategically at the junction of Upper and Lower Egypt, and Menes is credited with having diverted the Nile in order to obtain the best situation for it. This political act demonstrates the intention of the earliest ruler of a united Egypt to govern the country as a whole and not to treat Lower Egypt simply as a conquered territory of Upper Egypt. But Upper Egyptian traditions, particularly religious ones, were evidently given greater prominence; the king was the embodiment of the Horus falcon, the great deity of Hierakonpolis (the old capital of the southern realm).

Enterprises of the new regime

Under the new regime great enterprises were undertaken of a kind impossible in earlier times. Menes conducted operations in the Eastern Desert; Aha, his successor, waged a campaign against the Nubians; the third king, Djer, sent expeditions to the east, probably to Sinai, and to the south as far as the Second Cataract; Den, the fifth king of the 1st dynasty, seems not only to have sent expeditions abroad but also to have done much to regulate the administration of the country. This last king of the 1st dynasty also apparently introduced the double crown as the royal diadem of the joined kingdoms.

The 2nd dynasty. The circumstances that led to a change of dynasty are not known, but shreds of evidence suggest that the last reigns of the 1st dynasty were troubled by internal dissension and probably that the royal line had lost the initial strength of purpose, which had helped it in the beginning to succeed so brilliantly. The first king of the 2nd dynasty gained power in about 2890 BC. His name was Hetepsekhemui, meaning "the

two powers are at peace," possibly a reference to the end of the struggle between the two parts of the country. Little is known about the first reigns of the new dynasty. The fourth king, Peribsen, may have been an usurper; he chose to call himself a Seth-king, not a Horus-king, which may indicate that he represented an opposing faction. Khasekhem, who is usually taken as Peribsen's successor, may have been in part his contemporary and rival for power; he was a Horus-king, and his monuments suggest that he was not only an energetic and warlike monarch but also a very successful one. The last king of the dynasty, Khasekhemui, designated himself a Horus- and a Seth-king, and it is possible that he is not to be distinguished from Khasekhem; it has been suggested that when Khasekhem overcame Peribsen he changed his name and designation as acts of reconciliation.

In spite of its generally shadowy character, there is enough evidence to prove that the Early Dynastic Period was in almost all respects the great formative epoch for Egypt. Government, administration, artistic convention, technical skills, writing, and building all were nurtured and developed at this time; they were to come to fruition in the Old Kingdom.

THE OLD KINGDOM AND THE FIRST INTERMEDIATE PERIOD

The Old Kingdom (c. 2686–c. 2160 BC). Under the rule of Khasekhemui, Egypt again made great strides forward in cultural development; and there is good reason to regard this king as the founder of the Old Kingdom and as the man who prepared the ground for the triumph of his successor Djoser.

The 3rd dynasty. The reign of Djoser (Zoser), which because of its monuments stands out particularly as a high-water mark of achievement, assumed a similar importance for the ancient Egyptians; in the king list of the Turin Canon his name is selected for the rare honour of being written in red. Yet he was not, apparently, the first king of the 3rd dynasty. After Khasekhemui it seems that Sanakht initiated the new dynasty. But of Sanakht's reign the only important event now known is an expedition to Sinai, possibly the first of many similar expeditions during to Old Kingdom, to search for the much prized turquoises of the region. Djoser, the immediate successor of Sanakht, also sent an expedition to Sinai; and it was he who probably established Egypt's southern boundary at the First Cataract on the Nile.

Djoser's funerary monument

But Djoser is especially important because he can be seen as the first Egyptian king who so organized and developed the resources and administration of the country that he was able to embark on the construction of a mighty funerary monument. In demonstrating his power in this way he established a new tradition, one by which his successors for many hundreds of years would demonstrate their own claims to glory. Building was the Egyptian medium of expression, above all. Whether it was Djoser or his principal official, Imhotep, who deserves the credit for this innovation cannot be decided.

Tradition assigns to Imhotep the credit for the achievements of Djoser's reign. According to contemporary records, Imhotep was a royal official, bearing important secular and priestly titles; by tradition he was much more—a great seer, physician, thinker, and, above all, an architect. Of these posthumous elements in his reputation only the last can be certainly attached to the Imhotep of Djoser's reign. Manetho made him responsible for the invention of building with hewn stone, and there seems no reason to doubt that as a senior official of Djoser's court, perhaps the most important nonroyal person of the time, he was in fact the builder of the Step Pyramid and the wonderful complex of buildings that surround it.

Simply, the Step Pyramid was to be Djoser's tomb; it consisted of a stepped structure built of the local limestone of Ṣaqqārah and faced with fine limestone from Turā across the Nile. It was set in an enclosure surrounded by a fine limestone wall, within which, in addition, were many smaller structures. The whole may have reproduced in stone part of the palace complex at Memphis. It was a noble conception, wonderfully executed—the first great stone building. It demonstrated what could be accomplished by fine craftsmanship allied with good organization; it was a permanent testimony to Djoser's greatness. Imhotep may have been the architect, but Djoser provided the raison d'être.

Precedent for Djoser's successors

The long-term effect of this partnership was fundamental to the development of Egypt and, in particular, to the growth of the idea of divine royalty. The Step Pyramid provided positive proof of what the king could do, and it demonstrated what he symbolized. It served as a notable precedent for Djoser's successors.

Sekhemkhet, Djoser's immediate successor, is known principally from the remains of his unfinished step pyramid, also at Ṣaqqārah, which was only discovered in 1951 and partially excavated in the years following. It is evident that Sekhemkhet embarked on a funerary monument that would at least equal that of Djoser. Parts of the great enclosure wall have been uncovered, and in scale it rivals Djoser's wall. The masonry, however, is less well prepared than that of the earlier monument. From this last fact it may perhaps be deduced that less care was taken in constructing Sekhemkhet's monument, that it was built in a great hurry, or that standards of construction had been relaxed. The complete absence of other evidence about Sekhemkhet's reign, apart from the record of a visit to Sinai, prevents any certain reason to be given for the difference in quality between the two monuments. It is, however, strange that not only Sekhemkhet but also the other kings of the 3rd dynasty left so few positive memorials of their reigns. The lack of material remains is, of course, to a great extent due to the passage of time and the natural causes of destruction. Nevertheless, the general impression remains that after the high peak of achievement attained during the reign of Djoser, a time of moderate development followed in which the rapid successes early in the dynasty were no more than consolidated.

Some indication of the general prosperity of the country during the reigns of Sekhemkhet, Khaba, and Huni—who are the best identified successors of Djoser—can be gained from the size and surviving remains of their funerary monuments and from the mastaba tombs of the nobles of the dynasty at Ṣaqqārah. The tombs in particular demonstrate the abilities of the craftsmen of the time and reveal something of the great men who served the kings of Egypt; these men were at least allowed some independence and the freedom to express their pride in the construction of grandiose monuments. Their status was to be reduced in the years to come, even as the power of the king was to increase.

The 4th dynasty. With the accession of Snefru (Sneferu) in about 2613 BC came a change in dynasty and the beginning of an epoch, the surviving manifestations of which still strongly influence the general impression of ancient Egyptian culture. Snefru was the builder of the first true pyramid, and his two vast structures at Dahshūr usher in the Pyramid Age as it is now generally understood. The change in design from step pyramid to true pyramid may represent a change in religious beliefs; it certainly represents a change in building methods. The design of the true pyramid was successfully achieved only after much experimentation, in which great engineering problems were solved. Huni, the last king of the 3rd dynasty, was possibly the first to attempt a true pyramid; and the present ruined state of his construction at Maydūm testifies to his long-term lack of success. Snefru's two pyramids reveal developments: one shows how the angle of inclination had to be changed in the course of construction, probably to avoid a structural collapse; the other possesses an unusually shallow angle of inclination in which the builders were clearly aiming at safety in construction. The ideal pyramid was eventually built by Snefru's successor, Khufu, and the first—the Great Pyramid at Giza—was the finest and most successful. But in the size of its base the Khufu pyramid was only a little larger than Snefru's northern pyramid at Dahshūr.

The beginning of the Pyramid Age

Achievements in mighty constructions during the first reign of the 4th dynasty show in striking manner how

capable the Egyptian administration was at this time. It may be assumed that the relatively quiet years of the 3rd dynasty left the country in a sound economic state and that Snefru took control of a land whose resources were capable of supporting great undertakings. Literary tradition of the Middle Kingdom suggests that, in his time, Snefru was a well-loved king with whom his more autocratic son and successor, Khufu, was subsequently compared, to the latter's detriment. It was probably Khufu's misfortune that he followed up too successfully the policies and practices instituted in his father's reign. Snefru had been an active king. From the laconic entries on the Palermo Stone it is known that he sent expeditions southward to Nubia and westward into Libya. Traces of a settlement in which copper smelting was carried on have been discovered in the region of the Second Cataract in Nubia; scraps of inscribed material show that it was occupied as early as the 4th dynasty, and its foundation may have been due to Snefru. Inscriptions in the Wadi Maghara in Sinai also record expeditions sent there by Snefru; so important were his activities there that his name remained associated with enterprises in Sinai for hundreds of years. He also sent expeditions by sea to Byblos to obtain timber for shipbuilding and other great works; balks of Syrian timber have been found inside his pyramids at Dahshūr.

In his great projects Snefru relied to a considerable extent on the support of relatives; most of the high officials in the land seem to have been princes. In the formal inscriptions set up in the chapels of their tombs at Maydūm and Dahshūr, they list the honours and titles that they held in their lifetimes. The enumerations are long and varied, and they cover all branches of activity; some titles indicate administrative duties, others are concerned with religious or military matters. These princely officials covered all departments of state; and while it is unlikely that they actively exercised the duties implied by their titles, there can be no doubt that they were ultimately responsible for all that happened. Their responsibility stemmed directly from the king, and they were answerable directly to him. This closely centralized form of government provided the great strength of the kings of the 4th dynasty. All the important officials of the time, whether their duties lay with national or provincial matters, fulfilled their offices at the royal residence, where they could be observed and prevented from becoming too powerful. Throughout Egyptian history the periods in which the country was strongest were those in which the ruler held the reins firmly in his own hands, curbing the activities of his provincial officers. The physical nature of Egypt, a land that for 500 miles (800 kilometres) had length but scarcely any breadth, rendered it difficult to rule, even under the most favourable conditions. A king who delegated control in the nomes to powerful local officials courted disaster.

Centralized government

Some idea of the artistic quality, refined taste, and superb workmanship of the products of the 4th dynasty can be gained from the residual funerary furniture and ornaments found in the second tomb of Hetepheres, the mother of Khufu, which that king had transferred to Giza from Dahshūr after the burial had been found plundered. Otherwise the majority of tombs of this time, whether royal or private, have yielded little, apart from fine sculpture and inscriptions of a funerary character. The great cemeteries at Giza, laid out around the pyramid of Khufu, reflect vividly the way everything circulated around the king. These cemeteries, with neat rows of mastaba tombs of the princes, notables, and high officials of the court, provided the means by which the king's servants were kept under control even in death. All the evidence suggests that it was in Khufu's reign that Egypt was most strictly centralized. The king, the living Horus, was Egypt; everyone else was subservient; everything planned was for the glory of the king; everything completely reflected this glory.

The Great Pyramid at Giza

The Great Pyramid is in every way the embodiment of this autocratic idea. It also offers the complete demonstration of what can be done when all resources are controlled by one man and directed solely for his purposes. In a society like that of Egypt at this remote time, the Great Pyramid could only have been built if national resources were concentrated and used with great efficiency. It is now thought that the construction of this pyramid was not effected by slaves but by the ordinary peasants of Egypt, employed in the months of the inundation when work on the land was impossible. Comparatively small gangs of quarrymen cut stone throughout the year; and it was moved by the bulk of the labour force during the inundation, when transport by water over the flooded land was relatively easy. Nevertheless, the transporting, preparing, and laying of about 2,300,000 blocks of an average weight of 2½ tons was no small task of organization; it could not have been completed without first-class planning. The completed monument on the desert plateau at Giza testifies to the efficiency of the whole operation.

Only the smallest of indications suggest that important public works other than funerary monuments were carried out during the 4th dynasty. But inscriptions in some tombs show that much effort was directed into the agricultural activities of the land—the improving of irrigation, of crops, of cattle; the farming of great estates in the Delta. Such unspectacular activities did not form part of the public image of the dynasty's kings, but they undoubtedly contributed immensely to the prosperity of the land and to the happiness of its inhabitants.

Khufu's successor, Redjedef, may have been an usurper of the throne. Little is known of his reign except that he probably supervised the funeral of Khufu and built his own pyramid at Abū Ruwaysh, a few miles to the north of Giza. After about nine years he was succeeded by Khafre, a younger son of Khufu, who was a king more in the tradition established for the dynasty. Khafre's pyramid at Giza is only a little smaller than that of Khufu; very considerable remains have survived, both of the funerary temple beside the pyramid and of the so-called valley temple on the edge of the desert escarpment, in which the final obsequies were celebrated before the royal burial took place. The vast granite blocks that make up most of the surviving parts of the valley temple present a grim and awesome aspect and also show how rapidly Egyptian stonemasons had developed the craft of stoneworking and how ready the Egyptian architects were to make use of intractable materials. Near the valley temple is the great Sphinx, generally thought to have been cut from the living rock in Khafre's reign as a symbolic guardian of the royal necropolis. Both Khufu and Khafre were remembered as hard and difficult kings in traditions that survived until Greek times. It seems not improbable that the immense resources in material and manpower needed to build their funerary monuments impoverished the country and imposed great strains on the people. They clearly attempted to do more than was reasonably within the competence of their times, both in building bigger than ever before and in using materials that required great effort to work. After their time such display and virtuosity were much modified. Menkaure, who built his pyramid at Giza like his great predecessors, was content with something quite modest; and Shepseskaf was satisfied with a stone-built mastaba, admittedly of great size and embellished with granite, at Ṣaqqārah.

Development of stoneworking

In many ways the 4th dynasty marks a summit in the history of Egyptian achievement, of which the great monuments provide the most obvious testimonies. More is revealed by the remarkable sculptures that dignified both the royal monuments and the tombs of nobles and high officials. This was the period of strong, impressive statuary, of portraiture, and of wonderfully subtle carving of hard stones. They sprang from a situation in which the king, the centre and pivot of the realm, was supreme. As the embodiment of divine royalty, he controlled everything. Such a powerful situation could not last for long; it could not easily be improved upon, but it could deteriorate.

The 5th dynasty. No contemporary records preserve any account of the events that led to the change of dynasty, which occurred in about 2494 BC. The change,

however, was very distinctive and marked by political and religious developments of far-reaching consequence. A romanticized tradition concerning the change is preserved in a story according to which the first three kings of the 5th dynasty were the sons of the wife of a priest of Re (Ra) and of the god Re himself. This tradition finds its source in the undoubted predominance given by the kings of the new dynasty to the worship of Re, the sun god, whose principal shrine was at Heliopolis, a few miles to the north of Memphis on the eastern side of the Nile. The adherence to Re was further emphasized by the regular adoption by the 5th-dynasty kings of the title "'Son of Re," which had only occasionally been used before. From the 5th dynasty this title and its concomitant name, in a cartouche, became a constant element in the royal titulary. Userkaf, the first king of the 5th dynasty, also built a special sun temple in honour of Re at Abū Jirāb (Abu Gurab), some miles to the south of Giza; five of his successors followed his example.

Worship of Re

The kings of the dynasty built pyramids at Abū Ṣīr and Ṣaqqārah, but in these they scarcely emulated the works of their predecessors. But if the pyramids of the 5th dynasty were small, they were made unusually handsome and interesting by the varied architectural detail of the dependent temples and by the fine reliefs with which they were decorated. Although they are now much damaged and largely lost, these reliefs reveal some of the activities in which the kings of the dynasty engaged. From the pyramid of Sahure, the successor of Userkaf, it emerges that a large punitive expedition had been sent against the Libyans, in the course of which much booty was taken; other scenes suggest expeditions perhaps by land and sea to Syria. The activities of Sahure outside Egypt are confirmed by entries on the Palermo Stone recording the bringing of produce from Syria and from Punt (an exotic, unidentified southern land, possibly part of Somalia); evidence of operations during Sahure's reign has also been found in Nubia.

Scenes of the smiting of foreign enemies regularly decorate royal funerary monuments in this dynasty and they soon become stereotyped and conventional. In consequence, it is hard to judge to what extent they truly record events that actually took place. In later times such scenes undoubtedly were wholly conventional and magical in intention; that they may have retained an element of truth in all earlier cases may be postulated. Inscriptions in Sinai show that at least four kings of the 5th dynasty sent expeditions there, and a text of Djedkare Izezi, the eighth king of the dynasty, preserves the first detailed account of the personnel making up an expeditionary force. Other inscriptions record forays into Nubia for raw materials and perhaps also for military purposes by this same Izezi.

Some of the most interesting scenes from the royal monuments of the 5th dynasty occur in the sun temple of Neuserre, the sixth king of the line. Here are shown episodes in the celebration of the Sed festival of the king—the ceremony by which an Egyptian monarch renewed his royal power after many years (perhaps 30) on the throne. The Sed festival was of great antiquity and may have had its origin in a barbaric, primitive custom of slaying the ruling monarch after a certain time in order to avoid having the throne occupied by an elderly, useless, infertile ruler. Buildings that can clearly be associated with the Sed festival were included within the enclosure of the Step Pyramid of Djoser at Ṣaqqārah.

Modification of Egyptian government in the 5th dynasty

The style of Egyptian government was modified during the 5th dynasty, much more freedom being granted to the nobility and high officials. Devolution of power to provincial administrators weakened the central hold on the country; highest office was no longer restricted to members of the royal house. In death high officials were allowed to enjoy more individual and lavish tombs than formerly. From the lively scenes carved on the chapel walls of these tombs, a vivid picture of the daily life on the great estates can be obtained. In the inscriptions set up in their tombs the same officials were able to record their deeds and honours.

Scribal activity flourished as never before, and from this time has come the earliest surviving collection of inscribed papyri. They were found in the funerary temple of Neferirkare, the third king of the 5th dynasty, and were written in the reign of Djedkare Izezi. Relating to the administration of the temple foundation, they contain much material about temple ceremonies, priestly duties, temple equipment and its periodic inspections, income and expenditure—in fact all the details concerning the running of a large establishment of many years' standing. The documents produce a profound impression of careful and well-organized administration.

With regard to religion and to an understanding of the peculiarly complicated beliefs concerning the king and his destiny, the most important written records of the 5th dynasty are the Pyramid Texts, first inscribed monumentally in his pyramid by Unas, the last king of the dynasty. This body of religious texts, consisting partly of ancient incantatory hymns and partly of more recent invocatory utterances designed to help the king achieve his solar destiny, was compiled for royal use only. The humble Egyptian could scarcely expect to share in such an apotheosis; his expectations in death were more modest. But already the conception of a judgment of the dead by the god Osiris was gaining ground and providing a form of hope for posthumous grace. The possibility of enjoying some kind of identifiable existence after death added a new dimension to the spiritual life of the Egyptian, who was already finding life on earth considerably freer and easier than had his ancestors. This expansion of an individual's potential opportunities was to increase greatly during the 6th dynasty and was ultimately to become a significant element in the dissolution of central power at the end of that dynasty.

Major religious texts

The 6th dynasty. As is so often the case regarding the first 1,000 years of Egyptian history, the most positive information about the reign of Teti, first king of the 6th dynasty, can be derived from the burials of high officials. Why a new dynasty began with Teti in about 2345 BC is not known, but the later king lists suggest that a new epoch began with his reign. In the shadow of his own modest pyramid at Ṣaqqārah, two of his viziers, Kagemni and Mereruka, built their mastaba tombs. In the proximity of their tombs to that of their king, they might seem to be following the custom of the 4th dynasty nobles, who derived what advantages they had in death from a burial close to the pyramid of their royal ruler. In the case of Kagemni and Mereruka, however, their tombs seem so placed as if to vie with that of Teti. Already in the 5th dynasty great nobles had been buried in tombs with elaborate, well-decorated rooms in the superstructures above the burial chambers. But such tombs were tactfully built apart from the king's pyramid. Teti's viziers seem to flaunt their own power blatantly by their display. The two tombs are huge, that of Mereruka in particular having more rooms than any other Old Kingdom mastaba. The varied scenes that decorate the walls of these rooms demonstrate clearly how powerful Mereruka was. No informative texts tell of his career, but little imagination is needed to appreciate that the vizier had in the 6th dynasty attained a position which was closer than ever before to that of the king in actual power.

Decentralization of power

The vizier, however, was not the only official to enjoy power and influence. Decentralization of administration led to the growth of provincial centres as independent focuses of power. This decentralization was based on the system of nomes, and the nomarchs were the agents of royal rule throughout the land. They administered justice locally, supervised irrigation, saw to the rehabilitation of the land after the annual flood, maintained the nome sanctuaries, raised men for military and other expeditions as directed by the king, and generally looked after the welfare of the inhabitants of their districts. These high officials were not buried in the royal necropolis; they had developed roots in their provincial homes, and they built their tombs nearby. In their tombs—often decorated in unusual ways and sometimes in a gauche, provincial style—they felt free to inscribe texts of a more personal nature than their ancestors had ever

dared. They described their principal triumphs and extolled without shame their own good qualities. It emerges that nomarchs could institute important activities in quarries and even undertake expeditions; there is always the implication that everything is done in the king's name, but this implication may in some cases have been more a fiction than a true expression of fact. In addition to such freedom of action, the successful, dutiful noble would also expect rewards from the king of a kind unheard of in earlier times.

Served by active, ambitious officers, who for most of the dynasty appear to have served their masters with loyalty, the kings of the 6th dynasty felt capable of entering upon continuous activity in foreign countries. The biographical inscriptions in private tombs tell of many enterprises. Teti set the example; his agents visited Byblos and, probably, Punt; they traded with Nubia and exploited the quarries in the Eastern Desert. The officers of Pepi I continued these activities on his behalf and extended their attentions eastward to the mines of Sinai and into Palestine, where military action was conducted by sea and land. Merenre, son and successor of Pepi I, who may have served as co-regent with his father for a few years, records a visit to the southern boundary of Egypt proper at Elephantine to receive the homage of Nubian chiefs. His personal interest in the land to the south typified the principal concern of foreign policy during the 6th dynasty—the development of trade with Nubia.

Trade with Nubia

It is arguable that this exploitation of trade with Nubia was the direct result of another development—the appointment of the nomarchs of Elephantine as governors of Upper Egypt, with exceptional powers over the southern nomes of Egypt and with a commission to lead expeditions into Nubia. Cause and effect cannot be disentangled now, but it is evident that a powerful noble at Elephantine would seek to enhance his position by taking full advantage of his opportunities to march southward. Trade with Nubia yielded ivory, incense, fine woods, oil, and animal skins; later it was to yield gold as well. But trade in Nubia was hazardous, so caravans needed protection. The governor of Upper Egypt, therefore, had to raise troops and maintain a garrison ready to meet emergencies. He was ideally placed to improve his own position both directly and indirectly by gratifying the king through his activities. Thus the governor Harkhuf ingratiated himself with the child-king Pepi II by acquiring a dancing dwarf for him on one of his expeditions to Nubia.

Pepi II, coming to the throne in about 2294 BC at a very early age, is reputed to have reigned for 94 years. His was the last effective reign of the 6th dynasty; in it there took place many activities characteristic of the dynasty, but in it also a fatal weakness developed, which ultimately led to a complete breakdown of central government. The existence of many tombs of important provincial officials shows that for much of Pepi II's reign there was no noticeable diminution in the style of life and control in the country. The funerary monument of the king himself equally testifies to the fact that he came to the end of his long reign without apparent mishap. Nevertheless, within a few years of his death, Egypt was a divided country.

Division of Egypt

The First Intermediate Period (c. 2160–c. 2040 BC). The weakness that led to the dissolution of centralized control after the reign of Pepi II was compounded of several elements, of which one was political and another religious. The gradual decentralization of government and the consequent increase in the power of provincial nobles has already been noted. The culmination of this movement seems to have coincided with the end of the reign of a very old king. In addition, the temper of the times had been much affected by the growth of the religious ideas associated with Osiris, to the detriment of the old, authoritarian, state religion of Re, in which the king was central and omnipotent. At the same time life in Egypt had become much more humane for the peasant as well as for the noble; much emphasis was placed on justice and moral virtues in the inscriptions of the nomarchs. Egyptians had become more conscious of the possibility of happiness in the earthly life and of a dignified existence in the life to come. These desirable changes in the Egyptian way of life seem not to have brought contentment, however, but instead a loss of orientation and troubled consciences.

The physical and mental anguish brought about by the crisis through which Egypt passed at this time is reflected in literary compositions, in which a doleful picture is painted of the land. For the first time men considered their human lot in a thoughtful and even pessimistic way. The happy, carefree attitude found in so much Egyptian art of earlier and later periods is absent. The malaise was general; the country had lost its sense of direction, and the struggle for power was on. The sequence of events at this time is hard to determine. It is unlikely that foreign invaders contributed to the overall disruption. Some infiltration undoubtedly took place into the Delta from Asia and also, possibly, from Libya, but the numbers involved were probably not great. Politically they may have been influential, but their presence was a symptom of the unrest rather than a cause.

The 7th and 8th dynasties. Attempts to reconcile wholly the dynastic traditions recounted by Manetho with the names in the king lists and the surviving monuments have been unsuccessful. There are, however, patches of firm ground in the morass of uncertainty; inscriptions found at Coptos in Upper Egypt give substance to kings ruling at Memphis and controlling at least part of Upper Egypt. These kings, none of whom seems to have reigned more than a few years, may be identified as the 8th dynasty. The preceding 7th dynasty may have been nothing more than an ephemeral series of princely nobles who attempted to carry on a form of royal power, also from Memphis. Meanwhile in the country the general collapse of authority led to a proliferation of local rulers, grouping and regrouping in alliances of mutual protection or aggrandizement.

Proliferation of local rulers

The 9th and 10th dynasties. Out of this turmoil the most positive coalition to emerge was based on the 20th nome of Upper Egypt, the Heracleopolitan nome, the ruler of which at this time (about 2160 BC) was Akhtoy (Achthoes). He declared himself king of Egypt and began to win over to his side by energetic methods the neighbouring nomes. Gradually he extended his power to the Delta and southward to Elephantine. His victories were not, apparently, maintained by his immediate successors, who made up the 9th dynasty. The situation in Egypt remained disturbed, although it appears, from surviving literary texts of the Middle Kingdom, that the Heracleopolitan rulers were able to support a court in which literature could flourish. Nevertheless the success of the 9th dynasty was limited, and in about 2130 BC the 10th dynasty, an extension of the 9th, succeeded to power at Heracleopolis. The change may have been due to a deterioration in the fortunes of the earlier dynasty, brought about by the rise of a new grouping of nomes in the south led by the nomarchs of Thebes.

A struggle for control in the southern part of Upper Egypt had resolved itself into a conflict between Thebes and Edfu. Theban victory in time led to an inevitable clash with Heracleopolis. The kings of the 10th dynasty seemed well-established in their realm; they even conducted expeditions to Syria. But the southern threat was considerable.

The 11th dynasty. The first Theban to claim the title "king of Upper and Lower Egypt" was Sehertowy Intef. He and his successors were to establish the new 11th dynasty in Upper Egypt, which in turn was to oust the vigorous Heracleopolitan line completely. Little in the known early history of Thebes hints at the power that was to emerge toward the end of the 3rd millennium BC. Nothing especially distinguishes the Old Kingdom nomarchs of Thebes from their contemporaries elsewhere. Sehertowy Intef's father Mentuhotep II was traditionally credited with establishing the victorious line of the 11th dynasty, but, as has been seen, he was not the first to assume the royal titles.

The three 11th dynasty kings named Intef were periodically engaged in extending their control northward, eroding the territory nominally under the sway of the kings at Heracleopolis. But the struggle was not continuous. References in texts suggest that there were long periods of relative peace, in which trade continued between south and north. The situation, however, was unsatisfactory for Egypt as a whole; and the land suffered in consequence, particularly because no general policy for irrigation could be enforced. Inscriptions of the period speak of famine and of steps taken by local officials to combat the shortages. Intef II ruled for more than 50 years and succeeded in establishing a fairly firm frontier north of the Thinite nome in which Abydos lay.

The end of the struggle came eventually in the reign of Mentuhotep II, who came to the throne at Thebes in about 2060 BC. An insurrection in the Thinite nome seems to have provoked the final clash. The details of the hostilities have not been preserved, but it is plausibly thought that the death of Merikare, the last-but-one ruler of the 10th dynasty, may have left the northern regime in a parlous state politically. The Thebans were now set on a victorious course, and the final defeat of the Heracleopolitans took place in about 2040 BC. Mentuhotep II had reunited Upper and Lower Egypt under a single king; he had founded the Middle Kingdom. In retrospect the Egyptians of the New Kingdom linked his name with those of Menes and Ahmose—the three were unifiers of the Two Lands.

Reunification of Upper and Lower Egypt

THE MIDDLE KINGDOM AND THE SECOND INTERMEDIATE PERIOD

The Middle Kingdom (c. 2040–1786 BC). Mentuhotep II marked the reunification of Egypt by adopting a new Horus name, Smatowy, which means "He who unites the Two Lands." While it is known that the reunification was secured only by force, there is no evidence to suggest that the new, effective king of Upper and Lower Egypt imposed a harsh rule on the parts of the country not formerly under his control. Support for the Heracleopolitan house was a political matter, as it was for the Theban house. No rival ideologies were involved in the struggle; throughout its ancient history the two conflicting forces were toward unity and disunity. A united Egypt functioned better and provided superior living conditions for peasant and noble alike; consequently the ending of disunity was undoubtedly welcomed by the majority of Egyptians, who would give no trouble to the new regime if it functioned well.

Reorganization and expansion. All the surviving evidence suggests that Mentuhotep II set about to reorganize the land with great energy, and he soon was able to undertake expeditions into Nubia, Libya, Syria, and Sinai. To satisfy the revived needs of sculptors and architects, he reopened many of the quarries, which had been little exploited for many years. He paid particular attention to Nubia, where an independent kingdom had been established in the days when expeditions southward from Elephantine were discontinued. In accord with the martial character of the times, an effort seems to have been made to occupy the northern part of Nubia. It was no longer enough just to control the trade routes and to police the oases in the Western Desert from which trouble might come.

Foreign expeditions under Mentuhotep II

The active foreign policy of Mentuhotep II was continued by his successor Sankhkare Mentuhotep III, whose principal claim to fame lies in the reopening of the transdesert route from Coptos through the Wadi Hammamat to the Red Sea, from which expeditions could be dispatched to Punt. This approach to Punt may have been an innovation of this reign; for the voyages, large boats of the kind normally found on the run to Byblos were employed. The quarries of the Wadi Hammamat contain many inscriptions of royal officials who came there in the late 11th dynasty. One such official, sent to obtain stone for the royal sarcophagus by Mentuhotep IV, bore the name Amenemhet. He was vizier and governor of Upper Egypt, and he apparently toppled his royal master to become first king of the 12th dynasty in about 1991 BC. The change seems to have been accompanied by some violence, but the struggle may have been more between Amenemhet and other powerful nobles than between Amenemhet and Mentuhotep.

The 12th dynasty. Under Amenemhet began one of the most brilliant periods in the whole of Egyptian history. The kings of the 12th dynasty, fired by a remarkable spirit and endowed with great administrative talents, succeeded in inspiring their subjects to great deeds in all fields. Politically, the 12th dynasty rulers were most adroit in using and then discarding institutions that could be manipulated for their special purposes. The initial success of Amenemhet seems to have depended to some extent on support from powerful nomarchs; in consequence he reinforced the nome system by securing provincial boundaries and re-establishing powers and privileges. With their positions enhanced in this way, the nomarchs became influential and, in time, dangerous. Eventually their power was suppressed again by Sesostris III, but for the time being Amenemhet was happy to reward his loyal nobles. A further shrewd move was to establish a new capital in the north for the reunited kingdom. Although a native of Upper Egypt and a devotee of the god Amon of Thebes, who gained new and significant importance in the 12th dynasty, Amenemhet appreciated the need for a capital from which he could rule all Egypt efficiently. The new capital was built some miles to the south of Memphis and was called Itj-towy, "Seizer of the Two Lands." In the neighbourhood of Itj-towy (near modern al-Lisht) he built his pyramid tomb, and, in the Old Kingdom manner, his courtiers constructed their mastabas around it. In foreign affairs a more positive policy was adopted toward Nubia. Egypt's need of southern products and the necessity of protecting the trade routes led Amenemhet to annex the northern region of Nubia and to place fortresses in the Second Cataract area. The use of permanent fortifications also provided a solution for the perennial problem of Asiatic infiltration into the Delta; in the Wadi Tummilat on the eastern edge of the Delta the "Walls of the Ruler" were constructed—a fortified position that remained a strategic asset for many years.

Increased power of monarchs

Amenemhet I was also an important innovator in another respect. After 20 years of rule he made his son Sesostris co-ruler with himself, and for ten years they ruled together. In this way he ensured that Sesostris would succeed him and so provided a means of continuity for his line, which contributed greatly to the stability of the state. The practice of co-regency was followed by his successors in the 12th dynasty, and it was resumed as an important element of government in the New Kingdom. The co-regency of Amenemhet I and Sesostris I undoubtedly prevented trouble when the old king was assassinated while Sesostris was on campaign in Libya. Sesostris, hearing of his father's death, returned in haste to the capital, anticipating the takeover planned by the conspirators. His prompt action established him as a strong and astute ruler.

For the rest of his reign Sesostris was able to concentrate on reviving Egypt's greatness. He sent punitive expeditions into Nubia and Libya and secured peace in those quarters; he established farther outposts in the south; he developed diplomatic relations with western Asiatic countries; he opened up and actively exploited many quarries. Internally, the effects of his policy were seen in the building activities he initiated throughout the land. Unhappily, only fragments of the structures erected in his reign have survived, apart from an exquisite small shrine recovered from the filling of a pylon of later date at Karnak.

The internal development of Egypt continued increasingly in the reigns of Sesostris II and Amenemhet II. Perhaps the most important project begun at this time was the reclamation of land in the Fayyūm depression—an area neglected in earlier times—which was to become one of the most fertile provinces of the country. Associated with this project was the fostering of the cult of Sebek (Sobk), the crocodile god, in Shedet, the capital of the Fayyūm. Sebek was already much honoured dur-

ing the 12th dynasty at another cult centre a few miles to the south of Thebes. Prosperity, which accompanied the successful policies of the early 12th dynasty kings, attracted to Egypt many foreigners, especially Asiatics, who were now to be found in domestic and artisan situations throughout the land. Trade and diplomacy both with Asia and with the Aegean world flourished, and for the first time Egypt was exposed to foreign influences and ideas in a significant way.

Sesostris III, who became king in about 1878 BC, was the greatest ruler of the 12th dynasty. His energy was especially needed to deal with situations that had developed during the peaceful reigns of his predecessors. Renewed trouble in Nubia obliged him to take various steps to improve arrangements for the control of that land and to facilitate communications. Armed forays were followed by the building of a series of strong forts in the Second Cataract region, the establishing of permanent garrisons and customs posts to control the entry of foreigners, and the cutting of channels through the granite rapids of the First Cataract to enable ships to sail between Elephantine and the south.

Sesostris III's reforms were far-reaching for the future of government in Egypt. The power of the nomarchs, boosted by great responsibilities placed on them early in the dynasty, had again developed to a point where some clearly presented a threat to the state. By means that are no longer clear, Sesostris III clipped their wings completely; the great provincial houses suddenly came to an end. In place of the old system, a new bureaucratic regime was created, based on the capital at Itj-towy. The whole land was divided into three departments, each controlled by high officials responsible for law, agriculture, finance, and general administration; the new system came under the jurisdiction of the vizier.

The golden age of Sesostris III and Amenemhet III

The golden age made possible by the activities of Sesostris III continued through the reign of his son Amenemhet III. With peace abroad, and with the possibility of insurrection at home stifled, Egypt was able to enjoy an unprecedented period of prosperity. The surviving monuments testify in no uncertain way to the artistic brilliance of the period. Sculpture reflected the level of achievement; the fine royal statues of the two reigns are distinguished by a spirituality and humanity quite lacking in the royal sculpture of earlier periods. The kings are not shown as quintessential royalty—divine, aloof, unemotional. Likewise the statues of private persons have qualities of individuality never before found in Egyptian sculpture. But apart from sculpture, the surviving products of order arts and crafts show complete mastery of a great variety of materials. In many respects this was a great age in which the Egyptian emerged as a literate, humane person, appreciative of fine and beautiful things. This was the time when many of the classics of Egyptian literature were written, when the ability to read and write was honoured particularly, and when scribal activity, stimulated by a growing bureaucracy, increased in public and private life. The cult of Osiris, which had become the dominant element in funerary religion, gave hope of salvation for all; everyone who could afford it sought to enjoy the closest communion with the god by setting up a cenotaph, or memorial stela, at the most sacred sanctuary of the god at Abydos.

The 45-year reign of Amenemhet III provided such a long period of peace and prosperous growth that its effects were not easily eradicated by the unstable reigns that followed. The energies of the Egyptians had been channelled so successfully into large-scale, profitable activities that the advantages derived therefrom were not willingly discarded. Outstanding among the public works of the period was the further development of the Fayyūm, where by careful control of irrigation large tracts of land were recovered for cultivation. The attachment of Amenemhet III to that region led him to construct his pyramid at Hawāra, near its northern entrance. Near the pyramid he built a vast structure known in classical times as the Labyrinth, which may have been partly funerary temple and partly administrative centre. It was the great memorial by which the king would be remembered. For many centuries it fulfilled this function, but now only the smallest traces indicate where it once was.

Amenemhet IV and Queen Sebeknefru, the two last rulers of the 12th dynasty, were able in their short reigns to do little more than maintain the character of rule so brilliantly instituted by Sesostris III and Amenemhet III. Egypt's reputation stood high among the nations of the ancient world; its envoys and traders were welcomed in Asia and Crete. The great trading post of Karmah at the Third Cataract in Nubia was a flourishing entrepôt for goods from tropical Africa; the demands of its industries and crafts made Egypt a good customer. Chiefly by peaceful activities, after the more belligerent policies of Sesostris III, did Amenemhet III achieve this peak of reputation for his land. The prosperity did not go unnoticed by foreigners, who were inevitably attracted toward a land offering great opportunities. They provided new skills and techniques; they also introduced an alien element, which in time was to have dire effects on Egypt politically.

Egypt's world reputation at the end of the 12th dynasty

The Second Intermediate Period (1786–1567 BC). The Second Intermediate Period—the time between the end of the 12th dynasty and the beginning of the 18th dynasty—was very different in character from the First Intermediate Period. The First Intermediate Period, although a very confused time, was distinguished more by political troubles than by social upheaval; but the Second Intermediate Period was a time when central control collapsed, and with it the whole fabric of Egyptian life. The reasons for the dissolution of the centralized, well-administered state of the 12th dynasty are to be found, first, in the nature of the bureaucracy set up after the abolition of the nomarchic system and, second, in the growth of foreign enclaves in the Delta. Under the new system of internal government the greatest power in the land rested in the hands of the vizier, the king's principal officer of state.

The 13th–16th dynasties. During the 13th dynasty, which is credited by Manetho with 60 kings, the vizier seems to have been the person who provided not only the power for ruling but also continuity. A great vizier like Ankhu was more important than Sebekemsaf I and Sebekhotep III, the kings he theoretically served. It has been suggested that the kings of this dynasty were appointed or elected for short terms of office. Nevertheless, there is plenty of documentary evidence to show that the state continued to operate efficiently for many years and that the influence of Egypt was still recognized in Syria and Nubia.

Signs of disintegration were already apparent in the emergence in the western Delta of a small but tenacious regime centred on Xois. Manetho describes this line of petty rulers as the 14th dynasty; it had, according to him, 76 kings, and it lasted 184 years. The existence of this line is not confirmed by much evidence, and it can only have been of local importance. More significant were developments in the eastern Delta, where, as in the First Intermediate Period, had grown up concentrations of Asiatics, organized probably on a tribal basis and controlling their domestic affairs with very little reference to the central government of Egypt. The chiefs of these Asiatics were known in Egyptian as *heqau khasut*, "princes of foreign uplands," and it is from this designation that the Manethonian term Hyksos is probably to be derived.

Although the kings of the 13th dynasty remained in general control of most of Egypt for the first 100 years of their rule, there can be no doubt that their authority began to run rather sluggishly in the north from about 1720 BC, when the Asiatics occupied the town of Avaris in the eastern Delta and began to foster the worship of Seth, the local deity, as their own particular cult. The old idea that the Hyksos regime was established as the result of an invasion by Asiatic tribes is now generally discounted in favour of the view that it evolved from a confederation of small tribal groups already settled in the Delta from the time of the 12th dynasty. The growth of the Hyksos power was gradual, and it increased significantly as central control weakened. The steps by which

The growth of Hyksos power

this power developed are not known. But in about 1674 BC Memphis fell to them. At that time the Theban king was Dudimose and the leader of the Hyksos, according to Manetho, was Salitis. It is now thought that this Salitis was Mayebre Sheshi, the first Hyksos ruler to assume the formal trappings of an Egyptian king, to enclose his name in a cartouche, and to use the Egyptian royal titles. He is taken to be the founder of the 15th dynasty, the "Great Hyksos." Perhaps contemporaneously with this line there existed in the Delta also the 16th dynasty, a second Hyksos line; but beyond some of their names scarcely anything is known of the kings that made up this dynasty, and they may have been ephemeral.

Hyksos domination. The extent of Hyksos rule in Egypt has not been established with certainty. The discovery of objects with Hyksos names in Upper Egypt and in Nubia as far south as Karmah has led some historians to conclude that all Egypt was at first conquered by Salitis, as later tradition reported, and that the later Hyksos preferred to confine their active control to the Delta and Middle Egypt. Alternatively, it is suggested that the Hyksos only exercised a nominal suzerainty over Upper Egypt and that the Theban line of kings continued to rule their reduced kingdom uninterruptedly. The situation in Egypt may, however, have been very different from that which was described in later tradition. It is possible that once the Hyksos were in firm control of the north, they established a *modus vivendi* with the Theban kings, which allowed much freedom for travel and trade between the two parts of Egypt and also with Nubia.

Hyksos rule was not, apparently, inimical to Egyptians or to the Egyptian way of life. Under the Hyksos kings no attempt was made to suppress Egyptian religion or culture; in fact, evidence suggests that Egyptian culture flourished more in the Hyksos half of the country than in Thebes at this time. As far as Nubia was concerned, the political developments in Egypt allowed local Nubian princes to set up an independent kingdom, which was supported in alliance by the Hyksos. Relations between the Nubian princes and their Theban neighbours may have been friendly; the evidence of Nubian burials throughout Upper Egypt in graves known as "pan-graves," shows that large numbers of southerners entered Egypt in the late Middle Kingdom and Second Intermediate Period, possibly to serve as mercenaries under the Theban kings. Their rather primitive grave goods form a marked contrast to the highly developed possessions of their Egyptian contemporaries.

The Hyksos rulers Khayan and Auserre Apopi I, particularly the latter, who ruled Egypt for at least 50 years, maintained a rule that was strong and largely Egyptian in character. Good relations were fostered with foreign countries, objects bearing Khayan's name having been found as far away as Baghdad and Knossos. These kings used the title "king of Upper and Lower Egypt," and it is possible that they even allied themselves with the Theban house.

New rulers at Thebes

The 17th dynasty. A new line of rulers had meanwhile assumed control at Thebes, beginning (in *c.* 1650 BC) what Manetho called the 17th dynasty. Although probably descended from the kings of the 13th dynasty, they may be distinguished from their predecessors by being liberated from the control of high officials. At first they were unable to achieve much more than the building of power within the reduced Theban domain. Their situation was similar to that of the Theban nomarchs in the First Intermediate Period, before the reunification of the two lands by Mentuhotep II. By the reign of Seqenenre II, in about 1580 BC, the Theban rulers had so far improved their political position that a clash with the Hyksos became increasingly inevitable. The only record of the fighting that eventually broke out is contained in a literary composition of the New Kingdom. From this biased account it appears that the Hyksos king Apopi picked a quarrel with Seqenenre over a matter interpreted as a religious insult. In the ensuing struggle, no details of which have survived, it is possible that Seqenenre secured some local successes in Middle Egypt but may have lost his life. His mummy, which has survived, bears wounds which were perhaps obtained on the battlefield.

Hyksos–Theban war

Under Kamose, the son and successor of Seqenenre, began the ultimate clash between the Hyksos and the Theban kingdoms. It seems likely that some years had elapsed since the conflict between Apopi and Seqenenre. A contemporary account suggests that fairly peaceful relations existed for the time being, so that the Thebans were even able to pasture their cattle in the Delta. Pride, however, and the desire to reunite the whole of Egypt under a Theban king led Kamose to declare war on Apopi. Initial surprise brought considerable success. The Hyksos adherents in Middle Egypt put up little resistance. Kamose anticipated a move by Apopi to bring in help from his Nubian allies; Kamose also succeeded in penetrating as far as the Hyksos stronghold of Avaris. The final stages of Kamose's victorious campaign are not known, but it is possible that he was obliged to withdraw after Apopi had recovered from the surprise of the initial onslaught and successfully counterattacked. The end of Hyksos rule was, however, now in sight; while it was to remain for Kamose's successor, Ahmose, the first king of the 18th dynasty, to complete the expulsion of the hated foreigners, to Kamose belongs the credit for striking the first really victorious blow. (T.G.H.J.)

II. From the beginning of the 18th dynasty to c. 330 BC

THE NEW KINGDOM

The 18th dynasty. Although King Ahmose (ruled 1570–1546 BC) had been preceded by his older brother, King Kamose, Egyptian tradition regards Ahmose as the founder of a new dynasty because it was he who first unified Egypt under a native ruler. Continuing a practice inaugurated late in the 17th dynasty, Ahmose married his full-blooded sister Ahmose-Nofretari. Such consanguineous marriages were accorded important theological significance in that the queen was given the title of God's Wife of Amon. Further, they reflected a renewed emphasis on divine kingship, which had suffered during earlier periods. Although full-blooded consanguineous marriages failed to survive beyond the first two reigns of the 18th dynasty, the title of God's Wife was thereafter bestowed upon principal heiresses in childhood, and the pharaoh customarily married the most legitimate heiress to validate his right to the throne. Like her predecessors of the 17th dynasty, Queen Ahmose-Nofretari was an influential and highly honoured woman. A measure of her importance is her posthumous veneration at Thebes, where later pharaohs were depicted offering to her as a goddess among the gods.

Consanguineous marriages of the pharaohs

Ahmose. Upon his accession Ahmose had to expel the Hyksos from the delta and regain former Egyptian territory to the south. Destroying the Hyksos stronghold at Avaris, in the eastern delta, he finally drove them beyond the eastern frontier and then besieged Sharuhen (Tall al-Far'ah) in southern Palestine. Ahmose's officers and soldiers were handsomely rewarded with spoil and captives, who became personal slaves. This marked the creation of an influential military class. Later, Ahmose campaigned in the south as far as the Second Cataract (rapids) of the Nile, where a new temple was built at the old fortress of Buhen. For the administration of the regained territory he created a new office—overseer of southern foreign lands, which ranked second only to the vizier. Its incumbent was accorded the honorific title of king's son, indicating that he was directly responsible to the king as deputy, supervising military installations and the delivery of tribute.

The Hyksos domination of Egypt had not been as cataclysmic as Egyptian sources later tended to portray it; thus, the founders of the New Kingdom did not face such difficult problems of internal reconstruction as those at the end of the First Intermediate Period. The early New Kingdom bureaucracy was modelled after that of the Middle Kingdom. The vizier was the chief administrator and the highest judge of the realm. By the end of the 16th century BC the complexity of administration required two viziers, one for Upper and one for Lower

Administration

Egypt. During the 18th dynasty, young bureaucrats received their scribal education in the temple school. There was no essential cleavage between civil servants and priests, and state administrators frequently received priestly appointments (and vice versa). Early in the dynasty many administrative posts were inherited, but royal appointment of capable officials, often selected from military officers who had served the king on his campaigns, later became the rule. The trend was thus away from bureaucratic families and the inheritance of office.

Amenhotep I. Initially, Ahmose's son and successor, Amenhotep I (ruled 1546–1526 BC), directed his attention to the south, where, motivated by the desire to expand Egypt's frontiers, he pushed beyond the old Middle Kingdom boundaries. Such an incentive may also have prompted his little-known ventures in Asia, though in these he may have been responding to pressure from soldiers eager for more rewards. The emerging kingdom of Mitanni perhaps threatened Egyptian security to the north. Mitanni is mentioned on a stela of Amenemhat, who served Amenhotep I.

Foreign wars increased Egypt's wealth. The state god Amon-Re benefitted as the spoils of war entered the god's treasuries or were utilized to finance construction of sacred monuments at Thebes. The New Kingdom saw continued devotion to Amon-Re, whose cult gave the king, as his representative, the mission of expanding Egypt's frontiers. Under Amenhotep I the pyramidal form of royal tomb was abandoned in favour of a rock-cut tomb, which he hoped would provide a more secure burial place, and, except for Akhenaton, all subsequent New Kingdom rulers were buried in concealed tombs within the famous Valley of the Tombs of the Kings in western Thebes. Separated from the tombs royal mortuary temples were erected at the edge of the desert. Perhaps because of this innovation, Amenhotep I later became the patron saint of the workmen who excavated and decorated the royal tombs.

Thutmose I. Lacking a surviving heir, Amenhotep I was succeeded by one of his generals, Thutmose I (ruled 1525–*c.* 1512 BC), who had married the royal heiress and God's Wife Ahmose. The new king first directed his attention to maintaining military domination in the south. Assisted by able generals, he then executed a brilliant campaign into Syria and across the Euphrates, where he erected a victory stela near Carchemish.

Thutmose II. Although born of a minor wife of Thutmose I, Thutmose II's (ruled *c.* 1512–1504 BC) claim to the throne was validated by his marriage to the God's Wife Hatshepsut. In his first year of rule Thutmose II launched an attack against rebels in the south, wherefrom gold and other costly products, now vital to the economy, were obtained. During his reign there was still no forceful assertion of empire in Asia, military activity there being limited to a raid among the Bedouin, who were an irritation to overland trade; the port of Byblos (in Lebanon) supplied Egypt with pine timber, used, for example, in the construction of columned halls at Karnak.

Joint rule of Hatshepsut and Thutmose III. At Thutmose II's death his major queen, Hatshepsut, had only a young daughter; but a minor wife had borne him a boy, now nearly ten years old, who served as a priest in the Temple of Amon. This son, Thutmose III (ruled 1504–1450 BC), created the Egyptian Asiatic empire and became Egypt's greatest ruler. During his first two years as king Thutmose III theoretically controlled the land, but actually it was the mature Hatshepsut who governed as regent. In Thutmose III's second regnal year she boldly assumed the kingship herself. According to one version of the history of the event, the oracle of Amon proclaimed her king at Karnak, where she was crowned. A more propagandistic account, promulgated in texts and reliefs of her splendid mortuary temple at Dayr al-Baḥrī, ignores the reign of Thutmose II and asserts that her father, Thutmose I, announced her accession. Upon becoming king, Hatshepsut became the dominant partner in a joint rule that lasted until her death in 1482 BC; there are monuments dedicated by Hatshepsut that depict both pharaohs. She had the support of various powerful personalities, who significantly did not form a homogeneous faction. In styling herself king, Hatshepsut adopted the royal titulary but avoided the epithet mighty bull, regularly employed by other kings. Although in her reliefs she was depicted as a male, pronominal references in the texts generally reflect her womanhood. Similarly, much of her statuary portrays her as male, but there are rarer examples that render her as a woman. In less formal documents she was referred to as "King's Great Wife," while Thutmose III was "Pharaoh." There is, thus, a certain ambiguity in the treatment of Hatshepsut as king.

Hatshepsut's assumption of the kingship

Her temple gives the impression of a monarch concerned with pacific enterprises, such as the transporting of obelisks for Amon's temple and a commercial expedition to Punt. It has been suggested that her policy and art reflect a return to Middle Kingdom ideals. There is evidence of bellicose tendencies, however, and it is probable that she appeared on the battlefield in Nubia; as she grew older, it was natural that her participation in warfare lessened, so that her monuments generally lauded her ability in other directions. The notion of a struggle between a pacific queen backed by a peace party and a warrior king supported by a war party can no longer be maintained, nor does the romantic tale of Thutmose III's minority being filled with hatred of Hatshepsut really reflect the true situation (the persecution of Hatshepsut's memory was not pursued until late in his reign, and not all her officials were purged by Thutmose III at her death).

Thutmose III. During Hatshepsut's floruit Egypt's position in Asia deteriorated because of the expansion of Mitannian power and influence in Syria. Shortly after her death, the Prince of Kadesh, representing Mitannian interests, stood with troops of 330 princes of a Syro-Palestinian coalition at Megiddo; such a huge force was more than a defensive massing of troops, and probably the intention was to advance against Egypt.

Thutmose III proceeded to Gaza with his army and then to Yehem, subjugating rebellious Palestinian towns along the way. At a famous consultation with his officers concerning the best route to reach the coalition stationed on the other side of the Mt. Carmel ridge, the King overruled his officers' advice and selected a shorter but more dangerous route through the ʿArūnah Pass and then led the troops himself. The march went smoothly; and, when the Egyptians attacked at dawn, they prevailed over the enemy troops and besieged Megiddo. Egyptian sources describe the battle only briefly; but it must be considered one of the decisive ones in world history, because it was fundamental to the establishment of Egyptian Empire in Asia.

Establishment of an Egyptian Empire in Asia

Thutmose III meanwhile coordinated ancillary operations involving the landing of other army divisions on the Syro-Palestinian littoral, whence they proceeded inland, so that the total strategy resembled a pincer technique. The seige ended in a treaty by which Syrian princes swore an oath never to rebel against Thutmose III. Significantly, this oath was binding only upon the parties immediately concerned and not upon future generations. This reflects an aspect of Egyptian bureaucracy —that officials were bound only to the reigning pharaoh, and old appointments required reconfirmation upon the accession of a new king.

By the end of the first campaign Egyptian domination extended northward to a line linking Byblos and Damascus. Although the Prince of Kadesh still remained to be vanquished, Assyria sent lapis lazuli as tribute; Asiatic princes surrendered their weapons, including a large number of horses and chariots, at Megiddo. By contrast, Thutmose III took only a limited number of captives; and in this respect he was magnanimous, eschewing the large-scale deportations of prisoners. The King appointed Asiatic princes to govern the towns and took their brothers and sons to Egypt, where they were educated at the court. Most eventually returned home to serve as loyal vassals, though some remained in Egypt as

court pages. In order to insure the continued fealty of Asiatic princes, Egypt maintained garrisons that could quell insurrection and supervise the delivery of tribute. This organization was necessary in order to prevent an excessive drain on Egyptian manpower, for capable administrators were needed more at home than abroad.

Thutmose III conducted numerous subsequent campaigns in Asia. The submission of Kadesh was finally achieved, but Thutmose III's ultimate aim was the destruction of Mitanni. He used the navy to transport troops to Asiatic coastal towns, thus avoiding arduous overland marches from Egypt. His great eighth campaign led him across the Euphrates; although the countryside around Carchemish was ravaged, the city apparently was not taken, and the Mitannian prince fled to another land. The psychological gain of the eighth campaign was perhaps greater than its military success, for Babylonia, Assyria, and the Hittites all sent tribute in recognition of Egyptian dominance in Syria. Although Thutmose III never subjugated Mitanni, he firmly established an empire in Asia. This necessitated constant campaigning that contrasts with the forays of his predecessors. Egyptian royal inscriptions generally abound in exaggerations and pomposity; but the annals of Thutmose III are remarkably succinct and accurate, and autobiographies of his officials display a similar tendency toward factual reporting. The objectivity of narrative texts is a measure of the greatness of the King, who brought back to Karnak Asiatic flora and fauna, which were carefully depicted in his festival hall.

Amenhotep II. About two years before his death the aging king appointed his 18-year-old son, Amenhotep II (ruled 1450–1425 BC), as coregent. Just prior to his father's death Amenhotep II set out on a campaign to an area near Kadesh, in Syria, whose city-states were now caught up in the power struggle between Egypt and Mitanni; Amenhotep II killed seven princes, whose bodies he shipped back to Egypt to be suspended from the ramparts of Thebes and Napata in the Sudan. In his seventh and ninth years Amenhotep II made further campaigns into Asia, where the Mitannian king pursued a more vigorous policy. The revolt of the coastal city of Ugarit was a serious matter, because Egyptian control over Syria required bases along the littoral for inland operations and the provisioning of the army. Ugarit was pacified, and subsequently the fealty of Syrian cities, including Kadesh, was reconfirmed. The number of prisoners of war recorded for Amenhotep II's ninth-year campaign is an extraordinary 89,600. Although it had an estimated population of 4,500,000, the amount of free cultivable land in Egypt was relatively small; thus, the situation was economically unfavourable to the institution of large-scale slavery (the bulk of the land was tilled by native tenant farmers, who delivered from one-third to one-half of their crop to the landowners). Amenhotep II's enormous figure cannot represent a large-scale introduction of slaves into Egypt but, rather, probably a census of Syrians under Egyptian domination.

Amenhotep II's campaigns into Asia

Thutmose IV. Under Amenhotep II's son Thutmose IV (ruled 1425–1417 BC), the Asiatic empire was maintained but not expanded. The Mitannian king Artatama, having been successful against the Hittites, posed a problem for Egypt, and Thutmose IV sought to establish peaceful relations with him through a diplomatic marriage. Artatama gave his daughter in marriage, the prerequisite for which was probably an Egyptian concession clarifying which Syrian city-states belonged within the Mitannian sphere. This was the first of such diplomatic marriages, paving the way for the age of Amenhotep III, when the emphasis shifted from war to diplomacy and the enjoyment of the luxury of empire.

Foreign influences during the early 18th dynasty. During the empire period Egypt maintained commercial ties with Phoenicia, Crete, and the Aegean islands. The Egyptians portrayed goods obtained through trade as foreign tribute. In the Theban tombs there are representations of Syrians bearing Aegean products and of Aegeans carrying Syrian bowls and amphorae—indicative of close commercial interconnections among Mediterranean lands. Egyptian ships trading with Phoenicia and Syria journeyed beyond to the Aegean islands and Crete, a route that explains the occasional confusion of products and ethnic types in Egyptian representations. The most prized raw material from the Aegean world was silver, which was lacking in Egypt, where gold, especially from the Nubian mines, was relatively abundant.

One result of empire was a new appreciation of foreign culture. Not only were foreign *objets d'art* imported into Egypt but Egyptian artisans copied Aegean wares as well. Imported textiles inspired the ceiling patterns of Theban tomb chapels, and Aegean art with its spiral motifs and rendition of movement influenced Egyptian artists. Under Amenhotep II a deeper recognition of the value of things foreign is manifest in the acceptance of Asiatic gods in Egypt, where Astarte and Resheph became revered for their reputed potency in warfare; and, later, Astarte was honoured in connection with medicine, love, and fertility. Some Asiatic gods were eventually amalgamated with similar Egyptian deities; thus, Astarte was combined with Sekhmet, the goddess of pestilence, and Resheph with Mont, the war god. Just as Asiatics themselves tended to be accepted as members of Egyptian society and could on occasion rise to important positions, so their gods, though represented as foreign, were worshipped according to Egyptian cult practices. The breakdown of Egyptian isolationism and increased cosmopolitanism in religion were also reflected in Amon-Re's concern for the welfare of Asiatics.

Late in his reign Thutmose III extended the Egyptian boundary southward to the Fourth Cataract of the Nile. Egypt's superior military equipment, notably the horse and chariot, had already pacified the south, and only occasional forays into the desert were required to protect gold-mining operations. In Nubia there was peaceful symbiosis of Egyptians and Nubians, who gradually became Egyptianized—sons of chiefs were educated at the Egyptian court, some remaining as pages, others returning to Nubia to serve as administrators and be buried in Egyptian fashion. Gradually, Nubian fortresses lost their strategic value and became commercial and administrative centres. Open towns developed around the forts, and in several of the temples outside the walls the cult of the divine pharaoh was established. Lower Nubia supplied Egypt with raw materials: gold, semiprecious stones, and building stone; from further south came African woods, perfumes, oils, ivory, panther skins, and ostrich plumes. With the introduction of the shadoof, an irrigation device, Nubia's agricultural productivity increased, and Nubian farmers worked on lands belonging to the king or temples, raising grain for Egypt.

Extension of Egypt's boundary southward

Amenhotep III. Queen Mutemuia, a commoner, was the honoured mother of Thutmose IV's son Amenhotep III (ruled 1417–1379 BC), who acceded to the throne at about the age of 12. Amenhotep III soon wed a commoner, Tiy, who became his principal wife. This marriage represented a departure from the concept of God's Wife and the principle of marrying the royal heiress, who would become the chief queen. Earlier in the dynasty military men had served as royal tutors; but Tiy's father was a commander of the chariotry, and now the royal line was even more directly influenced by the military. In his fifth year Amenhotep III claimed to have achieved a considerable victory over Kushite rebels, but the Viceroy of Kush probably actually led the troops. Several temples, erected under Amenhotep III in the Sudan, attest to the prosperity of the south under Egyptian domination; peaceful relations also prevailed with Asia during this reign. A commemorative scarab from the king's tenth year announced the arrival of the Mitannian princess Gilukhepa, along with 317 women; thus, another diplomatic marriage helped maintain friendly relations between Egypt and its former foe. Another Mitannian princess was later received into Amenhotep III's harem, and during his final illness the Hurrian goddess Ishtar of Nineveh was sent to his aid. Although military activity was at a minimum, Egypt maintained garrisons of troops in Asia to deter local insurrection. At the expense of older bureaucratic families and the principle of in-

heritance of office, military men acquired high posts in the civil administration. Most influential was the scribe and commander of the elite troops Amenhotep, son of Hapu, whose reputation as a sage survived into the Ptolemaic period.

Amenhotep III sponsored much building, especially in the Theban area. At Karnak he erected the huge third pylon, and at Luxor he dedicated a magnificent new temple to the Amon of Opet. The King's own mortuary temple in western Thebes was unrivalled in its size; little remains of it today, but its famous Colossi of Memnon reflect a penchant for the colossal in art and architecture. One of these statues served as an object of public veneration, before which men could appeal to the king's *ka* (spiritual self), representing the transcendent aspect of kingship. In Karnak, smaller but exquisitely carved statues of Amenhotep, son of Hapu, were placed to hear the petitions of passersby.

Thutmose III had shifted the centre of administration to Memphis, though Thebes had retained its religious importance. Amenhotep III renewed the emphasis upon Thebes and erected a huge palace complex on the west side of the river. Among the highest-ranking officials at Thebes were men of Lower Egyptian background, who received tombs in the Theban necropolis that were both large and highly refined in their decoration. An academic quality is evident in the tombs, certain scenes of which were copied from Old Kingdom reliefs. The subsequent revolutionary art of the Amarna period perhaps reflects a reaction against the studied perfection of Theban artists. Antiquarianism is evidenced in Amenhotep III's celebration of his jubilees (anniversaries of his accession), which were performed at his Theban palace in accordance, it was claimed, with ancient writings. Tiy participated in these ceremonies connected with the renewal of kingly power. Although Amenhotep III's benefactions to Amon were many, he was not thereby placating a powerful priesthood at Thebes; in fact, the important post of second prophet of Amon was held by the Queen's brother, who was of military background. Amenhotep III's last years were spent in ill health. To judge from his mummy and less formal portraits of him from Amarna, he was obese when, in his 38th regnal year, he died and was succeeded by his son Amenhotep IV (ruled 1379–1362 BC), the most controversial of all the pharaohs.

Renewed emphasis upon Thebes

Amenhotep IV (Akhenaton). Problems surrounding the reign of Amenhotep IV, who in his fifth year of rule altered his name to Akhenaton, have been complicated by the possibility that a major portion of his reign was spent as a coregent with his father, a view unsupported by the weight of evidence. Akhenaton's earliest monuments are entirely conventional in their iconography and style. Certain high officials initially continued in the posts they had acquired under Amenhotep III, and the high priest of Amon functioned as late as Akhenaton's fourth year. In his first years a temple was erected to the god Re-Harakhti at Karnak. Although the style of art was conventional, the hawk-headed sun god was given a didactic name mentioning Aton, the solar disk. This term had long been in use, but under Thutmose IV there had been a reference to the Aton as a god, and under Amenhotep III instances of the Aton treated as a deity became more frequent. Akhenaton, therefore, did not create a new god but rather singled out this god from among others.

A major change occurred in about his fourth year, when, to the east of the great Amon temple at Karnak, a large temple to the Aton was erected. In this temple the radical art of the King's reign is found in its most extreme form, and large statues of Akhenaton depict him as the primordial god in a most bizarre fashion. Adorning the temple walls were scenes of the jubilee that the king cocelebrated with the Aton, whose names were written as a king's within cartouches. No anthropomorphic representations of the deity were used; the supreme god was depicted as a sun disk, from which arms bearing the symbol of life emanate. During the construction of this temple the cult of Amon and other gods was suspended, and the worship of the Aton in an unroofed sanctuary superseded that of Amon, who had dwelt in a dark shrine of the Karnak temple. Frequently portrayed in the brightly painted reliefs was the King's wife Nefertiti, whom he had married before his accession. Like Queen Tiy, she may have come from a military family. Her father was possibly Ay, who became prominent at Amarna and later held the throne briefly.

At about the time that he altered his name to conform with the new religion, the King transferred the capital to a virgin site at Tell el-Amarna in Middle Egypt. There, he constructed a well-planned city—Akhetaton—comprising temples to the Aton, palaces, official buildings, villas for the high ranking, and quarters for the labourers. In the eastern desert cliffs surrounding the city, tombs were excavated for the courtiers; and deep within a secluded *wādī* the royal sepulchre was prepared. Reliefs from these tombs have been invaluable for reconstructing life at Amarna. The murals stress the life of the royal family more intimately than heretofore. In Akhenaton's ninth year a more monotheistic didactic name was given to the Aton, and an intense persecution of the older gods, especially Amon, was pursued. Amon's name was excised from many older monuments throughout the land, and occasionally the word gods was expunged. Evidence suggests that the King's monotheistic fervour intensified.

The transfer of the capital to Amarna

Akhenaton's will undoubtedly was a major factor in this religious and cultural revolution. There was an emphasis upon Maat—truth as determined by the king's view of things—so that, for example, his peculiar physiognomy became the norm for representing not only members of the royal family but also commoners as well. In religion the accent was upon the sun's life-sustaining power, and naturalistic scenes adorned the walls and floors of Amarna buildings. The king's role in determining the composition of the court is expressed in epithets given to officials he selected from the lesser ranks of society, including the military. Few officials had any connection with the old aristocracy, and certain courtiers who had been accepted at the beginning of Akhenaton's reign were purged.

The view that Akhenaton was too concerned with religious matters to be interested in the fate of the empire and that as a pacifist he opposed military action is questionable. Wall scenes at Amarna frequently depict soldiers at the capital; and cuneiform tablets dealing with Egyptian foreign relations with Asia show evidence of the maintenance of Egyptian garrisons in Asia. Akhenaton even responded on occasion to the frantic pleas of Asiatic princes for additional military assistance, although rival princes tended to portray the situation as worse than it actually was.

Akhenaton's religion was not entirely monotheistic—he himself was a god, the son and manifestation of the Aton; thus, the Amarna theology bolstered the King's divine status. As the King was the Aton's high priest, so there was a high priest of Akhenaton; and the Aton and Akhenaton uniquely shared their jubilees. Because the King had the most direct access to the Aton, he worshipped the Aton on the commoner's behalf. With the discarding of traditional myths, old beliefs were undermined, and the religious needs of the people were poorly provided for. Actually, there is evidence that commoners in their private lives had not completely abandoned the old deities. In the field of ethics there is no discernable advance over the older polytheistic religion.

Smenkhkare. Perhaps because of the supremely exalted status accorded to Akhenaton in his association with the Aton, his religious experiment was short-lived. Texts from Amarna speak of an eternal process of daily revitalization of Akhenaton by the Aton's rays, but unfortunately the King was mortal. Toward the end of his reign he selected Smenkhkare (ruled 1364–61 BC), his son by a minor wife, to serve as coregent. By this time Queen Nefertiti, who had achieved unprecedented status in the iconography of Amarna, had died; and Smenkhkare's wife, Akhenaton's eldest daughter, Meritaton, became the first lady. Although the royal residence was still maintained at Amarna, by Smenkhkare's third year the cult of Amon had been revived at Thebes, where the King's mortuary temple was built.

Tutankhamen. Smenkhkare was succeeded by his child brother Tutankhaton (ruled 1361–1352 BC), who likewise married one of Akhenaton's daughters, Ankhesenpaaten. For two years the young couple lived at Amarna, pursuing the worship of the Aton. Amon-Re, however, becomes increasingly manifest in the documentation as there was a transition back to orthodoxy. Finally, the decision was made to abandon Amarna, and the King and Queen altered their names to Tutankhamen and Ankhesenamen. An inscription of Tutankhamen's fourth year testifies to the complete restoration of the old cults. The Aton is not mentioned, and the Amarna period is described as one of misery and of the withdrawal of the gods from Egypt.

Restoration of the old cults at Thebes

Under Tutankhamen a considerable amount of building was accomplished in Thebes. His Luxor colonnade bears detailed reliefs depicting the great traditional Beautiful Feast of Opet. Tutankhamen honoured Amenhotep III, the builder of the Luxor temple, whom he even referred to as his father. Although Tutankhamen's inscriptions credit him with the restoration, in point of fact there were actually two powerful individuals behind the scene: the God's Father, Ay, who served as vizier and regent (his title "the God's Father," indicates a close relationship to the royal family), and the general Horemheb, who functioned as royal deputy.

The Amarna period evidently had not occasioned as marked a withdrawal from empire as early historians, stressing Akhenaton's pacifism, maintained. Military campaigns during Tutankhamen's reign are difficult to trace, but Horemheb probably fought in Asia, supporting vassal princes against the Hittites. Effective administration continued in Nubia and the Sudan. Tutankhamen achieved fame because of the discovery of his rich burial in the Valley of the Tombs of the Kings. There is some reason to suspect that because of his return to orthodoxy his burial equipment was extraordinarily lavish.

Ay. Tutankhamen's funeral was conducted by his successor, the aged Ay (ruled 1352–1348 BC). Hittite sources indicate that, upon the plea of the widowed Ankhesenamen, the Hittite king Suppiluliumas dispatched to Egypt a prince for her to wed, but Egyptians treacherously murdered the prince on his journey. Although some scholars hold Ay responsible for this action, Horemheb was more likely involved in the plot. Ay was succeeded by Horemheb (ruled 1348–1320 BC), who married Queen Nefertiti's sister Mutnedjemet to secure his claim to the throne.

Horemheb. Horemheb dismantled many monuments erected by Akhenaton and his successors and used the blocks as fill for huge pylons at Karnak. At the Luxor temple, Horemheb appropriated Tutankhamen's reliefs by surcharging the latter's cartouches with his own. Although certain scholars have minimized Horemheb's devotion to Amon and his persecution of the memory of the Amarna pharaohs, it is significant that he was crowned in the Amon temple at Luxor. Horemheb appointed new officials, not from established families but from the army. Horemheb concentrated on domestic problems. He issued a series of police regulations dealing with the misbehaviour of palace officials and military personnel, and he reformed the judicial system, reorganizing the courts and selecting competent new judges. As a reaction against the Amarna period, Horemheb's royal tomb reflects a trend away from the theology of the solar religion and emphasizes Osirian beliefs of the afterlife.

The 19th and 20th dynasties. Having no son, the aged Horemheb selected his general and vizier, Ramses, to succeed him.

Ramses I and Seti I. This new king, as Ramses I (ruled 1320–1318 BC), hailed from the eastern delta, and with the new 19th dynasty there was a noticeable political shift to the north. Ramses I was succeeded by his son and coregent, Seti I (ruled 1318–1304 BC), who completed his father's burial and provided him with mortuary chapels at Thebes and Abydos. Seti I was a competent warrior and a builder of impressive monuments.

Political shift to the north

A noteworthy development in narrative art is evident in scenes showing Seti in various stages of battle, on the north wall of the Great Hypostyle Hall at Karnak. These reliefs and stelae found in Palestine testify to Egypt's determination to reassert authority over its weakened empire. The Hittites had become the dominant Asiatic power. Before tackling them, Seti I laid the groundwork for military operations in Syria—he fought against the troublesome Bedouin and brought important Palestinian cities to submission; then, following the strategy of Thutmose III, he secured the coastal cities and gained Kadesh. Although the engagement with the Hittites was successful, Egypt acquired only temporary control of a portion of the north Syrian plain. A treaty was concluded with the Hittite king Muwatallis, but the Hittites subsequently pushed farther southward and gained Kadesh by the time of Ramses II. Seti I ended a new threat to Egyptian security when he defeated Libyans attempting to enter the delta.

Seti I's reign was called a renaissance, and Egypt's prosperity is reflected in his considerable monuments and their refined decoration. In his fourth year he issued a lengthy decree intended to prevent arbitrary action by government officials against the personnel and property of his mortuary temple at Abydos, invoking harsh physical penalties and the curse of the gods upon malefactors. Some scholars regard this as a reflection of the tensions of a church–state rivalry, but it may simply have been a means of insuring good government.

Ramses II. Seti I appointed his son Ramses II (ruled 1304–1237 BC) as coregent. Early Egyptologists called Ramses II "the Great." During his long reign, there was a prodigious amount of building, ranging from religious edifices throughout Egypt and Nubia to a new cosmopolitan capital, Pi-Ramesse, in the eastern delta; his cartouches were carved ubiquitously, even on monuments that were not his own; his penchant for decorating vast areas of temple walls with battle scenes gives the false impression of a warrior-king mightier than Thutmose III. But scholars today are cautious in evaluating his exploits. His most famous series of reliefs are those pertaining to the Battle of Kadesh, fought in his fifth year; and the texts accompanying the pictorial record a comprehensive account of the battle. Egyptian sources regard the battle as an Egyptian victory, but the Hittite coalition actually fared better than did the Egyptian army. After the arduous struggle, his officers advised him to make peace, saying, "There is no reproach in reconciliation when you make it." In succeeding years Ramses II campaigned in Syria; after a decade of stalemate, a treaty in his 21st year was concluded with Hattusilis III, the Hittite king.

Treaty with the Hittites

The rise of Assyria and unrest north of the eastern Mediterranean had forced the Hittites into accepting this treaty, while Ramses II may have feared a new Libyan threat to the western delta. Egyptian and Hittite versions of the treaty survive. It contained a renunciation of further hostilities, a mutual alliance against outside attack and internal rebellion, and the extradition of fugitives. Illustrative of cosmopolitanism in religion was the role of gods of both lands as witnesses. The treaty was further cemented 13 years later by Ramses II's marriage to a Hittite princess. The second half of his long reign was, thus, relatively peaceful and prosperous. By his numerous wives the King had an immense family. Although Isinofre was the important queen for the succession, he especially honoured Nefertiri by dedicating a temple to her at Abu Simbel, in Nubia, and by supplying her with a magnificent tomb in the Valley of the Tombs of the Queens.

Merneptah. Ramses II's 13th son, Merneptah (ruled 1236–1223 BC), was his successor. Several of Merneptah's inscriptions, of unusual literary style, treat an invasion of the western delta in his fifth year by Libyans, supported by a number of groups of Sea Peoples who had travelled from Anatolia by sea to Libya in search of new homes. The Egyptians defeated this confederation and settled captives in military camps to serve as Egyptian mercenaries. One of the inscriptions concludes with a poem of victory, famous for its words "Israel is desolated and has no seed." This is the earliest documented mention of

Israel; and, on the basis of it, many have assumed that Ramses II was the pharaoh of the oppression and Merneptah the king at the time of the Exodus. Certain cities in Syria, such as Tyre and Simyra, were still in Egyptian hands, but nomadic movements in southern Palestine were a source of concern for Egypt. Early in his reign Merneptah asserted Egyptian suzerainty in Palestine and destroyed Gezer. In the Libyan war, slain soldiers were enumerated by counting their severed hands or phalli, and certain captives were impaled through the stomach; but Merneptah displayed more humane qualities when he dispatched grain to the Hittites during a famine. Recognizing the terms of the Hittite treaty, the Egyptians aided the Hittites militarily in the protection of Syrian possessions.

Last years of the 19th dynasty. Upon the death of Merneptah, competing factions contended for the succession—possibly an unfortunate consequence of Ramses II's immense family. Following a five-year reign of a usurper, Amenmesse (ruled 1222–1217 BC), Merneptah's son Seti II (ruled 1216–1210 BC) acquired the throne. Although little is known of his political activities, his presence is well attested on monuments at Karnak, some of which he usurped. References to an enemy may reflect the unstable political situation, because his successor, Siptah (ruled 1209–1203 BC), was actually installed in the kingship through the efforts of a powerful Syrian royal butler who had become chancellor of Egypt. The young Siptah was married to Tausert, who, upon Siptah's death, herself ruled as king in 1202–1200 BC, appropriating Siptah's regnal years as her own and restoring her first husband's name (Sethos II) over Siptah's. A description in a later papyrus of the end of the dynasty alludes to a Syrian usurper, probably the chancellor, who subjected the land to harsh taxation and treated the gods as mortals with no offerings in their temples.

Beginning of the 20th dynasty

Setnakht. Order was restored by a man of obscure origin, Setnakht (ruled 1200–1198 BC), the founder of the 20th dynasty, who ruled only briefly and appropriated Tausert's tomb in the Valley of the Tombs of the Kings.

Ramses III. Setnakht's son Ramses III (ruled 1198–1166 BC) was the last great pharaoh. There are problems in evaluating his achievements because he emulated Ramses II and copied numerous scenes and texts of Ramses II in his mortuary temple at Madīnat Habu, one of the best preserved temples of the empire period. Thus, the historicity of certain Nubian and Syrian wars depicted as accomplishments of Ramses III is subject to doubt. There are, however, certain battles actually fought by Ramses III that were more decisive in world history than any of Ramses II's. In his fifth year Ramses III defeated a Libyan invasion of the delta. The figures given in one scene on the walls of his mortuary temple indicate that several tens of thousands of the Libyan enemy perished; but another, similar scene gives less inflated figures, in the thousands.

A greater menace lay to the north, where a confederation of Sea Peoples was progressing by land and sea toward Egypt. These barbarians, who had overrun the Hittite kingdom, had camped in Amor in Syria. In his eighth year Ramses III engaged them successfully on two frontiers—a land battle in Palestine and a naval engagement in one of the mouths of the delta. These two victories were significant for the civilized world; they permitted the survival of Egyptian culture for another millennium. Forced away from the borders of Egypt, the Sea Peoples sailed farther westward, and there is reason to believe that Sicilians, Sardinians, and Etruscans originated from groups of them. The Philistine and Tjekker Sea Peoples, who had come by land, were established by the Egyptians in military camps in the southern Palestinian coastal district in an area where the overland trade route to Syria was threatened by Bedouin attacks. Initially created to protect Egyptian commercial interest, these groups later became independent of Egypt. Ramses III utilized certain groups of these peoples as mercenaries, even in battle against their own kinfolk. Reliefs at Madīnat Habu differentiate among the various groups of Sea Peoples on the basis of their costumes and weapons. In his 11th year Ramses III successfully repulsed another great Libyan invasion by the Meshwesh tribes. Meshwesh prisoners of war, branded with the king's name, were settled in military camps in Egypt, and their descendants subsequently became politically important because of their military role.

These great defensive wars in the first half of Ramses III's reign were less profitable economically than the Asiatic campaigns of the 18th dynasty. Under Ramses III the estate of Amon received only one-fifth as much gold as in Thutmose III's time. Although there are masterpieces of Egyptian art at Madīnat Habu, much of the relief inside the temple and the quality of the masonry betray a decline. Toward the end of his reign, administrative inefficiency in coping with a deteriorating economic situation resulted in the government's failure to deliver grain rations on time to necropolis workers, whose dissatisfaction was expressed in demonstrations and in the first recorded sit-down strikes in history. A different sort of internal trouble originated in the royal harem, where a minor queen plotted to murder Ramses III so that her son might become king. Involved in the plot were palace and harem personnel, government officials, and army officers. A special court of 12 judges was formed to try the accused, who received the death sentence. It is noteworthy that among both the accused and the judges were persons of foreign origin—a clear illustration of how far foreigners could advance in influence. An indication of moral decay is the conduct of certain judges, five of whom were punished for carousing with the women on trial. There is no evidence that this harem conspiracy caused Ramses III's death.

Harem conspiracy against Ramses III

Ramses IV. Ramses III was succeeded by his son Ramses IV (ruled 1166–1160 BC). In an apparent act of piety, Ramses IV saw to the compilation of a long papyrus in which the deceased Ramses III confirmed the temple holdings throughout Egypt; Ramses III had provided the largest benefactions to the Theban temples, in terms of donations of both land and personnel. Of the annual income to temples, 86 percent of the silver and 62 percent of the grain was awarded to Amon. The document demonstrates the economic power of the Theban temples, for the tremendous landholdings of Amon's estate throughout Egypt involved the labour of a considerable portion of the population; but the ratio of temple to state income is not known, and there is some later evidence that, unless exempted by special decree, the temples were obliged to deliver taxes to the state. Indeed, the Amon temple, which figures prominently in the papyrus, is the King's own mortuary temple, for the deified Ramses III was himself a form of Amon-Re, known as United with Eternity.

The later Ramesside kings. The Ramesside period saw a tendency toward the formation of high-priestly families, which the pharaohs sometimes tried to control by appointing outside men to the high priesthood. Such a family had developed at Thebes in the second half of the 19th dynasty, and Ramses IV tried to control it by installing Ramessesnakht, the son of a royal steward, as Theban high priest. Ramessesnakht participated in affairs normally outside the scope of the high priest's office; he personally led an expedition to the Wadi Hammamat (modern Wādī Rawḍ ʿĀid) quarries in the eastern desert, and at Thebes he supervised the distribution of rations to the workmen decorating the royal tomb. Under Ramses V (ruled 1160–1156 BC), Ramessesnakht's son not only served as steward of Amon but also held the post of administrator of royal lands and chief taxing master. Thus, this high-priestly family had acquired extensive authority over the wealth of Amon and over state finances; but to what degree this represented a threat to the crown and led to the decline of royal authority is uncertain. Part of the problem in evaluating the evidence is that Ramesside history is viewed from a Theban bias, because Thebes is the major source of information. Documentary evidence from Lower Egypt, where the king normally resided, is meagre because of unfavourable conditions there for the preservation of monuments or papyri.

A long papyrus from the reign of Ramses V contains

valuable information on the ownership of land and taxation. In Ramesside Egypt most of the land belonged to the state and the temples, while most peasants served as tenant farmers. Some scholars interpret this document as indicating that the state retained its right to tax temple property (at an estimated one-tenth of the crop).

Ramses VI (ruled 1156–1148 BC), a probable son of Ramses III, next took the throne. He usurped much of his two predecessors' work, including the tomb of Ramses V; a papyrus refers to a possible civil war at Thebes. Following the death of Ramses III the Asiatic empire had rapidly withered away, and Ramses VI is the last king whose name appears at the Sinai turquoise mines. The next two Ramses (ruled 1148–1140 BC) were obscure rulers, whose sequence has been questioned. During the reigns of Ramses IX (ruled 1140–1121 BC) and Ramses X (1121–1113 BC) there are frequent references in the papyri to the disruptive presence of marauding Libyans near the Theban necropolis.

Loss of the Asiatic empire

By the time of Ramses IX the Theban high priest had attained great influence, though textual references suggest that royal authority still outranked that of the high priest. Early in the reign of Ramses XI (ruled 1113–1085 BC), during a civil war at Thebes, the high priest Amenhotep, the son of Ramessesnakht, was suppressed from his office for nine months; evidence indicates that the King called upon the Viceroy of Kush and his Nubian troops to restore order and reinstate the high priest to his office. For many years scholars have believed that the growth of priestly power culminated with the assumption of kingship by the high priest Herihor at Thebes. This interpretation has, however, been seriously challenged, and it is now believed that Herihor had no connection with the preceding high-priestly family but was an army officer, who subsequently not only assumed the high priestship but also the secular offices of viceroy of Kush and vizier. In the Khons temple at Karnak, Herihor was represented as high priest of Amon in scenes adjoining those of Ramses XI. This in itself was unusual, but subsequently he took an even bolder step in having himself depicted as king to the exclusion of the still-reigning Ramses XI. Herihor's kingship was a purely Theban matter, and at Thebes these years were referred to as a renaissance, which lasted a decade.

With the shrinkage of the empire, the supply of silver and copper was cut off, and the amount of gold entering the economy was reduced considerably. The economically distressed inhabitants of western Thebes took to pillaging the tombs of kings and nobles during the reign of Ramses IX; the despoiling continued into the reign of Ramses XI, and even the royal mortuary temples were stripped of their valuable furnishings. Nubian troops, called in to restore order at Thebes, themselves contributed to the depredation of monuments. This pillaging brought fresh gold and silver into the economy, and there was a rise in the price of copper. The price of grain, which had been inflating, dropped.

While Ramses XI was still king, Herihor died and was succeeded as high priest by Piankh, a man of similar military background. A series of letters from Thebes tell of Piankh's military venture in Nubia against the former Viceroy of Kush, while Egypt was on the verge of losing control of the south. With the death of Ramses XI, the governor of Tanis, Smendes, became king, founding the 21st dynasty.

Society and culture in the New Kingdom. Although Egyptians readily personified abstract concepts and at an early period represented their cities and local areas as divine personifications, there was no such divinization of Egypt itself.

The king. Instead, the divine king embodied the state, so that Egypt's success in war was conceived of as an accomplishment not of the state or the people but of the pharaoh, who was always regarded as the victor, even in battles he personally had not fought. Theologically, the birth and death of the king were linked with the concepts of order and chaos in nature and society, for without a king no organized state and society could be maintained. Upon the king the people pinned their hopes for the success of their collective endeavours; anarchy not only could produce genuine hardship in a land so much dependent upon a centrally controlled irrigation system but it partook of the chaos of nonexistence.

In addition to the many inscriptions referring to the king as a god, there are texts in which he is treated as a human, subordinate to the gods. An argument that such humanizing texts detract from the pharaoh's divinity fails to account for the fact that the highest gods themselves displayed very human qualities in Egyptian myths. Although Egyptians of the New Kingdom frequently prayed to gods, they did not pray directly to the person of the king but rather to his transcendent essence, the *ka*, which might be embodied in a statue. The dual nature of the king, both divine and human, can be explained by distinguishing between the institution of divine kingship and the actual person filling the divine office. During the 18th dynasty there was a tendency, especially in texts describing the king engaged in sport, to portray him as a superhuman among humans. Perhaps as a reaction to this trend, Akhenaton stressed his divine nature through his unique relation with the Aton. In Ramesside times the divine aspect of the king's personality was emphasized in royal inscriptions and through the cult of the king in temples.

Dual nature of the king in the New Kingdom

Actually, much depended upon the strength of the individual ruler. Law was simply the expression of the king's will through royal decrees regulating the functioning of society. During the height of the New Kingdom, kings had the power to make appointments in the civil service, the army, and the priesthood. Courtiers surrounding a king might present their plans to him, but he would frequently ignore their advice and act independently; this became a literary device serving to emphasize the pharaoh's superior wisdom. Only in the king were such vast legislative, executive, and judicial powers concentrated, for the dualism of the vizierate and the rest of the administration served as a check upon any one person acquiring too much authority. The New Kingdom town mayors, who were subject to the two viziers, became the principal provincial administrators, and their large number tended to consolidate the king's authority over them.

The civil service. Efficient administration demanded the perpetuation of an educated class. Literary texts, designed during the Middle Kingdom to make the scribal profession attractive, were constantly utilized in schools of the New Kingdom; in addition, new texts extolling the scribal occupation were composed to counteract the allure of the military and priestly professions (the hardships of the soldier's life were emphasized). During the New Kingdom an individual of humble background could receive a scribal education and enter the civil service, sometimes rising to a position of considerable importance. Although there was a tendency toward the formation of families of high officials, especially early in the 18th dynasty, the buildup of such an aristocracy was moderated by the king's power to appoint and remove government officials. Frequently, the names and representations of high officials were destroyed in Theban tombs because of a family's fall from royal favour. Increased social mobility, emphasizing an individual's merit, made the New Kingdom officialdom more fluid in composition than it had been in earlier periods.

The military. A distinguishing feature of New Kingdom society was the emergence of an influential military class and the increased number of career military men; the empire depended upon the armed forces for its growth and maintenance. Military tutors educated young princes; and, when the king set out to war, he was assisted by skilled officers and soldiers, who were rewarded for their valour on the battlefield with lands, captured slaves, and other booty. In order to persuade sons to follow their father's military profession, families could hold the lands as long as each new generation provided a son who would make the army his career. Thus did military families develop. In addition, a small percentage of the male population was conscripted; apparently, only civil-service officials were exempted from the draft. Dur-

Development of military families

ing the 18th dynasty the bulk of the army was composed of native Egyptian soldiers, supplemented with Nubian troops; under Amenhotep III the Sherden group of Sea Peoples entered the army. The use of prisoners of war increased during the Ramesside period, when Libyans, Sea Peoples, and Bedouins were settled in military camps.

The military was the closest to the king; it rivalled the civil service and the priesthood in terms of individual advancement on the basis of merit. The fact that a distinguished military officer or a king's old companion at arms could receive royal appointment to a high position in the civil administration or in the priesthood was a potential source of tension with the established bureaucracy. During the Ramesside age, royal butlers, who functioned as a sort of royal bodyguard on the king's campaigns, were frequently foreigners; they had considerable political influence and often more effective power than the vizier.

The priesthood. In practice, the king's role as the god's high priest was delegated to members of the priesthood, whose activities included the daily ritual of provisioning the god. In return, the priests received income from the temple endowments, which was ritually offered to the god but then became the priests' sustenance. As temple property increased through the expansion of the empire and royal donations, the personnel connected with the temples and their estates multiplied. During the 18th dynasty the administration of the Amon temple was generally supervised by a high state official such as the vizier. An ecclesiastical background was not a prerequisite for the highest priestly posts; the pharaoh often appointed high state officials or even military men to top positions in the priestly hierarchy. Priestly positions were often inherited, but the power of royal appointment served as a check upon the growth of priestly power. At least the lower grades of priests and temple personnel were subject to conscription into the army or for corvée labour. Today, scholars are more cautious in evaluating the evidence from the Ramesside period that had once been interpreted as demonstrating a church–state conflict.

Artisans. Another important segment of the society consisted of the craftsmen. During the New Kingdom the number of specialized artisans producing wares for a more urban and cosmopolitan society increased. Most were freemen producing goods for their patrons—*e.g.*, the government, the temples, or nobles—in return for wages. A wealthy patron might honour the best of his artisans by naming them in his tomb or by rewarding them with land or a funerary endowment. That craftsmen took pride in their profession is indicated by instances in which son followed father in his skilled trade. Artisans also produced products for the market. A seller would evaluate his ware in terms of copper or silver weights or pieces, and the purchaser would in turn offer some product or products of equal evaluation. Thus, trading was not simple barter but involved reckoning in terms of a sort of abstract monetary unit.

Commerce and trade. Private industry in the New Kingdom apparently did not expand significantly beyond that which an individual household might produce. Private commerce was limited to trading at the village marketplace and to waterfront transactions of private merchants plying the Nile with very modest commodities. More developed industry and commerce was a function of the state or temples, and international trade was a royal monopoly.

The common people. The masses consisted of uneducated labourers, with peasants forming the backbone of a basically agrarian economy. Slavery was not of fundamental importance to New Kingdom agriculture; most of those who were free serfs or tenant farmers tilled the fields and received a share of the harvest. At those times of year when a larger work force was required for irrigation or harvesting, the institution of the corvée was utilized. Only the official class was generally exempted from such compulsory labour, and even priests could be compelled to perform menial work; but, because wealthier Egyptians could afford to hire substitutes, the corvée was hardest on the poor.

Corvée labour

Slaves. Slaves were either acquired in foreign wars or bought from merchants. The number of slaves per household rarely exceeded two. A considerable number were donated to the temples and employed in the temple workhouses, and state mining and quarrying operations made use of slave labour. An owner could sell or even hire out his slaves, and a slave could be emancipated by his owner. Occasionally a slave even married a member of his household.

Social mobility. The potential for social mobility existed, so that New Kingdom Egypt was devoid of anything like a caste system. Many persons rose to prominence from humble beginnings, and the careers of many officials indicate that the classes were not mutually exclusive. Some historians of the New Kingdom have emphasized power struggles among the various classes, but little positive evidence exists, and the propagandistic quality of Egyptian texts might easily lead to wrong interpretations. The evidence also suggests that such conflicts as there may have been were utilized by the king to support his supreme authority.

Women. The Egyptian wife enjoyed considerable prestige as "mistress of the house." So far as is known, marriage involved no formal religious ceremony but was a contractual arrangement, initiated by the suitor applying to the girl's father, and it could be dissolved relatively easily by the husband and probably by the wife. Among commoners, brother–sister marriages did not exist, and monogamy prevailed, though there were concubines who did not have the legal status of wife. A woman's legal status, whether single or married, was in law and in fact equal to a man's, and she could initiate and perform a broad range of legal acts, including the willing of property and the bringing of suit. Women also functioned significantly in religious affairs; many wives of Theban officials were chantresses of Amon, and the role of women in the religious education of the family must have been important.

Culture and religion. As a result of the empire and increased foreign contacts, the ethos of Egyptian culture became more cosmopolitan, and foreigners and elements of their culture and religion were accepted. While the empire was being fashioned, there was confidence and optimism, reflected in autobiographical inscriptions; the appreciation of life and the acceptance of new trends are discernible in scenes from Theban tomb chapels. The beautifully executed tomb chapels of the 18th dynasty are testimonials that Egyptians enjoyed life and sought its prolongation beyond death.

Both the Amarna revolution and altered political circumstances had a reaction on the culture of the Ramesside period. Few tombs of the 19th and 20th dynasties contain biographies of the tomb owner, and there is a significant shift to the depiction of the gods and the netherworld. Ramesside tombs present a sombre impression of withdrawal from earthly life.

The Ramesside period has often been characterized as an age of personal piety. An increased dependence of man upon god is clearly evident in the wisdom literature that stresses man's fragile nature and god's omnipotence. Long introductions in private letters often contain references to the writer's piety and prayerful approach to problems. The common expression in letters, "I am all right today; tomorrow is in the hands of god," reflects the ethos of the age.

Role of the oracle in daily affairs

A manifestation of the increased dependence upon god was the frequency with which the divine oracle of the god entered into daily affairs. The oracle of Amon was consulted on matters pertaining to the state and temple as well as on community problems. To some extent, Amon made decisions in conformance with the king's will; but not all oracular decisions were rigged to conform to the king's will or the priests' wishes. The method of making oracular decisions involved a portable bark, or litter, of the god on the shoulders of lay priests in a public appearance. The god's pronouncements on a wide variety of matters, including legal disputes, were conveyed by the motions of the bark indicating approval or disapproval to questions requiring a yes or no answer. After the New

Kingdom, oracles became even more prominent in determining the course of events.

The religiosity of the Ramesside period is reflected in surviving biographical inscriptions. They are chiefly autobiographies of high priests inscribed on sections of the temple wall or on statues set up in the temple. Such inscriptions had a didactic function in urging future priests to follow the same paths of piety and devotion to the god. The religious context for the expression of personal achievement was carried over into the postempire period.

EGYPT FROM 1085 BC TO THE MACEDONIAN INVASION

Weakness and foreign domination to the mid-7th century. During the Tanite 21st dynasty (1085–945 BC), certain Theban high priests locally assumed the title of king. There is no indication of conflict, however, between the priests and the Tanite pharaohs. Indeed, the dating of documents, even at Thebes, was in terms of the Tanite reigns, and apparently there were close family ties between the Tanite kings and the Theban high priests. Piankh's son, Pinudjem I, who relinquished the office of high priest and assumed the kingship at Thebes, was probably the father of the Tanite pharaoh Psusennes I. Beginning with Herihor and continuing through the 21st dynasty, the high priests' activities included the pious rewrapping and reburial of New Kingdom royal mummies. The ransacking of the royal tombs during the 20th dynasty necessitated the transfer of the royal remains to two caches—the tomb of Amenhotep II and a cliff tomb at Dayr al-Baḥrī—where they remained undisturbed until modern times. Dockets pertaining to the reburial of these mummies contain important chronological data from the 21st dynasty.

The 21st dynasty. Little is known of the political activities of the Tanite pharaohs, although the burials of King Psusennes I and his successor, King Amenemope, were discovered at Tanis. This was a period of the usurpation of statuary and the reuse of material of earlier periods. At Karnak, Pinudjam I, who decorated the facade of the Khons temple, usurped a colossal statue of Ramses II, and Psusennes I's splendid sarcophagus from Tanis had originally been hewn for King Merneptah. Much of the archaeological remains from Tanis, the capital of the 21st and 22nd dynasties, comprises material transported from other sites, notably from Pi-Ramesse. During the 21st dynasty, Egypt lost control of Nubia; and, with the demise of Egypt's Asiatic empire, the kingdom of Israel developed under the kings David and Solomon. During David's reign, Philistia served as a buffer between Egypt and Israel; but upon David's death the next to the last king of the 21st dynasty, Siamon, invaded Philistia and captured Gezer. If Egypt had any intention of attacking Israel, Solomon's power forestalled Siamon, who presented Gezer to Israel as a dowry in the diplomatic marriage of his daughter to Solomon. This is indicative of the reversal of Egypt's status in foreign affairs since the time of Amenhotep III, who had written the Babylonian king, "From of old, a daughter of the king of Egypt has not been given to anyone."

The 22nd and 23rd dynasties. During the 20th dynasty captive Meshwesh Libyans had been settled in military colonies in Egypt; the leader of the colony at Bubastis was designated great chief of the Meshwesh. Toward the end of the 21st dynasty the great chief of the Meshwesh Sheshonk (the biblical Shishak) secured special privileges from King Psusennes II and the oracle of Amon for the mortuary cult of his father at Abydos. The oracle proffered good wishes not only for Sheshonk and his Libyan family but, significantly, also for his army. With a strong military backing, Sheshonk eventually took the throne, founding the 22nd dynasty (945–*c.* 730 BC). Military controls were established, with military garrisons under Libyan commandants serving to quell local insurrections, so that the structure of the state became more feudalistic. The dynasty's success is also attributable to the Libyans' willingness to adopt Egyptian mores and to express their piety to the Egyptian gods through the building of temples. The dynasty tried to cement relations with Thebes through political marriages with members of priestly families. King Sheshonk's son Osorkon married Psusennes II's daughter, and their son eventually became high priest at Karnak. By installing their sons as high priests and promoting such marriages, kings of the 22nd dynasty strove to terminate the inheritance of important priestly offices at Thebes. But frequent conflicts arose over the direct appointment of the Theban high priest from the sons of Libyan kings and over the inheritance of the post by men of mixed Theban and Libyan descent. This tension took place against a background of Theban resentment of the northern dynasty. During the reign of Takelot II, strife concerning the high priestship led to civil war at Thebes. The King's son was appointed high priest, and he achieved some semblance of order during his visits to Thebes.

Libyan rule

The 22nd dynasty revived Egyptian influence in Asia. Five years after Solomon's death (*c.* 931), Sheshonk I entered Palestine and plundered Jerusalem; booty from this campaign financed an enlargement of the Amon temple, and Egypt regained some of its lost commercial prestige. There is no evidence of further military intervention in Palestine. Under King Osorkon II Peywed Libyans posed a threat to the western delta, perhaps necessitating a withdrawal of attention from Asia. The latter part of the dynasty was marked by increasing fragmentation of the land. The kings of a contemporary 23rd dynasty (*c.* 817–*c.* 730 BC) were recognized at Thebes.

Renewed influence in Asia

With the collapse of the New Kingdom, Egypt ceased to control territories to its south; but Kushites (Ethiopians) in the area of the Fourth Cataract in the Sudan maintained a brand of Egyptian culture. The earliest known member of the Kushite royal family was Alara, whose piety toward Amon is mentioned in several inscriptions. His brother and successor, Kashta, proceeded into Upper Egypt, forcing Osorkon III to retire to the delta. Although Kashta did not adopt a pharaonic titulary, he assumed the title of king and compelled Osorkon III's daughter Shepenwepe I, the God's Wife of Amon at Thebes, to adopt his daughter Amonirdis I. While the Libyans sought to control Thebes through the high priesthood, the Kushites stressed the role of the God's Wife of Amon, who no longer was the spouse of a king but remained a virgin as the consecrated partner of Amon.

The 24th and 25th dynasties. Meanwhile, the eastern delta capital, Tanis, lost its importance to the commercial city of Sais in the western delta. A Libyan prince of Sais, Tefnakhte, attempting to gain control over all Egypt, proceeded southward to Heracleopolis after acquiring Memphis. This advance irritated the Kushites, whose ruler Piankhi (more correctly, Piy) proceeded into Egypt (*c.* 730 BC), where he received the submission of the northern coalition. Piankhi, in his victory stela, is portrayed as a man conforming strictly to Egyptian traditions; and his piety, especially to Amon, is stressed. In this campaign an internal slave trade developed, involving the sale of prisoners of war from the north in Thebes.

Upon Piankhi's return to Kush, Tefnakhte reasserted his authority in the north, where he was eventually succeeded by his son Bocchoris, the sole king of the 24th dynasty (720–714 BC), identified by the historian Manetho (flourished *c.* 300 BC). Piankhi's brother Shabaka meanwhile founded the Ethiopian 25th dynasty (716–656 BC); he burned Bocchoris alive and brought all Egypt under his rule. In this period Egypt's internal politics were affected by the growth of the Assyrian Empire. In Palestine and Syria frequent revolts against Assyria were aided by Egyptian forces. Against the power of Assyria, the Ethiopian troops met with little success, partly because the base of the Egyptian military lay in the Sudan and partly because the Libyan–Egyptian princes of the delta, strategically the most important part of Egypt, rendered the Ethiopians only nominal allegiance.

Growth of the Assyrian Empire

Although the earlier years of King Taharqa, who as second son of Shabaka had succeeded his brother Shebitku, were prosperous, the confrontation with Assyria became acute. In 671 BC the Assyrian king Esarhaddon entered Egypt and drove Taharqa into his homeland. Two years later Taharqa regained a battered Memphis,

but in 667 BC Assyria's King Ashurbanipal forced Taharqa to Thebes, where the Ethiopians held ground. Taharqa's successor, Tanutamon, defeated at Memphis a coalition of delta princes who supported Assyria, but Ashurbanipal's reaction to this was to humiliate Thebes, which the Assyrians plundered. The Kushites then withdrew from the Egyptian political scene, although Kushite culture survived in the Sudanese Meroitic kingdom for another millennium.

Independence and cultural revival under the 26th dynasty. Assyria, unable to maintain a large force in Egypt, supported several loyal delta vassal princes, including the powerful Psamtik. But the Assyrians faced serious problems closer to home, and Psamtik was able to assert his independence and extend his authority as king over all Egypt without extensive use of arms, inaugurating the Saite 26th dynasty (664–525 BC). Psamtik I took Thebes peacefully, and his policy toward Thebes was wise. The most influential man there was its mayor and the fourth prophet of Amon—Montemhat. Psamtik I allowed Montemhat to retain his post and, in an effort to appeal to pro-Ethiopian sentiments at Thebes, allowed the God's Wife of Amon and the Votaress of Amon (the sister and daughter of the late king Taharqa) to remain. Psamtik I's own daughter Nitocris was adopted by the incumbent Votaress of Amon and thus became heiress to the important position of God's Wife. In the text relating to this adoption, Psamtik I carefully acknowledged the preceding Ethiopian dynasty. Essential to the settling of internal conflicts was the Saite dynasty's superior army, composed of Libyan soldiers, whom the Greeks called Machimoi (warriors), and Greek and Carian mercenaries. Because Greece was overpopulated, many Greeks emigrated to Egypt. Greek pirates raiding the delta coast were induced by Psamtik I to serve in his army and were settled like the Machimoi in colonies at the delta's strategically important northeastern border. Trade developed between Egypt and Greece, and more Greeks emigrated to Egypt.

Foreign policy under Saites

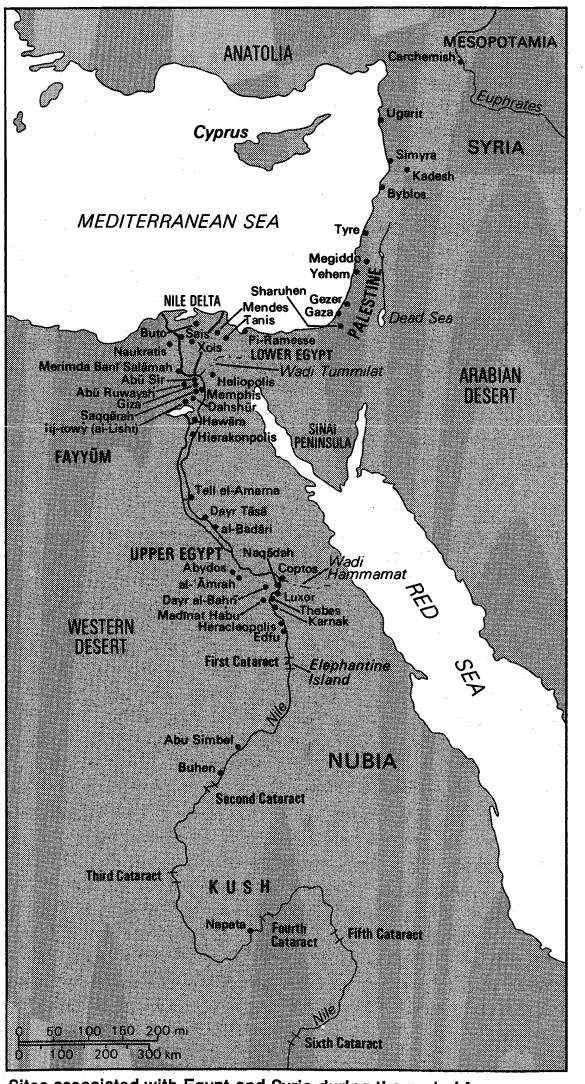
The Saite dynasty generally pursued a foreign policy that avoided territorial expansion and tried to preserve the status quo. Assyria's power was waning. In 655 BC Psamtik I marched into Philistia in pursuit of the Assyrians, and in 620 BC he apparently repulsed Scythians from the Egyptian frontier. During the reign of his son Necho, Egypt supported Assyria as a buffer against the potential threat of the Medes and the Babylonians. Necho was successful in Palestine and Syria until 605 BC, when the Babylonian Nebuchadrezzar inflicted a severe defeat on Egyptian forces at Carchemish. After withdrawing his troops from Asia, Necho concentrated on developing Egyptian commerce; the grain that was delivered to Greece was paid for in silver. Although a canal linking the Nile with the Red Sea remained unfinished, the circumnavigation of Africa by Phoenician ships sent by Necho may have succeeded. During the short reign of Psamtik II, there was a campaign into Nubia, involving the use of Greek and Carian mercenaries who left their inscriptions at Abu Simbel; the reason for this unusual campaign may be found in the erasure at this time of the names of the Ethiopian kings from monuments in Egypt. Psamtik II also made an expedition to Phoenicia accompanied by priests; whether it was a goodwill mission is subject to question.

The next king, Apries, tried unsuccessfully to end Babylonian domination of Palestine and Syria. With the withdrawal of Egyptian forces, Nebuchadrezzar destroyed the temple in Jerusalem. Apries' army was then defeated in Libya when it attacked the Greek colony at Cyrene; this led to an army mutiny and to civil war in the delta. A new Saite king, Ahmose (Amasis), usurped the throne and returned to a more conservative foreign policy in a reign characterized by internal tranquillity. In order to secure the loyalty of Greek soldiers who had sided with Apries in the civil war, Ahmose withdrew them from the military colonies and transferred them to Memphis, where they formed a sort of royal bodyguard. At the same time he limited Greek trade in Egypt to Sais, Memphis, and Naukratis, the latter becoming the only port to which Greek wares could be brought so that the government could enforce taxes on imports and on business. In following a status quo foreign policy, Ahmose supported a waning Babylonia, now threatened by Persia; but not long after his death the Persian Cambyses penetrated Egypt, reaching Nubia (525 BC).

The political policy of the Saite kings was to utilize foreigners as mercenaries in order to prevent foreign invasions. An element within Egyptian culture, however, resisted this; and the presence of foreigners in Egypt, both as invaders and settlers, led to the rise of a nationalism that provided psychological security in days of political uncertainty. An important manifestation of this nationalism was a cultural revival recalling Egypt's glorious past. In this revival the Old Kingdom tended to provide the norm. The god Seth, who had been an antithetic element in Egyptian religion, came to be detested as the god of foreign lands, and his worship was proscribed. While pursuing their pragmatic foreign policy, the Saite kings were careful not to offend the nationalistic sensitivity of the people and priests. One problem that arose was the financing of restorations and construction of new temples; in addition to state contributions for this purpose, private people, inspired by nationalistic fervour, donated parcels of land to the temples in return for a priestly office and its income.

Nationalism and cultural revival

Egypt under Achaemenid rule. According to the account of the Greek historian Herodotus, Cambyses' conquest of Egypt was ruthless and sacrilegious. Contemporary Egyptian sources, however, treat him in a more favourable light.



Sites associated with Egypt and Syria during the period from the Old Kingdom to the Macedonian conquest.

The 27th dynasty. As founder of the Persian 27th dynasty (525–404 BC), Cambyses assumed the full titulary of an Egyptian pharaoh and paid honour to the goddess Neith of Sais. Cambyses' unfavourable later reputation probably resulted from adverse propaganda by Egyptian priests, who resented his reduction of temple income. Cambyses' successor, Darius I, was held in higher esteem because he was actively concerned with improving the temples and restored part of their income, and he codified laws from the earliest times to Ahmose. Darius I's tolerance and his attempt to win over priests and learned Egyptians was part of his politics to retain Egypt as a lasting part of the Persian Empire. Egypt, together with the Libyan oases and Cyrenaica, formed the sixth Persian satrapy (province), whose satrap resided at Memphis, while Persian governors under him held posts in cities throughout the land. Under Darius I the tax burden upon Egyptians was relatively light, and Persians aided Egypt's economy through irrigation projects and improved commerce, enhanced by the completion of the canal to the Red Sea.

The Persian defeat by the Athenians at Marathon in 490 BC had significant repercussions in Egypt. In 486 BC a revolt broke out in the delta, perhaps instigated by Libyans of the west delta. The result was that the Persian king Xerxes reduced Egypt to the status of a conquered province. Egyptians dubbed him the "criminal Xerxes." He never visited Egypt and showed no interest in the temples; he appears not to have utilized Egyptians in high positions in the administration. Xerxes' murder in 465 BC was the signal for another revolt in the western delta by a Libyan dynast, Inaros, who succeeded in acquiring control over the delta and was supported by Athenian forces against the Persians. Inaros was crucified by the Persians (454 BC), who regained most of the delta. Herodotus made a famous visit to Egypt around 450 BC. During the reign of Darius II there was a fanaticism against the Jewish colony at Elephantine by Egyptians who resented the exclusiveness of the colony's religion; the priests of Khnum bribed the Persian commander there to lead Egyptian troops and destroy the Jewish temple in 410 BC.

The 28th, 29th, and 30th dynasties. The death of Darius II prompted another rebellion by Libyan dynasts in the delta, and Amyrtaeus alone formed a Saite 28th dynasty (404–399 BC); his rule was recognized in Upper Egypt at a time when Persia's troubles elsewhere forestalled an attempt to regain Egypt.

Native rule

Following internal troubles in the delta, a 29th dynasty (399–380 BC) emerged from the eastern delta city of Mendes. Achoris, its third ruler, was especially vigorous, and the prosperity of his reign is indicated by many monuments in Upper and Lower Egypt. Once again Egypt was active in world politics, a result of the balance of power between Persia and Greece and a buildup of the Egyptian army and navy. Achoris' policy was based on the principle that any enemy of Persia was a friend of Egypt, so his envoys were busy setting up alliances abroad. The Egyptian army comprised not only Greek mercenaries but also Greek commanders; the mercenaries were not permanent residents of military camps in Egypt but native Greeks seeking payment for their services in gold. Gold coins were minted to pay these new mercenaries, but gold and silver were scarce in Egypt, which no longer possessed the Nubian gold mines. Aided by the Greek commander Chabrias of Athens and his elite Greek troops, Achoris prevented a Persian invasion; but upon Achoris' death his son Nepherites II lasted only four months before a general, Nectanebo I (Nekhtnebef) of Sebennytos, usurped the throne, founding the 30th dynasty (380–343 BC). In 373 BC the Persians attacked Egypt, and, although Egyptian losses were heavy, disagreement between the Persian satrap Pharnabazus and his Greek commander over strategy and an inundation of the delta saved the day for Egypt. With the latent dissolution of the Persian Empire under the weak Artaxerxes II, Egypt was relatively safe from further invasion, and relative stability is reflected in the considerable building achieved in Nectanebo's name.

Egypt had a more aggressive foreign policy under Nectanebo's son Tachos. Possessing a strong army and navy composed of Egyptian Machimoi and Greek mercenaries and supported by Chabrias and the Spartan king Agesilaus, Tachos (also known as Zedhor) invaded Palestine. But friction between Tachos and Agesilaus and the cost of financing the venture proved to be Tachos' undoing. In an attempt to secure quick money, Tachos had imposed taxes and seized temple property. Egyptians, especially the priests, resented this increased burden and supported Tachos' nephew Nectanebo II (Nekhtharehbe) in his usurpation of the throne. The ill-fated venture into Asia proved too costly for a basically agrarian economy, and Egypt henceforth had to be content with defensive military action.

Agesilaus supported Nectanebo in his defensive foreign policy, and the priests sanctioned the new king's building activities. Meanwhile, Persia enjoyed a resurgence under Artaxerxes III; but a Persian attack upon Egypt in 350 BC was repulsed. In 343 BC the Persians once again marched against Egypt. The first battle was fought at Pelusium and proved the superiority of Persia's strategy. Eventually the whole delta, then the rest of Egypt, fell to Artaxerxes III, while the Egyptian king fled to Nubia.

The 31st dynasty. Artaxerxes founded the 31st dynasty; he dealt harshly with Egypt, razing city walls, rifling temple treasuries, and removing sacred books. Persia acquired rich booty in its determination to prevent Egypt from further rebelling.

Return of Persian rule

After the murder of Artaxerxes III, in 338 BC, there was a brief obscure period during which a Nubian prince, Khababasha, seems to have gained control over Egypt, but Persian domination was re-established in 335 BC under Darius III. In 333 BC Alexander the Great of Macedonia defeated the Persians at Issus; and, although Gaza resisted his advance, Egypt was handed over to him in 332 BC by the Persian satrap. Alexander was acclaimed pharaoh at Memphis and greeted as the son of Amon by the priest of the god's oracle at the oasis of Siwa.

(E.F.W.)

III. Hellenistic and Roman Egypt

MACEDONIAN AND PTOLEMAIC EGYPT (332–30 BC)

The Macedonian conquest. When Alexander the Great marched his army of Macedonians and Greeks into Egypt in the fall of 332, he found a country ready to throw off Persian control and a people who looked upon him as something of a liberator: the ancient kingdom of the pharaohs fell to him without a battle. Alexander at once visited the oasis of Siwa, the seat of the oracle of Amon that had been famous throughout the Greek world for generations. There he was told by the oracle that he was the son of Amon and thereby acquired in Egyptian eyes confirmation of his newly acclaimed position as pharaoh.

Perhaps Alexander's most important act was the founding of a new city, Alexandria, west of the Nile Delta near the small village of Rakotis, on a site suitable for a major port. Offshore was an island that, connected to the shore by a causeway, would act as protection from the open waters of the Mediterranean to afford two good harbours for shipping. The city flourished in later years, became the seat of government of Egypt, and, as a major cultural centre of the East for centuries, was by far the most famous and successful of Alexander's city foundations.

Founding of Alexandria

In the spring of 331 Alexander marched east, leaving Egypt under the divided administration of Macedonians, Greeks, and Egyptians. The Persian satrap was replaced by two Egyptian governors or nomarchs, but the military forces were divided among a number of commanders. The system, however, does not seem to have long survived Alexander's departure. A Greek, Cleomenes of Naukratis, seems to have been entrusted with taxation and soon gained, with Alexander's consent, absolute control of the country. Ancient writers attest his deviousness in securing revenue, but Cleomenes nevertheless seems to have played a major part in the completion of the construction of Alexandria.

Alexander's sudden death in Babylon in June 323 altered everything. His campaigns of ten years had given rise to a number of generals who if nothing else were eager to enjoy their hard-won conquests; it took them a week to divide Alexander's empire among themselves. The division was intended to be amicable. The old general Perdiccas was to have served as regent for Alexander's yet unborn son and successor, Philip Arrhidaeus. But the generals, who for a number of years acted as satraps (governors) under the regency, behaved increasingly like independent dynasts. Civil war ensued, and soon the fiction of a unified kingship was abandoned. By 305 the generals had become kings in their own territories, and the empire built by Alexander had fallen into a number of independent states. Egypt fell to Ptolemy I, the son of an obscure Macedonian named Lagus.

The Ptolemaic dynasty. The history of Egypt in the next century was deeply influenced by these changes in the nature of Alexander's kingdom. When Alexander assumed control of Egypt, he had assumed also the trappings of divinity of the pharaohs; so too did Ptolemy after he proclaimed himself king. But though he and his successors played the part of Egyptian kings to Egyptians, their need for Greeks in administration and in the army kept them a part of the Hellenic world. A creative tension emerged between Hellenic and Egyptian cultures whereby each preserved its own characteristics but affected and changed the other.

During the 3rd century the country was wholly reorganized. Ptolemy I had fought a number of battles against other Macedonian generals in the 4th century, defeating Antigonus at Gaza in 312, and finally, at Ipsus in 301, in a battle which brought to an end the wars of the Diadochoi ("successors"). But the turmoil of these wars little changed the political picture that had obtained after the death of Alexander when Ptolemy took over Egypt from Cleomenes. By the end of the century Ptolemy still held Egypt securely, had extensive interests and influence in the Aegean, controlled Cyprus and Cyrene, and was quarrelling with his neighbour over control of Palestine.

Battle of Ipsus

Ptolemy II, 284–246. Ptolemy's son, Ptolemy II, was associated on the throne with his father in 284 and became sole ruler in 282. His reign saw the continuance of the first Ptolemy's foreign policy, although, with Alexander's generals now dead, struggles were now between the kings of the second generation. Ptolemy II engaged in a massive administrative and economic reorganization of Egypt. His father may have begun it, but it is from the time of Ptolemy II that there are extensive remains of papyri of official documents, private letters, royal decrees, and the like, that make it possible to read directly the evidence of the nature of the bureaucratic, social, and economic life of the country during the period.

Ptolemy II revived the custom of the Egyptians, followed later by many of his successors, of the king marrying his sister. His queen and full sister, Arsinoe II, was one of the most remarkable women of the Hellenistic Age. After being involved in several intrigues in Asia Minor, she returned to Egypt when the prospect of success, or even survival, outside seemed bleak and, with the determination typical of the women of this dynasty, supplanted her predecessor and became queen. She was in the fullest sense co-ruler. Each of the pair was known as Philadelphus, "brother-loving," "sister-loving," and, so far as is known, while Arsinoe lived she played a considerable, perhaps dominant, role in the formulation of royal policy. She was displayed on the coinage, and soon after her death (270 BC), if not before, she was worshipped as a goddess in the Greek ruler cult in Egypt.

Ptolemy II Philadelphus did much to advance the cultural and economic life of his kingdom. He promoted scholarship and literature, bringing to Alexandria such prominent writers as Theocritus and Callimachus. The museum called the Temple of the Muses in Alexandria was perhaps founded by Ptolemy I but was certainly active by the reign of his son; it was a centre of scientific and scholarly activity during his reign, and its attached library attracted great men of learning from all over the Greek world. It was during Ptolemy II's reign that the Greek Septuagint translation of the Old Testament was begun, and the occurrence of some peculiarly Ptolemaic administrative terminology suggests an Egyptian background as well as location for the translators. Philadelphus' activity also extended to administrative organization and agricultural innovation. New crops were introduced into the country, and new irrigation works opened up new agricultural lands.

Ptolemy III, 246–222. When Ptolemy II died in 246, he left the country in good condition to his son, Ptolemy III Euergetes I. Euergetes ("doer of good") received his name from his success in war against his eastern neighbour, Seleucus II of Syria. The war was occasioned by the murder of Euergetes' sister, Berenice, in Syria. One of the last acts of Ptolemy II had been the arrangement of the marriage of Berenice to Antiochus II of Syria. With the birth of a son to Berenice, Ptolemy II died expecting peace between the two realms and hoping to have his grandson king of Syria. The arrangement was disturbed by Laodice, the former queen of Antiochus, who murdered Berenice and the young prince and set her own son, Seleucus, on the throne. To save, and then avenge, Berenice, Euergetes marched into Syria; his campaign ended in victory, and he gained great popularity with the Egyptians with the recapture of the statues of their gods, originally taken by the Persians.

War with Syria

Euergetes spent the remainder of his reign in Egypt, losing some of his conquests in northern Syria and eventually signing a peace in Palestine. Like his father, he nourished culture and began the building of the great temple at Edfu, which still survives intact.

Ptolemy IV, 222–205. When Euergetes died (221), he was succeeded by his 23-year-old son Ptolemy IV Philopator. An unfavourable Greek tradition portrays Philopator as a corrupt and weak ruler, controlled by a court circle led by an Alexandrian named Sosibius. At the very beginning of his reign Egypt was unsuccessfully attacked by Antiochus III, who later launched a more determined attack that ended, however, in a great Ptolemaic victory at Raphia in southern Palestine in 217, a battle in which for the first time Egyptians, equipped and trained in the Macedonian fashion, fought beside Macedonian and Greek troops against a Hellenic enemy.

Ptolemy V, 205–180. Philopator died in 205 and left an infant son, Ptolemy V Epiphanes. The palace was under the influence of the court circle, which announced Philopator's death and Epiphanes' accession only in the following year. Soon, however, the Alexandrians rose against the court group and installed Tlepolemus, a soldier, as regent. But Tlepolemus had no success in holding back Antiochus III, who took advantage of Egypt's internal instability to expel Ptolemaic forces from the Levant and its Aegean dependencies. Tlepolemus' successor, Aristomenes, was successful for a while in pushing Antiochus' forces back in Palestine. Epiphanes' reign saw the growing strength of native revolts that had begun under his predecessors. By 197, when Epiphanes reached his majority at age 12, there seems to have been enough trouble, both internally and externally, to necessitate that Epiphanes celebrate his coming of age, in a bid for the support of the Egyptians. Honours were accorded the Egyptian religion, as recorded on the celebrated Rosetta Stone, during a religious synod in 196. Throughout this period Antiochus III pressed on Egypt from the east, while Philip V of Macedon seized Egyptian territory overseas. The Battle of Panium (*c.* 200) transferred Palestine to the Seleucids. A peace with Antiochus was concluded in 193 whereby Epiphanes married Antiochus' daughter Cleopatra, making Egypt something of a Seleucid protectorate. Ephiphanes had more success in internal affairs and by 183 had managed to suppress the remnants of native revolt. He died in 180 and was succeeded by his son Ptolemy VI Philometor.

Native revolts

Ptolemy VI, 180–145. Philometor, like his father, came to the throne a child, but he had the benefit of the firm rule of his mother, Cleopatra I, who governed the country with a strong hand until her death in 176. Her death was followed by a rupture between the Seleucid and Ptolemaic courts. Philometor married his sister Cleo-

patra II, and after some disturbances in Alexandria, which resulted in the reuniting of Philometor's younger brother with the reigning couple, the Egyptians sent an expedition into Palestine that was defeated by Antiochus IV. Antiochus returned the compliment, and by 169 he and his army were in Egypt, in control of part of the country and, in some way, of the person of Philometor as well. But Philometor escaped from the guard imposed by Antiochus, and when Antiochus returned with an army in the following year he faced a reunited trio.

He faced Rome as well. The Romans had been involved in the eastern Mediterranean for a long time, and their influence was growing. With the defeat of Philip V of Macedonia in 197, and his son Perseus in 168, Rome controlled the whole of Greece. After the defeat of Antiochus III at Magnesia in 190, Antiochus had withdrawn from conflict with Rome, and Rome's power was recognized throughout the East. The Ptolemies saw the advantage of friendship with Rome and placed themselves under its patronage. Thus when Antiochus IV returned to Egypt in 168, it was in line with Roman policy to oppose him. Gaius Popillius Laenas had been sent to Egypt, and as soon as he learned of the Roman victory over Perseus in 168 he knew it was safe to act. He ordered Antiochus to withdraw, and the latter's rapid compliance was a recognition of Roman dominance in the East. Thereafter, Roman protection of Egypt was to be a feature of international life as the Hellenistic kingdoms slid more and more toward impotence.

Friendship with Rome

Dynastic strife and the rise of Rome, 145–52 BC. The rest of the reign of Philometor, and much of that of his successor and brother, Ptolemy VIII Euergetes II, was marked by dynastic quarrelling. The association of Philometor, his wife, Cleopatra II, and Euergetes II lasted until 164, when Philometor was expelled; the solution imposed by the Romans gave Egypt to Philometor and Cyrenaica to Euergetes; he remained there until his brother died on campaign in Syria in 145. Philometor's attempt to have his young son Neos Philopator reign as Ptolemy VII was foiled by Euergetes, who returned to Egypt in that year to begin his long reign. Euergetes killed Neos Philopator and married Philometor's widow, Cleopatra II, who was sister to them both. Later he divorced Cleopatra II and married her daughter Cleopatra III. Disagreement between brother and sister persisted, and in 131 or 130 Cleopatra II was in formal revolt from Euergetes and was recognized as ruler in parts of Upper Egypt. By 118, after a long period of dissension, brother and sister made peace, and a long decree announced amnesty to the partisans of both sides.

Ancient historians portrayed Euergetes as a tyrant who massacred a number of Alexandrians and expelled the representatives of Hellenic culture. But monuments and papyri show him a capable administrator and a friend and patron of the native Egyptians. And, though detested by the Greeks, there is evidence that he was a supporter also of Greek culture.

Cleopatra II and her brother both died in 116. This left Cleopatra III, to whom Euergetes had bequeathed the throne, with two sons, Ptolemy, later to be known as Soter II, and Alexander. Cyrenaica had been given to Ptolemy Apion, an illegitimate son of Euergetes. Cleopatra wanted her younger and favourite son Alexander to share her throne, but it was Soter II who ruled with her as Ptolemy IX. In 110 Soter II was briefly supplanted by Alexander, but in 109 he returned, only to be expelled again in 107. Alexander ruled with Cleopatra III until her death in 101, and then with Cleopatra Berenice until his own expulsion in 88 by Soter, who ruled until his death in 81.

The years of these kings were marked by the growing power of Rome and the loss of more and more Ptolemaic independence. As early as the reign of Ptolemy VIII, Hellenistic kingdoms were falling to Rome: in 133, Attalus III of Pergamum bequeathed his kingdom to Rome, and in 96 Ptolemy Apion died, leaving the Romans Cyrenaica. In the latter kingdom the Romans left the cities independent for a while, but the area was formally annexed as a province in 74. After Soter's death, Cyprus became more or less independent and was eventually annexed by Rome in 58. Soter's son, who ruled as Ptolemy XII, succeeded to the throne in 80 after a rapid succession of unsuccessful claimants early in the year, and depended upon Rome for his tenuous hold on the throne during his life.

Increasing loss of independence

Ptolemy XII, nicknamed Auletes ("flute player"), was the last king to maintain the long-standing Ptolemaic policy of neutrality in the internal disputes of Rome. Soter II at the end of his reign had been able to keep Egypt out of the war between Rome and Mithradates, the strong king of Pontus in Anatolia. Auletes managed in general to avoid the struggle that centred on the Roman general Pompey's activities in the East. This did not mean that Egypt stayed out of the Roman sphere. On the contrary, in 58, after convincing Julius Caesar to recognize him as king in Egypt, Auletes was forced to go to Rome to plead his case and ask for support. He was away two years, during which time Cleopatra Tryphaena and his daughter Berenice IV took turns ruling. He spent so much money arranging support in Rome that when he was restored to the throne in 55 by one of Pompey's commanders, Aulus Gabinius, he had to bring his creditor, the banker Rabirius, back to Egypt to act as finance minister. The arrangement was probably struck to ensure that Rabirius would recover at least as much money as he had lent.

Cleopatra. In 52 Auletes associated with himself on the throne his daughter Cleopatra VII and his elder son Ptolemy XIII, both of whom succeeded him when he died the following year. The reign of Cleopatra VII, the famous queen of history and romance, was dominated by her struggle to survive her brother on the throne and by her attempts to join herself to the winning side in the civil wars of the late Roman Republic.

The arrival of Julius Caesar

Her first success was with Julius Caesar, when, after his victory over Pompey, he came to Alexandria in 48 in pursuit of his fleeing enemy. Caesar was trapped in the city by Egyptian forces who resented the Roman intruder. In control of the palace, the King, and Cleopatra, Caesar established a more than political relationship with the queen, so that after the siege was broken in 47 he established her firmly upon the throne with her younger brother, who ruled as Ptolemy XIV (47–44), Ptolemy XIII (51–47) having died.

Later in 47, Cleopatra gave birth to a son, who according to her assertion was the child of Caesar and to whom she gave the name Caesarion ("little Caesar"). In the following year Cleopatra followed Caesar to Rome and took up quarters there. With Caesar's support she was safely queen of Egypt and could come to Rome with her boy brother-king without risking trouble at home. Indeed, she bid fair as Caesar's consort to be queen of the Mediterranean world.

Her security ended abruptly with Caesar's assassination in 44. Cleopatra quickly got back to Egypt, in hope of riding out the storm that would surely come. For a while, she played the old Ptolemaic game of neutrality, trying to remain friendly to both the Caesarian and anti-Caesarian parties, hostile to neither. She did nothing to help Antony and his ally Octavian against Caesar's enemies at Rome until they had won, but when she met Antony at Tarsus in 41 he returned to Egypt with her, and a new liaison was begun. Antony left Egypt for campaign in 40, not to return until 36, in which year he signalled his affiliation with Cleopatra by giving her large tracts of the Levant.

In the following years a breach between Antony and Octavian opened. Antony's personal commitment to the East, staying as he did with Cleopatra, became useful propaganda against him in Italy that Octavian employed to build his strength for the coming showdown. It came in September of 31, in the naval battle at Actium in western Greece, where Cleopatra and Antony were both present. During the battle, Cleopatra and the Egyptian squadron left precipitately. Why, no one knew, in antiquity or today, but as Antony saw her ships pull away, he too left the battle, abandoning any possibility of victory. The two returned to Alexandria amid the trappings of pretended

triumph, while they were in reality only playing out the drama of defeat as they waited for Octavian, now master of the Roman world, to arrive.

On Octavian's arrival in August of 30, Antony stabbed himself, and Cleopatra is said to have committed suicide by the bite of an asp. Much as Octavian might have liked to exhibit the notorious queen at Rome, the last of the Ptolemies controlled her own end and died in Egypt. Though less successful than her ancestor, she had all the verve and vigour of that first Ptolemy who had commanded the troops of Alexander the Great and started the longest and most successful of the Hellenistic dynasties.

Government and civilization of Hellenistic Egypt. The multitude of papyri, both private and public, from the Ptolemaic period, has left a remarkably clear picture of the nature of everyday life. The absence of any connected ancient source makes knowledge of the period as a whole rather fragmentary, but the historian can get a better sense of the internal working of the country than he can for any other place in the ancient world. Fortunately, the documentation is particularly heavy for the mid-3rd century BC, when Ptolemy II was in the process of organizing the country's administration that served his successors so well.

Administration. The early Ptolemies had numerous problems. They needed Greeks for their armies and administration. Greeks who served in the army were assigned plots of land in Egypt to use for their support, and these soldiers, called cleruchs, initially formed the backbone of the legal, economic, and administrative system. Many took care of minor and major official duties in an economic system arranged to produce the highest possible revenue for the crown. This administrative structure was quite elaborate. There were parallel bureaucracies, each with its own hierarchy, for the various departments of government: a department of revenue oversaw the collection and transfer of taxes to the crown; a department of land administration assigned acreage for seeding of various crops and saw to the proper physical operation of agriculture; and a clerical department kept records of landholding, sowing assignments and requirements, and the assessment and payment of taxes. These three departments operated at various levels: at the level of the nome, the country's major subdivision (of which there were about 40), the *oikonomos* supervised revenue administration, while the *nomarchēs* supervised the land, and the *basilikos grammateus* (royal scribe) was in charge of records. Junior to them were assistants to the *oikonomos*, juniors to the *nomarchēs* called *kōmarchēs* (in charge of villages), and village scribes. A *strategus* (general) not only commanded the military in each nome but also, and more importantly, administered justice. Above all these was an official in Alexandria called the *dioikētēs* (finance minister), who directed economic activities for the entire country.

The parallel bureaucracies of Hellenistic Egypt

Such was the theoretical organization of the bureaucracy, but it is clear from the papyri that the tripartite structure did not always work as neatly in practice. There were not always enough officials, and the complexity of administration in some areas required a more articulated division of authority. As a consequence, there were more divisions and more levels in some areas than the structure seems to have called for, and fewer in others. All in all, however, it seems to have worked well.

Both Greeks and Egyptians were employed in bureaucracy. Even at the highest ranks, and in the earliest period of Ptolemaic control, native Egyptians served the king, as attested by Egyptian-style statues with the Egyptian names and ranks of their dedicators. The papyri indicate also numerous Egyptians in the lower levels of government.

Economy. The economic system established by the Ptolemies was a mixture of royal monopolistic control and private enterprise. Royal revenues came both from ordinary taxes and from the proceeds deriving from exploitation of the monopolies. A long papyrus of the 3rd century BC preserves many of the regulations imposed by Ptolemy II to ensure that the oil monopoly brought him maximum revenues. Fragments of the papyrus preserve references to other royal enterprises: the production of textiles, the manufacture and sale of beer (the national drink of the Egyptians), and the distribution of salt. What was not monopoly was taxed. The crown imposed taxes at different rates, depending upon the product, so that there was a special tax, for example, on vine and orchard lands and another on grain-producing lands.

Sources of royal revenue

While it is not possible to determine the direction of expenditure of these revenues, it is clear that, for the 3rd century at least, a good deal of money must have gone to maintain the large Ptolemaic military establishment. Within the country itself, some of the money went for the administration, although in such circumstances that the expenditure for staff was undertaken in pursuit of revenue. There are directions in a papyrus detailing the revenue laws for the exploitation of the oil-bearing plants, and as part of the instructions the king specifies the salaries to be paid for the various tasks. Many papyri mention salaries to be paid to underlings, and although many of these seem only quasi-official or even unofficial personages, in many cases the payments come from the royal account or from crown officials. Even the tax farmers who collected the oil tax were paid by salary rather than out of the profit of excess collection—the method favoured by the Romans.

Role of banking in economy

The economy itself was highly sophisticated and included a number of specialized occupations. Bankers, for example, served both the king and private persons. The royal bank, with branches throughout the country, served in a way as a treasury, receiving all payments due the king and making official disbursements on proper authority. At the same time, the banks carried on all kinds of private business, such as making loans, accepting deposits, and making payments on private accounts. All this was carried on with written instructions and drafts, so that the system bore something of a modern character. In addition to banking services, there was extensive use of private credit, and the papyri preserve many loan transactions. Land itself was the object of many transactions, and even though in theory ownership of much of it was vested in the king, there are records of many leases of land as well as of houses and even parts of houses.

Religion. Throughout the Ptolemaic period the relations between the king's officials and the priestly bureaucracy created problems. From the beginning, the temple land seems to have been managed by government officials, much as royal land was managed. Out of the revenues, and perhaps in some instances in addition to the revenues, the government paid the expenses of the temples. This close economic involvement between state and religious enterprises illustrates both how closely the Ptolemaic system supervised important activities and how carefully the Ptolemies maintained the religious practices of Egypt.

The Greeks had always respected the antiquity of Egyptian religion, and individual Greeks were fully prepared to worship Egyptian deities as they had Greek; so that in Egypt, as in much of the rest of the Hellenistic world, the growing cults of Isis, for example, mark the spread of mystery, salvation, and healing religions three centuries before Christ.

The Ptolemies did nothing to discourage Egyptian cults, although from time to time there seems to have been some attempt to restrict the political influence of the priesthood. From the beginning of their dynasty, the Ptolemies built new temples in the Egyptian style and fostered the worship of Sarapis, a god encountered by the Greeks perhaps as early as Alexander's founding of Alexandria, and under the Ptolemies represented as a Hellenic deity, often as Zeus. The centre of his worship was the new Ptolemaic capital of Alexandria, where a number of Sarapeums were supported by the Ptolemies with an elaborate priesthood.

Worship of Sarapis

The Ptolemies too were worshipped as gods, as had been the pharaohs before them, and Egyptian inscriptions refer to Ptolemies in the same language that had been addressed to pharaohs of native descent. New to Egypt, however, was a formalized royal cult, more Greek than

Egyptian, dedicated to the worship of Alexander the Great and of each of the Ptolemies. By the end of the period, the number of priests in the cults had swelled to serve the many divinities of the dynasts who had reigned. Inscriptions record the synods of these priests meeting to issue decrees in the service of the reigning monarchs, and it is easy to see that the royal cult could serve political as well as religious and psychological needs.

Law. Although there was a blending of Greek and Egyptian religion, Ptolemaic law seems to have kept the two strains of culture apart. Egyptian custom received the status of law by royal decree of the Ptolemies. Egyptians were judged by their native judges, and their contracts were judged according to native custom. Greeks wrote Greek contracts to be assessed in different courts. In general, the language of the contract determined the court of resort in civil law. Much law, however, was by administrative decision, and, in an economic system strongly controlled by the monarchy, much private activity and many private disputes were resolved by bureaucrats rather than by judges.

Culture. Greek culture flourished under the Ptolemies. Many famous writers came to Alexandria, particularly during the reigns of the first Ptolemies, and the city became a centre of writing and culture. Major poets, such as Theocritus, Callimachus, and Apollonius of Rhodes, took up residence under the patronage of Ptolemy II and formed a literary school, though they did not always like each other's work. Callimachus, for example, one of the most esteemed writers of the ancient world, referred disparagingly to the lengthiness of the poems of Apollonius. The original literary product of Alexandria was considerable; much of the writing in Greek of the Hellenistic period was done there. In addition, a school of scholarship was active (from the mid-3rd century on) that is probably responsible for the preservation of a great deal of earlier Greek writings. Such scholars as Zenodotus and Aristarchus, famous for their work on Homer, made new editions, corrected and refined manuscripts, and wrote commentaries on the great writings of the past. The Library of Alexandria collected all the manuscripts it could, while throughout the period new texts were made and sold to Greeks throughout the country.

Alexandrian poets

Alexandria was also a centre for scientific research. The great mathematician and astronomer Eratosthenes did his work there in the 3rd century, serving as librarian and carrying on his experiments. The school of mathematics at Alexandria is said to have been established by Euclid. As the home of so many scholars, philosophers, and literary figures, Alexandria became one of the great centres of education and culture, rivalling Athens, and it retained its importance and status long after the end of the Ptolemaic dynasty.

The Ptolemies endured through the great shift in the centre of power in the Mediterranean world from Greece and the East to Rome and the West. The security of Egypt permitted Greek culture to flourish, and the Greeks in Egypt developed and expanded. Although it has been common to see in the Hellenistic period, particularly in Egypt, a blending of cultures, and although that seems to have occurred to some extent in religious affairs, the two cultures in fact merely coexisted. The Ptolemies exploited the resources of the country and promoted Greek culture, but they allowed native institutions to continue, not only in religious affairs but in law, literature, and custom. There seems to have been no conscious "racial" discrimination against Egyptians, and there was much intermarriage. There does, however, seem to have been a deliberate policy to maintain the Greek language in the administration of the country. Egyptians who chose to learn Greek could rise in government, but then they and their descendants became Hellenized. The language policy may have been merely administrative convenience, but it must have contributed, as a side effect, to the great Alexandrian efflorescence of Greek studies.

Interaction of the two cultures

ROMAN EGYPT (30 BC–AD 642)

The Egypt taken over by Octavian after the suicide of Cleopatra in 30 BC was a country that, despite some internal struggles and dynastic quarrels, was still relatively well organized, and still wealthy. With the introduction of a new stability under the Romans, Egypt had great promise of prosperity and profit. In the century before it became the property of Rome, Egypt had been carrying on extensive trade with the West; and even though the Ptolemies had been forced on occasion to borrow from Roman bankers to maintain themselves against other claimants to the throne, Cleopatra had been able to lend financial support to Antony's adventures and supply a fleet in his final attempt to defeat Octavian. Much of this wealth must have come from shipments of grain to Italy. Egypt had long produced a grain surplus, and the land remained fertile and productive despite political unrest and international upheaval. As Italy became more urbanized and populous, and required increasing imports of grain, Egypt became more and more important as a reliable source of its food. Octavian (later known as Augustus) saw that the stability of his new empire depended, to some extent, upon the security of Egypt and its grain supply and upon the assurance that the supply would remain controlled by his government. He thus treated Egypt differently from other Roman provinces. The country was treated as his personal property, just as the country, the people, and all its resources had been the personal property of the Ptolemies. No Roman of senatorial class was allowed to enter Egypt without the explicit permission of the emperor. This reduced the danger that adventurers challenging imperial rule might fasten upon Egypt and its grain supply to control events in Italy.

Egypt's role as the granary of Rome

But Egypt was a special province for other reasons. Just as in Hellenistic times Egypt differed from its neighbouring states in its centralized bureaucracy and royally directed economy, so in Roman times it differed from other provinces in that it had greater geographic and administrative cohesion. In other provinces of the Roman Empire, administration seems to have been centred in the major cities, which seem to have continued as "free" states, governed under their own laws and impinged upon only as the Roman provincial governors saw need to provide for security for the continued collection of taxes. The tax collections in these provinces were undertaken by companies of private Roman citizens, supported by governors and armies. Egypt, on the other hand, continued to be governed by its Ptolemaic bureaucracy, and although more and more tasks came to be undertaken as "voluntary" duties by residents of the countryside and towns, government, through the prefect of Egypt, an agent of the emperor, was direct and pervasive. And in Egypt, the emperor was king, pharaoh, god, just as had been his predecessors, the Ptolemies.

Nevertheless, the influence of Egypt must have been considerable, particularly as this form of government was a success. Egypt by and large was one of the more stable parts of the empire, and although there were occasional revolts beginning there, or political movements to which the governors of Egypt adhered, more often than not the armies in Egypt did not initiate changes in emperors but rather joined the side of the successful claimant. And as the bureaucracy of Egypt continued to supply money and material through changes of imperial house and through convulsions of empire, the merits of a standing bureaucracy became more and more obvious. Egypt thus served to some extent as an impulse toward the growth of Roman bureaucracy.

Roman administration. Egypt was administered directly by the imperial house through a prefect of equestrian, or knightly, class appointed by the emperor. Emperors seldom visited the province, and visits by the imperial family were discouraged for the same reason as visits by those of senatorial rank. The emperor used servants and prefects rather than dangerous nobility to communicate with Egypt and carry imperial decisions; the emperor Tiberius reprimanded his nephew Germanicus for an unauthorized visit in AD 19. When residents of Egypt or Alexandria required direct audience with the emperor, they had to go to Rome, as did representatives of the citizens and Jews of Alexandria, petitioning against

each other about the recurring disputes and disorders that arose between them. The visit of the emperor Septimius Severus to Egypt in AD 199 or 200 was exceptional.

Duties of the Roman prefect

The Roman prefect headed the bureaucracy of the country; he commanded the army, dispensed justice, administered taxes, and directed the economy as a whole. Though much of the lower bureaucracy seems to have continued under the Romans, at least in name, a fundamental change in fact came over the country. It has been conventional to speak of Greco-Roman Egypt as though there were a continuum (and in some areas of life there was such); but in most matters affected by government, the advent of Roman control brought significant changes.

The most obvious changes were in law. Although the practices of Greek law, particularly in the private sphere, seem to have continued, the administration of that law was different. Jurisdiction was centred in the office of prefect, who was assisted by an official called the *juridicus*. The Ptolemaic courts were eliminated or radically changed; under the prefect, Egypt was divided into three judicial districts or assizes (*conventus*), where cases were heard and decisions rendered. For the country as a whole, edicts were issued to establish specific practices, and these edicts established precedents. Surviving records contain judgments and appeals citing particular decisions or directives by former prefects. Roman law was in effect for Roman citizens in Egypt, and in the Greek city of Alexandria the Roman government supplanted the law courts of Ptolemaic times.

City government

Egyptian cities also received different treatment by the Romans. Alexandria, which may have had a senate in Ptolemaic times, petitioned the emperor Claudius (ruled 41–54) for the establishment of one. The petition was denied, so that no such institution existed there in the first centuries of Roman rule. Septimius Severus, however, established senates (*boulai*) for all the Greek cities and for the metropolitan cities of the nomes as well. The establishment of a Greek form of government for the cities conformed to the Roman practice of relying upon local administration and magistracies insofar as possible, even though in Egypt those cities must have been less autonomous than they were in the less organized provinces of the empire. At the time of Augustus' conquest, there were only three autonomous Greek cities in Egypt, the old Greek *polis* of Naukratis, along with Alexandria and Ptolemais. In 130 the emperor Hadrian established a new city, Antinoöpolis, named after his friend Antinoüs, which was lavished with privileges and set up with all the institutions characteristic of a Greek city, including a *boulē*. But the other cities, although they had many of the magistracies associated with the autonomous city in the time of Greek independence, were not to have their *boulai* until they were set up for all cities, including nome capitals, by Septimius Severus at the beginning of the 3rd century.

The new *boulai* illustrate well the Roman system of local administration in Egypt. These councils (and their members) were responsible for carrying out, each within its own city, the duties of administration and the collection of taxes. They were responsible to the central bureaucracy and to superior officials in this, but they were to use their own resources in carrying out their tasks. This was the system of "liturgies": individual Egyptians were nominated by the civil authority to carry out administrative responsibilities and were designated to hold offices for fixed periods, the nominator's property and that of the "liturgist" serving as guarantee in case of failure. The *boulai* were responsible for nomination to offices, thus placing in local hands the unpopular task of distributing burdens. The system worked well for a while and gave the Romans a maximum return from Egypt, as the costs of running the country and the collection of taxes were paid by the residents, not by the Roman administration.

"Liturgy" system

Another characteristic of Roman administration was the introduction of a clear-cut and formalized classification of population. In Ptolemaic times the government recognized different ethnic backgrounds and provided for distinctions and different treatments for them under the law. But there was little formal attempt to keep the distinctions permanent. The Romans, on the other hand, made a very strong attempt to formalize and control the ethnic groupings, particularly as they affected Roman citizenship. The so-called *Gnomon*, or consolidation of rules, issued in the 2nd century by the office of the *idios logos* (controller), makes it clear that changes of status were discouraged, that cross-inheritance was limited, that Romans were not allowed to make Greek wills, and that Egyptians were not to be granted citizenship by serving in the legions. The manifold regulations of the *Gnomon* describe a situation in which citizenship, whether Roman or in the "Greek" cities, was protected and guarded as a privilege.

Throughout the first two centuries of Roman rule, Roman citizenship, in Egypt as elsewhere, could be granted a non-Roman either as a reward or as a result of serving in the legions. In 212, by the imperial decree known as the Constitutio Antoniniana de Civitate, better known as the Edict of Caracalla, the citizenship was broadened, but its interpretation remains problematical. It appears, however, that citizenship was granted to all but certain specified groups, and it is clear that citizenship was taken up extensively throughout Egypt after the issuance of that decree.

Although changes in detail of administration in Egypt can be traced in the first two centuries, down to the accession of Septimius Severus in 193, the systems established by Augustus and his Julio-Claudian successors continued to function without fundamental modification. By the time of Severus, however, the structure was showing signs of strain, and Severus' changes were aimed at shoring up an administrative structure that was under stress, although still working. By granting municipal councils to the chief city of each nome and to the Greek cities, Severus altered the bureaucratic structure without changing the nature or types of taxes collected, and he maintained the position and nature of Egypt as a single province in the empire without modification of the higher offices in the administration in Egypt. His change was aimed at throwing more of the responsibility onto the residents of the cities.

The end of Roman rule. By the end of the 3rd century the strain of political and economic unrest told not only upon Egypt but upon the whole Roman Empire. A succession of short-lived military adventurers holding power at Rome could not meet the needs of a complex empire; and the recurrent struggles for power had themselves contributed much to the troubles of government. The need for reform was met with the accession of Diocletian as emperor in 284. He was something of a genius at administration, and in his reorganization of the empire he broke completely with tradition. He divided the empire into East and West, each under a reigning Augustus and an heir apparent called a Caesar. All provinces were reorganized and given completely new bureaucracies, and many had their boundaries redrawn. Within the provinces, local administration was radically altered to fit a new concept of taxation, wherein each year each province was allotted its "indiction," an estimate based on 15-year assessments. The intention of the fiscal reform, based as it was on estimates of productivity, was to distribute the burden of taxation more fairly, and Diocletian hoped through his reforms to stabilize the capitation and land taxes and to rely less on the special taxes, which had been a particular burden in the years before his accession.

Reforms of Diocletian

Egypt itself was divided into three provinces—the Thebaid, Aegyptus Herculia, and Aegyptus Jovia—each under a governor called a *praeses*. The prefect of Egypt continued to function, but his responsibility was limited to the civil administration of the northern province, Aegyptus Jovia, including Alexandria. Military authority was invested in a new official called the *dux*, or duke, of Egypt. Another major change occurred soon after Diocletian's abdication in 305. Within the country, the administration was changed from the structure based on nomes, each with its capital and central administrators, to a structure based on self-governing municipalities. Each city was assigned a territorial district, and that district was divided into numbered sections, each called a *pagus*,

with each pagus under the administration of an official called *praepositus*. Administrative duties centred on the cities themselves, and thence extended to the higher reaches of government. Latin replaced Greek as the language of the upper bureaucracy.

Growth of Christianity

In these years and in the years following, even more changes took place. Christianity was spreading throughout the empire, and, despite attempts during the reign of Diocletian to suppress it, by the mid-4th century Egypt seems to have become predominantly Christian, and the religion had penetrated throughout the native population. The native Egyptian language, Coptic, newly rendered into a script based on the Greek alphabet, transmitted the Bible and the writings of the Christian Fathers to the peasants. The granting of tolerance in 313 to the Christians by the emperor Constantine I (306–337) gave impetus to the development of a formal Egyptian church that soon vied with that of Constantinople. The Orthodox Alexandrian patriarchs were often at odds with large segments of the Egyptian population, which tended to favour heterodox Christian teachings (*e.g.*, Arianism and Monophysitism). Throughout the countryside the growing enthusiasm for anchoritism, the hermit life, took vigorous and intellectual men into the desert.

Consolidation of large estates

Meanwhile, the economic structure of the country underwent a harsh transition in the 4th century. Large estates became concentrated in the hands of wealthy families throughout the country. Smaller landholders often turned to these powerful families for protection, and turned over their small holdings to them in return for the right to farm the land as lessees. This was the solution to the growing pressure of taxation, which large landholders could meet more easily than the small. With this transfer of land, the rich became more powerful, and despite the attempts of the state to restrict the trend the great estates grew apace, until they were officially recognized in the early 5th century.

The people themselves were officially bound to the land. In 415 a government decree restored vagrant tenants (*coloni*) to their masters and villagers to their towns. The restriction on freedom of movement and occupation characteristic of the empire as a whole was enforced in Egypt, and the extensive agricultural value of the countryside made it a useful adjunct to the exploitation of the large estates. This contributed to the growth of the power of the landowners, so that these great nobles became more and more part of the government structure, as it was through them that taxes were collected and law enforced.

Momentous events would prevent this system from evolving into the feudal structure of the West. Although Constantine reunited the empire at the beginning of the 4th century, it was divided permanently into East and West after the death of Theodosius in 395. Egypt was thenceforward administered from Constantinople. Under this administration the strength of Christianity and the great nobles increased. Byzantine Egypt was the Egypt of Christianity, monasticism, Coptic culture, and the gulf between wealthy noble and poor farmer. This was the structure against which the growing power of the Sāsānid Persians pressed. Justinian, in the mid-6th century, reorganized Egypt into four parts, each under a governor with both military and civil powers. The government, however, could not resist the attacks of hostile powers. In 616 the Persians invaded and conquered Egypt, holding it until the death of their king in 628, when, by terms of a treaty, Egypt reverted to Byzantium under Heraclius I. But this did not last. The armies of Islām swept out of Arabia, first against the Persians, then against Egypt and the West. In 639 the army of ʿUmar invaded Egypt, and after a series of campaigns Alexandria was evacuated by the Byzantine army in 642. Egypt was now part of the Muslim world, and transmitted to Islāmic scholars the traditions and science of Hellenism, on which the Muslims themselves built and expanded. (A.E.S.)

IV. Egypt from c. 630 to c. 1800

Medieval Egyptian history opens and closes with foreign conquests of Egypt: the Arab invasion led by ʿAmr ibn al-ʿĀṣ in 639 and the Napoleonic expedition of 1798 mark the beginning and end of an era. Within the context of Egyptian internal history alone, this era was one in which Egypt cast off the heritage of the past to embrace a new language and a new religion—in other words, a new culture. While it is true that the past was by no means immediately and completely abandoned and that many aspects of Egyptian life, especially rural life, continued virtually unchanged, it is nevertheless clear that the civilization of Islāmic Egypt diverged sharply from that of the Greco-Roman period and was transformed under the impact of Western occupation. The history of medieval Egypt is therefore largely a study of the processes by which Egyptian Islāmic civilization evolved, particularly the processes of Arabization and Islāmization. But to confine Egyptian history to internal developments is to distort it, for during the entire medieval period Egypt was a part of a great world empire; and within this broader context, Egypt's history is a record of its long struggle to dominate an empire—a struggle that is not without its parallels, of course, both in ancient and modern times.

FROM THE ISLAMIC CONQUEST TO 1250

Period of Arab and Turkish governers (639–868). The sending of a military expedition to Egypt from the caliphal capital in Medina came in a second phase of the first Arab conquests. Theretofore the conquests had been directed against lands on the northern borders of Arabia and were in the nature of raids for plunder; they had grown in scale and momentum as the Byzantines and Persians put up organized resistance. By 635 the Arabs had realized that in order to meet this resistance effectively they must begin the systematic occupation of enemy territory, especially Syria, where the Byzantine army was determined to halt the Arab forays.

The Arab expedition to Egypt

The Arab conquest. The Arabs defeated the Byzantines and occupied the key cities of Syria and Palestine, and they vanquished the Persian army on the eastern front in Mesopotamia and Iraq. The next obvious step was to secure Syria against a possible attack launched from the Byzantine province of Egypt. Beyond this strategic consideration, Arab historians call attention to the fact that ʿAmr ibn al-ʿĀṣ, who later conquered Egypt, had visited Alexandria as a youth and had himself witnessed Egypt's enormous wealth. In spite of the obvious economic gain to be had from conquering Egypt, the caliph ʿUmar, according to some sources, showed reluctance to detach ʿAmr's expedition from the Syrian army and even tried to recall the mission once it had embarked; but ʿAmr, with or without the caliph's permission, undertook the invasion in 639 with a small army of some 4,000 men (later reinforced). With what seems astonishing speed the Byzantine forces were routed and had withdrawn from Egypt by 642. An attempt by a Byzantine fleet and army to reconquer Alexandria in 645 was quickly defeated by the Arabs.

Various explanations have been given for the speed with which the conquest was achieved, most of which stress the weakness of Byzantine resistance rather than Arab strength. Certainly the division of the Byzantine government and army into autonomous provincial units militated against the possibility of a concerted and coordinated response. Although there is only dubious evidence for the claim that the Copts welcomed the Arab invasion in the belief that Muslim religious tolerance would be preferable to Byzantine enforced orthodoxy and repression, Coptic support for their Byzantine oppressors was probably unenthusiastic at best.

Arab policies

Early Arab rule. In Egypt—as in Syria, Iraq, and Iran—the Arab conquerors did little in the beginning to disturb the status quo; as a small religious and ethnic minority, they thus hoped to make the occupation permanent. Treaties concluded between ʿAmr and the *muqawqis* (presumably a title referring to Cyrus, archbishop of Alexandria) granted protection to the native population in exchange for the payment of tribute. There was no attempt to force, or even to persuade, the Egyptians to convert to Islām; the Arabs even pledged to preserve the

Christian churches. The Byzantine system of taxation, combining a tax on land with a poll tax, was maintained, though it was streamlined and centralized for the sake of efficiency. The tax was administered by Copts, who staffed the tax bureau at all but the highest levels.

To the mass of inhabitants, the conquest must have made little practical difference, because the Muslim rulers left them alone, in the beginning at least, as long as they paid their taxes; if anything, their lot may have been slightly easier, because Byzantine religious persecution had ended. Moreover, the Arabs deliberately isolated themselves from the native population, according to ʿUmar's decree that no Arab could own land outside the Arabian Peninsula; this policy aimed at preventing the Arab tribal armies from dispersing and at ensuring a steady revenue from agriculture, on the assumption that the former landowners would make better farmers than would the Arab nomads.

As was their policy elsewhere, the conquerors refrained from using an established city such as Alexandria as their capital; instead, they founded a new garrison town laid out in tribal quarters. As the site for this town they chose the strategic apex of the triangle formed by the Nile Delta—at that time occupied by the Byzantine fortified township of Babylon. They named the town Fusṭāṭ, which is probably an Arabized form of the Greek term for "encampment" and gives a good indication of the nature of the earliest settlement. Like garrison towns founded by the Arabs in Iraq—Basra and Kūfah—Fusṭāṭ became the main agency of Arabization in Egypt inasmuch as it was the only town with an Arab majority and therefore required an extensive knowledge of Arabic from the native inhabitants.

The process of Arabization, however, was slow and gradual. Arabic did not displace Greek as the official language of state until 706, and there is evidence that Coptic continued to be used as a spoken language in Fusṭāṭ. Given the lack of pressure from the conquerors, the spread of their religion must have been even slower than that of their language. A mosque was built in Fusṭāṭ bearing the name of ʿAmr ibn al-ʿĀṣ, and each quarter of the town had its own smaller mosque. ʿAmr's mosque served not only as the religious centre of the town but also as the seat of certain administrative and judicial activities as well.

Although Alexandria was maintained as a port city, it should not be forgotten that Fusṭāṭ, being built on the Nile bank, was itself an important port and remained so until the 14th century. ʿAmr enhanced its commercial significance by clearing and reopening Trajan's Canal, so that shipments of grain destined for Arabia could be sent from Fusṭāṭ to the Red Sea by ship rather than by caravan.

Egypt under the caliphate. For more than 200 years —that is, throughout the Umayyad caliphate and well into the ʿAbbāsid—Egypt was ruled by governors appointed by the caliphs. As a province in an empire, Egypt's status was much the same as it had been for centuries under foreign rulers whose main interest was to supply the central government with Egyptian taxes and grain. In spite of evidence that the Arab governors tried in general to collect the taxes equitably, taking into account the capacities of individual landowners to pay and the annual variations in agricultural yield, resistance to paying the taxes increased in the 8th century and sometimes erupted into rebellion in times of economic distress. Periodically, religious unrest was manifested in the form of political insurrections, especially in those exceptional times when a governor openly discriminated against the Copts by forcing them to wear distinctive clothing or, worse, by destroying their icons. Still, the official policy, especially in Umayyad times, was tolerance, partly for fiscal reasons. In order to maintain the higher tax revenues collected from non-Muslims, the Arab governors discouraged conversion to Islām and even required those who did convert to continue paying the non-Muslim tax. New Christian churches were sometimes built, and the government took an interest in the selection of patriarchs.

Resistance to taxation

More than just a source of grain and taxes, Egypt also became a base for Arab-Muslim expansion, both by land and sea. The former Byzantine shipyards in Alexandria provided the nucleus of the Egyptian navy, which between 649 and 669 joined in expeditions with the Syrian navy against Rhodes, Cyprus, and Sicily and defeated the Byzantine navy in a major battle at Phoenix in 655. By land, the Arab armies advanced both to the south and to the west. As early as 651–652 the governor of Egypt invaded Nubia and imposed a treaty that required the Nubians to pay an annual tribute and to permit the unmolested practice of Islām in the province. Raids against North Africa by Arab armies based in Egypt began in 647; by 670 the Arabs had succeeded in establishing a garrison city in Ifrīqīyah (now Tunisia), called al-Qayrawān (Kairouan), which thenceforth displaced Egypt as the base for further expansion.

While some Arabs were passing through Egypt on their way to campaign in North Africa, others were being sent to the Nile Valley on a permanent basis. In addition to tribal contingents that at times escorted newly appointed governors to Egypt (some of which settled in towns), tribesmen were sometimes imported and settled in an effort to increase the Arab-Muslim concentration in the vicinity of Fusṭāṭ. The settlement of large numbers of anarchic tribesmen in Egypt, with tribal ties and allegiances elsewhere in the empire, meant that Egypt became embroiled in political difficulties with the central government. Civil strife centring around the assassination of the caliph ʿUthmān (656) began in Egypt, where the tribesmen resented the favouritism shown by the caliph to members of his own family. Uprisings led by the dissident Khārijite sect (the Seceders) were frequent in the mid-8th century. In the 9th century the caliph Maʾmun himself led an army from Iraq to put down a rebellion raised both by tribesmen and by Copts; repression of the Copts accompanying their defeat in 829–830 is usually cited as an important factor in accelerating conversion to Islām.

Civil strife

The difficulty inherent in ruling Egypt from Baghdad, which was itself undergoing stress and turbulence, is evident from the rapid turnover in governors assigned to Egypt; the ʿAbbāsid caliph Hārūn ar-Rashīd (ruled 786–809), for example, appointed 24 governors in a reign of 23 years. Possibly as a means both of removing the governorship from the level of tribal strife and paying the central government's Turkish troops, the caliphs began assigning Egypt to Turks rather than to Arabs. But this policy resulted in no tangible improvement in the administration of Egyptian affairs until 868, when the reign of Aḥmad ibn Ṭūlūn inaugurated a new phase of medieval Egyptian history.

The Ṭūlūnid dynasty (868–905). Though short-lived, the Ṭūlūnid dynasty succeeded in restoring a measure of Egypt's ancient glory. For the first time since the pharaohs, Egypt became virtually autonomous and the bulk of its revenues remained within its borders. What is more, Egypt became the centre of a small empire when Ibn Ṭūlūn conquered Syria in 878–879. These developments were paralleled in other provinces of the ʿAbbāsid Empire and were the direct result of the decline of the caliph's power. In order to strengthen their armies, the ʿAbbāsid caliphs had begun early in the 9th century to form contingents of Turkish slaves. To finance these new military formations and, in particular, to pay the Turkish commanders who headed them, the caliphs began to give them administrative grants (*iqṭāʿ* in Arabic, usually translated "fief") consisting of tax revenues from certain territories. In 868 Egypt was granted as a fief to the Turkish general Babak, who chose to remain in Iraq but appointed his stepson, Aḥmad ibn Ṭūlūn, as his agent in Egypt. Ibn Ṭūlūn's great achievement was that he quickly established his own authority in Egypt and backed it up with an army of his own creation, powerful enough to defy the central government of Baghdad and to embark upon foreign expansion.

Ibn Ṭūlūn's first step was to eliminate possible rivals in Egypt. From an early date the administration of Egypt had been divided between the *amīr* (military governor), appointed by the caliph, and the *ʿāmil* (fiscal officer), who

was sometimes appointed by the caliph, sometimes by the governor. When Ibn Ṭūlūn entered Egypt in 868 he found the office of *ʿāmil* filled by one Ibn al-Mudabbir, who over a period of years had gained control of Egyptian finances, enriching himself in the process, and was therefore reluctant to acknowledge Ibn Ṭūlūn's authority. A struggle for power soon broke out between the two, which ended four years later with the transfer of Ibn al-Mudabbir to Syria and the assumption of his duties and powers by Ibn Ṭūlūn. An even more important step was the acquisition of an army that would be independent of the caliphate and loyal to Ibn Ṭūlūn. To build such an army, Ibn Ṭūlūn resorted to the same method the caliphs themselves used—the purchase of slaves who could be trained as military units loyal to their owner.

In 877, when Ibn Ṭūlūn failed to pay Egypt's full contribution to the ʿAbbāsid campaign against a black slave uprising in Iraq, the caliphal government, dominated by the caliph's brother al-Muwaffaq, realized that Egypt was slipping from imperial control. An expedition dispatched by al-Muwaffaq to remove Ibn Ṭūlūn from the governorship failed. Taking advantage of the caliphate's preoccupation with the revolt, Ibn Ṭūlūn in 878 invaded Syria, where he occupied the principal cities and garrisoned them with his troops. Thereafter he signified his autonomy by imprinting his name on the coinage along with the name of the caliph. Although the regent al-Muwaffaq lacked the resources to engage Ibn Ṭūlūn in battle, he did have him publicly cursed in the mosques of the empire as a means of retaliation.

Autonomy under Ibn Ṭūlūn

Internally, Ibn Ṭūlūn took active measures to raise Egyptian agricultural productivity and thereby to increase tax revenues; the huge surplus he left in the state treasury at his death in 884 is a measure of his success. Another tangible indication of his achievement for Egypt is an enormous mosque (still standing) that he erected in a suburb of Fusṭāṭ; in contrast, no building comparable in grandeur had even been contemplated by the governors who preceded him.

The great benefits Ibn Ṭūlūn had gained for Egypt by using its resources within the country were squandered by his son and successor, Khumārawayh. He expended huge sums on luxurious appointments for his residence and paid a fortune as a dowry for a daughter he married to the caliph al-Muʿtaḍid in 895. Nevertheless, Khumārawayh was able to maintain the Egyptian armies in the field, and he led them to victory both in Syria and in Mesopotamia. He resolved his father's conflict with the caliphate by a combination of arms and diplomacy, so that Khumārawayh's authority over Egypt, Syria, and Mesopotamia was given official caliphal recognition. This apparent strength evaporated when Khumārawayh was murdered in 896, leaving no funds with which his heir, a 14-year-old youth, could pay the troops. Both Egypt and Syria fell into anarchy, which lasted until 905 when a caliphal army invaded Egypt and momentarily restored it to the status of a province ruled by governors sent from Baghdad.

The Ikhshīdid dynasty (935–969). For 30 years the governors were unable to restore stability in Egypt. During this time, Egypt was subjected to attacks from the Fāṭimid state based in North Africa and to the rampages of an unruly domestic army. The appointment of Muḥammad ibn Ṭughj, from Sogdiana in Central Asia, as governor in 935 led to a repetition of Ibn Ṭūlūn's achievement; by bold measures Muḥammad established his authority over the treasury and the army, reasserted Egyptian influence in Syria, and won the governorship of the Holy Cities of Arabia (Mecca and Medina). In addition, he founded a dynasty; his sons inherited his Sogdian princely title of *ikhshīd*, but their authority was usurped by their Abyssinian slave tutor, Kāfūr, who ruled Egypt with the caliph's sanction. When Kāfūr died in 968 the Ikhshīdids were unable to maintain order in the army and the bureaucracy. In the following year the Fāṭimids took advantage of the disorder in Egypt to launch still another attack, this one so successful that it led to the occupation of the country by a Berber army led by the Fāṭimid general Jawhar.

The Fāṭimid dynasty (969–1171). The establishment of the Fāṭimid caliphate in 973 in the newly built palace city of Cairo had dramatic consequences for the evolution of Islāmic Egypt. Politically, the Fāṭimids went a step further than the Ṭūlūnids by setting up Egypt as an independent rival to the ʿAbbāsid caliphate. In fact, an avowed aim of the early Fāṭimid propagandists was to achieve world dominion, eradicating the ʿAbbāsid caliphate in the process. For a variety of reasons they achieved neither of these goals; nevertheless, at the height of Fāṭimid power in the mid-11th century, the Fāṭimid caliph could claim sovereignty over the whole North African coastal region, Sicily, the Hejaz and Yemen in Arabia, and southern Syria. Although actual politico-military control was never firm except in Egypt, allegiance paid to the Fāṭimids by their provinces was just as meaningful as that paid to the ʿAbbāsids and for a time was certainly more widespread. Even when the Fāṭimid state fell into decline in the 11th century and abandoned its imperial vision, Egypt continued to play an independent role in the Islāmic world under the leadership of Armenian generals who had gained control of the Fāṭimid armies.

Egypt set up as a rival to the ʿAbbāsid caliphate

Islāmization. It is difficult to estimate the religious change effected by the new dynasty except on the level of the governmental elite, which espoused the official doctrine of Ismāʿīlī Shīʿism—that the leader of the Islāmic community must be descended from ʿAlī, the fourth caliph. Because they believed that the Fāṭimid caliph was the only legitimate leader, the practice of Sunnī (orthodox) Islām was theoretically outlawed in Fāṭimid domains. But the practical difficulties which the Ismāʿīlī minority faced in imposing its will on the Sunnī majority meant that the Muslim population of Egypt remained predominantly Sunnī throughout the Fāṭimid period. Certainly there was no public outcry when Saladin, who founded the Ayyūbid dynasty, restored Egypt to Sunnī rule in 1171. Regarding non-Muslims, the Fāṭimids, with one notable exception, were known for their tolerance, and the Copts continued to serve in the bureaucracy. Several Copts held the highest administrative post—the vizierate—without changing their religion. Jews also figured prominently in the government; in fact, a Jewish convert to Islām, Ibn Killis, was the first Fāṭimid vizier and is credited with laying the foundations of the Fāṭimid administrative system. Christians and Jews even managed to survive the reign of the mad caliph al-Ḥākim (ruled 996–1021), who ordered the destruction of Christian churches in Fāṭimid territory, including the Church of the Holy Sepulchre in Jerusalem, and offered his non-Muslim subjects the choice of conversion to Islām or expulsion from Fāṭimid territory. This period of persecution undoubtedly accelerated the rate of conversion to Islām, if only on a temporary and superficial level.

In comparison with Iraq, Egypt contributed relatively little to Arabic literature and Islāmic learning during the early ʿAbbāsid period. But the Fāṭimids' intense interest in propagating Ismāʿīlism through a network of missionary propagandists made Egypt an important religious and intellectual centre. The foundation of the mosque–college of al-Azhar as well as of other academies drew Shīʿī scholars to Egypt from all over the Muslim world and stimulated the production of original contributions in literature, philosophy, and the Islāmic sciences.

Lack of literature and learning

Arabization. The Arabization of Egypt continued at a gradual pace. The early Fāṭimids' reliance on Berber troops was soon balanced by the importation of Turkish, Sudanese, and Arab contingents. The Fāṭimids are said to have used thousands of nomadic Arabs in the Egyptian cavalry and to have further stimulated Arabization by settling large numbers of Arabian tribesmen in Upper Egypt to deprive the Qarmaṭians—their Ismāʿīlī enemies in Iraq and Arabia—of Arab tribal support. On the other hand, the Fāṭimids reduced the Arab population of Egypt in the mid-11th century, when they incited the Banū Hilāl and the Banū Sulaym tribes to emigrate from Egypt into the neighbouring Berber kingdom of Ifrīqīyah.

Growth of trade. One of the most far-reaching changes in Fāṭimid times was the growth of Egyptian

Egyptian commerce

commerce, especially in Fusṭāṭ, which had become the port city for Cairo, the Fāṭimid capital. Theretofore, Iraq in the east and Tunisia in the west had been flourishing centres for trade conducted both within the Muslim world and between the Muslim and the Christian empires of the West. A number of factors contributed to alter this situation in favour of Egypt. As centralized power declined in Iraq, Mesopotamia, and Syria during the 9th and 10th centuries, traffic on the trade routes across these areas also declined. In Egypt, however, the establishment of a strong government, which soon controlled the Red Sea and maintained a strong navy in the eastern Mediterranean, offered an attractive alternative for the international transit trade between the Orient and Christendom. In addition to having the political stability essential for trade, the Fāṭimids encouraged commerce by their low tariff policy and their noninterference in the affairs of merchants who did business in Egypt. These factors, along with increased European mercantile activity in the Italian cities, helped restore Egypt as a great international entrepôt.

The end of the Fāṭimid dynasty. The Fāṭimid achievement in restoring to Egypt a measure of its ancient glory was remarkable but brief. Halfway through their history the political-religious authority of the Fāṭimid caliphs was vitiated by military uprisings that could be put down only by force. By 1163 the Fāṭimid caliph had been shunted aside in a power struggle betwen the vizier and the chamberlain, who were themselves so impotent that they had to seek help from the Sunnī and even from the crusader powers of Syria and Palestine. Thus began a series of invasions at the behest of Fāṭimid officials, which ended in 1169 with the occupation of Egypt by an army from Syria, one of whose commanders—Saladin—was appointed Fāṭimid vizier. Two years later Saladin restored Egypt to ʿAbbāsid allegiance, abolished the Fāṭimid caliphate, and, in effect, established the Ayyūbid dynasty.

The Ayyūbid dynasty (1171–1250). Under Saladin and his descendants, Egypt was reintegrated into the Sunnī world of the Eastern caliphate. Indeed, Egypt became champion of that world against the crusaders and, as such, chief target of the crusader armies. But this was a gradual process that required Saladin first to build an army strong enough to establish his power in Egypt and then to unite the factions of Syria and Mesopotamia under his leadership against the Franks. By so doing he reconstituted the Egyptian empire, which included, in addition to the areas just named, Yemen, the Hejaz, and, with the fall of Jerusalem (1187), a major part of the Holy Land.

Return to the Eastern caliphate under Saladin

The abolition of the Fāṭimid caliphate and the official reinstitution of Sunnī Islām seems to have caused little perturbation in Egypt except for an uprising by the Fāṭimid palace guard, quickly suppressed. This undoubtedly means that Ismāʿīlī Shīʿism was confined to Fāṭimid ruling circles.

Saladin's policies. Saladin's remission of all taxes not explicitly sanctioned by Islāmic law must have contributed to his own popularity as well as to the stability of his regime. To ensure the defense of his state against both internal and external enemies, he strengthened the fortifications of Cairo by building a citadel and extending the Fāṭimid city walls. Despite the major military and propagandistic efforts mounted against the crusaders, Saladin continued to treat the Christians of Egypt with tolerance; the Coptic Church thrived under the Ayyūbids, and Copts still served the government. Saladin even treated the Christians of Jerusalem with magnanimity after the conquest of that city.

Much to the consternation of the popes, trade between Egypt and the Italian city-states remained brisk, and the Egyptians were able to use raw materials provided by the Italian merchants to forge weapons against the crusaders. The administration of Egypt stayed in the hands of the vast bureaucracy; it was staffed mainly by scribes but was supervised by military officials.

Power struggles. The Ayyūbids introduced a significant change in the governance of their empire that was decisive for the history of their rule in Egypt. Though the Ayyūbids were themselves of Kurdish descent, Saladin followed the Turkish practice of assigning the provinces as fiefdoms to members of his family. In theory, such a measure would ensure the loyalty of the provinces to the central government of Egypt through the loyalty of Ayyūbid kinsmen to their family leader. In practice, however, the measure led to recurrent power struggles in which each governor used his province as a base from which to defy the supreme Ayyūbid power of Egypt. The sultans al-Malik al-ʿĀdil (died 1218) and al-Malik al-Kāmil (died 1238) eventually succeeded in reuniting Syria and Egypt under their leadership. Kāmil, especially, was able to exploit Frankish attacks—in the form of the Fifth Crusade, directed against Damietta—to rally family and provincial support for the defense of Egypt. Nevertheless, given the dissension within the Ayyūbid empire, it was clearly in the interest of the Egyptian sultan to reach a peaceful settlement with the crusaders; this was achieved in 1229 by a truce between Kāmil and the Holy Roman emperor Frederick II, by which Kāmil exchanged possession of Jerusalem and other territory in the Holy Land for Frederick's guarantee to support the sultan against aggression from any source.

Growth of Mamlūk armies. The only real security for Ayyūbid Egypt lay in its independent military strength. This explains why one of the last sultans, al-Malik aṣ-Ṣāliḥ Ayyūb (died 1249), resorted to increased purchase of Turkish slaves—Mamlūks—as a means of manning his armies. Although slave troops had formed an important part of Egyptian armies since the time of Aḥmad ibn Ṭūlūn, their strength had been checked by racial dissension among the various slave units and by the presence of nonslave elements. But after the death of aṣ-Ṣāliḥ Ayyūb in the course of the Sixth Crusade, which the Egyptians defeated thanks to the Mamlūk corps, the Mamlūks were able to exploit a palace feud and to elevate a member of their own ranks to the sultanate. Thus was established the Mamlūk sultanate, which lasted for two and a half centuries and brought Egypt to the peak of its evolution in the medieval period.

THE MAMLUK AND OTTOMAN PERIODS (1250–1800)

The Mamlūk dynasty (1250–1517). During the Mamlūk period Egypt became the unrivalled political, economic, and cultural centre of the eastern Arabic-speaking zone of the Muslim world. Symbolic of this development was the re-establishment in 1261 under the Mamlūks of the ʿAbbāsid caliphate in Cairo (the Mongols had abolished the caliphate when they invaded Baghdad in 1258). Although the caliph enjoyed little authority and had no power, the mere fact that the Mamlūks chose to maintain the institution in Cairo is a measure of their determination to dominate Arabic Islām. It is curious that the Mamlūks—all of whom were of non-Arab, non-Muslim origin and some of whom knew little if any Arabic—established a regime that saved a substantial portion of Muslim territory from pagan domination and established Egypt's supremacy in Arabic culture.

Egypt as the centre of eastern Islām under the Mamlūks

Political life. The political history of the Mamlūk state is complex; during their 264-year reign, no fewer than 45 Mamlūks gained the sultanate, and once, in desperate circumstances, a caliph (in 1412) was briefly installed as sultan. At times individual Mamlūks succeeded in establishing dynasties, most notably Sultan Qalāʾūn (ruled 1279–90), whose progeny ruled Egypt, with two short interruptions, until 1382. Often the Mamlūks chose to allow a sultan's son to succeed his father only for as long as it took another Mamlūk to build up enough support to seize the throne for himself. In reality there was no principle of legitimacy other than force, for without sufficient military power a sultan could expect to be overthrown by a stronger Mamlūk.

Nevertheless, several sultans succeeded in harnessing the energies of the Mamlūk system to establish internal stability and to embark on foreign conquests. Sultan Baybars I (ruled 1260–77), the true founder of the Mamlūk empire, led the Mamlūks to victory against the Mongols at ʿAyn Jālūt in 1260 and captured most of the re-

maining crusader fortifications in Palestine and Syria. During the long reign of al-Malik an-Nāṣir (ruled 1293–1341), the Mamlūks concluded a truce with the Mongols (1323) after several major battles and, despite widespread famine, outbreaks of religious strife, and Bedouin uprisings, maintained economic prosperity in Egypt and peaceful relations with foreign powers, both Muslim and Christian.

Although the state began to decline politically and economically after the death of Nāṣir in 1341, Egypt continued to dominate Eastern Arabdom. But the cumulative effect of the plague, which swept Egypt in 1348 and 1481, Timur's victory in Syria in 1400, and Egypt's loss to the Portuguese of control over the Indian trade, along with the sultans' inability to keep their refractory Mamlūk corps under control, gradually sapped the strength of the state. The best efforts of such a vigorous sultan as Qā'it Bāy (ruled 1468–96) failed to make Egypt strong enough to defend its Syrian empire against raids by the Turkoman states of Anatolia and Azerbaijan and campaigns of the Ottoman Turks.

Contributions to Arabic culture. By the time of the Mamlūks, the Arabization of Egypt must have been almost complete. Arabic had been the language of the bureaucracy since the early 8th century and the language of religion and culture even longer. Moreover, the prevalence of Arabic as a written and spoken language is attested by the discovery in the geniza (storeroom) of a Cairo synagogue of thousands of letters and documents —called the "Geniza Documents"—dating from the 11th through the 13th centuries. Though often written in Hebrew characters, the actual language of most of these documents is Arabic, which proves that Arabic was widely used even by non-Muslims. The main incentive for learning Arabic must have come from the desire of a subject population to learn the language of the ruling elite. The immigration of Arab tribesmen during the early centuries of the occupation, and their intermarriage with the indigenous inhabitants, must also have contributed to the gradual spread of Arabic in Egypt.

Prevalence of Arabic in Mamlūk Egypt

The specific Mamlūk contribution to Arabic culture, however, lay above all in the military achievement. By defeating the Mongols, the Mamlūks provided a haven in Syria and in Egypt for Muslims fleeing from Mongol devastation. The extent of this haven was narrowed by subsequent Mongol attacks against Syria, one of which led to a brief Mongol occupation of Damascus in 1294–95, so that Egypt received an influx of refugees from Syria itself as well as from areas farther east.

This accidental displacement of scholars and artisans into Egypt does not, however, wholly account for the efflorescence of certain types of cultural activity under the Mamlūks. In the same way as they supported the caliphate as a visible symbol of their legitimate claim to rule Islāmic territory, the Mamlūks cultivated and patronized religious leaders whose skills they needed in administering their empire and in directing the religious sentiments of the masses into safe (*i.e.*, nondisruptive) channels. Those divines who cooperated with the state were rewarded with government offices, in the case of the *'ulamā'* (religious scholars), and with endowed monasteries, in the case of the Ṣūfīs (mystics). On the other hand, those who dared criticize the prevailing social and moral order were thrown into prison (such was the fate of the famous legist, Ibn Taymīyah, who, having emigrated from Mesopotamia in order to escape the Mongols, was incarcerated in Cairo by the Mamlūks and their religious functionaries for spreading seditious doctrines).

Concrete evidence of the stimulus the Mamlūks gave to cultural life can be found chiefly in the fields of architecture and historiography. Dozens of public buildings erected under Mamlūk patronage are still standing in Cairo and include mosques, colleges, hospitals, monasteries, and caravansaries. Historical writing under the Mamlūks was equally monumental, in the form of immense chronicles, biographical dictionaries, and encyclopaedias.

Religious life. The Mamlūk period is also important in Egyptian religious history. With few and therefore notable exceptions, the Muslim rulers of Egypt had seldom interfered with the lives of their Christian and Jewish subjects so long as these groups paid the special taxes levied on them in exchange for state protection. Indeed, both Copts and Jews had always served in the Muslim bureaucracy, sometimes in the very highest administrative positions. Even the Crusades apparently failed to upset the delicate balance between Muslims and Christians. Trade with the Italian city-states had certainly continued, and there is no evidence that the local Christians were held accountable for the crusader invasions of Egypt. While it is true that Saladin dismissed all Copts from the bureaucracy and imposed sumptuary laws on them, this policy was abandoned by his successors in their desire to reach an accommodation with the crusaders.

With the establishment of the Mamlūk dynasty, however, it is generally agreed that the lot of the Christians, both in Egypt and in Syria, took a distinct turn for the worse. One indication of this change is the increased production of anti-Christian polemics written by Muslim theologians. A possible reason for the change may have been the association of Christians with the Mongol peril. Because the Mongols used Christian auxiliaries in their armies—Georgians and Armenians in particular—they often spared the Christian populations of towns they conquered, while slaughtering the Muslims. Also, the diplomatic efforts aimed at uniting the Mongols with Christian European powers in a joint crusade against the Muslims might have contributed to the Mamlūks' distrust of the Christians. But the dissatisfaction seems to have originated not so much with the Mamlūk rulers as with the masses, and it seems to have been directed not so much against Christians' sympathy for the Mongols as against their privileged position and role in the Mamlūk state.

Anti-Christian feelings

On several occasions popular resentment against the Copts' conspicuous wealth and their employment in the government was manifested in public demonstrations. Both Muslims and Christians resorted to arson, burning the others' sanctuaries, to express their hatred. Under such pressure, the Mamlūk government dismissed Christians from the bureaucracy on no fewer than nine occasions between 1279 and 1447, and in 1301 it ordered all the churches in Egypt closed. As a result of these intermittent persecutions and the destruction of churches, it is believed that the rate of conversion to Islām accelerated markedly in the Mamlūk period and that Coptic virtually disappeared except as a liturgical language. By the end of the Mamlūk dynasty, the Muslims may well have reached the same numerical superiority that they enjoy in modern times—a ratio of more than ten to one.

Economic life. In trade and commerce, the Mamlūk period marks the zenith of medieval Egyptian economic history. During the 13th and 14th centuries (as long, that is, as the sultanate was able to maintain order in Egypt), trade was heavy with Mediterranean and Black Sea ports and with India. The Oriental trade was controlled largely by a group of Muslim merchants known as the Kārimīs; the Mediterranean trade was left to European traders, whom the Mamlūks allowed certain privileges in Alexandria. By the 15th century, however, Egypt's commercial importance rapidly deteriorated as the result of population losses, increased government interference in commerce, Bedouin raiding, and Portuguese competition in the Indian trade.

The Ottomans (1517–1798). With the Ottomans' defeat of the Mamlūks in 1516–17, Egyptian medieval history had come full circle, as Egypt reverted to the status of a province governed from Istanbul. Again the country was exploited as a source of taxation for the benefit of an imperial government and as a base for foreign expansion. The economic decline that had begun under the late Mamlūks continued, and with it came a decline in Egyptian culture.

Economic decline under Mamlūks

Some historians attribute the lethargy of Ottoman Egypt solely to Ottoman domination. But although Ottoman policy was geared to imperial, not Egyptian needs, it was obviously to the rulers' benefit to provide a stable government that would maintain Egyptian agriculture at a high level of productivity and would promote the transit

trade. To a certain extent Ottoman actions served these purposes. The decisive factor that ultimately undermined Ottoman policies was the perpetuation of the former Mamlūk elite; though they collaborated with the Ottoman government, they often defied it and in the end they dominated it. By and large the history of Ottoman Egypt concerns the process by which the conquered Mamlūks reasserted their power within the Egyptian state.

The Ottoman conquest. From the conquest itself, the Ottoman presence in Egypt was entangled with Mamlūk factionalism. There is no doubt that the Ottomans invaded Syria in 1516 in order to break an incipient coalition against Ottoman expansion between the Ṣafavids of Persia and the Mamlūks of Egypt and Syria. The long-standing enmity between the Ottomans and the Mamlūks arose from their contest to control the Turkoman frontier states north of Syria. After the Ottomans strengthened their hold over eastern Anatolia in 1514, it was only natural that the Mamlūks should attempt to bolster their forces in northern Syria and exchange diplomatic missions with the Ṣafavids. The Ottoman Sultan Selim the Grim responded by attacking the reinforced Mamlūk army in Syria, probably as a preliminary step in a new campaign against the Ṣafavids. In 1516, after Selim had defeated the Mamlūks at Marj Dābiq (north of Aleppo), Ottoman goals had probably been met, especially since the Mamlūk sultan Qānṣūh al-Ghaurī died in the battle. But the Mamlūks rallied around a new sultan in Cairo, who refused to accept Selim's terms for a settlement. Spurred on by the Mamlūk traitor Khair Bey, Selim marched against Egypt in 1517, defeated the Mamlūks, and installed Khair Bey as Ottoman governor. Khair Bey died in 1552; thereafter, the Ottoman viceroy (called *vali*), with the title of pasha, was sent from Istanbul.

Mamlūk factionalism

Ottoman administration. In 1525 the Ottoman administration of Egypt was defined and codified by the Ottoman grand vizier, Ibrāhīm Pasha, who was dispatched to Egypt for this purpose by the Sultan Süleyman the Magnificent. According to the terms of Ibrāhīm Pasha's decree (*qanun-name*), Egypt was to be ruled by a viceroy aided by an advisory council (divan) and an army comprised of both Ottoman and local corps. The collection of taxes and the administration of the four provinces into which Egypt was divided were assigned to inspectors (*kashifs*). Although the Egyptian government was headed by bureaucratic officials sent from Istanbul, and supported by Ottoman troops, the Mamlūks were able to penetrate both the bureaucracy and the army. The *kashifs* were often drawn from Mamlūk ranks; three of the seven military corps formed by the Ottomans in the 16th century were recruited in Egypt, one of which—the Circassians—was composed of Circassian Mamlūks. Their service in the army enabled the Mamlūk amirs to secure high-ranking military posts that entitled them to serve on the divan itself.

The beys

By the 17th century a distinct elite bearing the title of bey had emerged, which consisted largely of Mamlūk amirs. These beys held no specific offices but were nevertheless paid a salary by the Ottoman government. The elite was perpetuated through the old Mamlūk system of purchasing slaves, giving them military training, then freeing them and attaching them to one of the great Mamlūk houses of Egypt. Thus, for all practical purposes, the Mamlūks maintained themselves as an elite throughout the Ottoman period. They were no longer the only politico-military elite, as they had been in the past, but they ultimately succeeded in re-establishing their dominance. Yet the chief obstacle to the growth of their power was not so much the Ottoman ruling hierarchy as it was their own factionalism. During the 17th and 18th centuries the Mamlūks were divided into two great rival houses—the Faqariyya and the Qasimiyya—whose mutual hostility often broke out into fighting and impaired the strength of Mamlūks as a bloc.

Mamlūk power under the Ottomans. In spite of internal dissension and the resistance of the non-Mamlūk hierarchy, the Mamlūks had emerged by the early 18th century as the supreme power in Egyptian politics. While the beys continued to acknowledge the authority of the Ottoman viceroy and to send tribute to Istanbul, the strongest single figure in Egypt was the bey who held the newly coined title of *shaykh al-balad* ("chief of the city"), which signified that he was recognized by the other beys as their chief. The Mamlūks' rise to power was climaxed by the careers of two amirs—ʿAlī Bey and Abū Dhahab—both of whom secured from the Sublime Porte (Ottoman government) de facto recognition of their autonomy in Egypt (1768–76) and even undertook military campaigns in Syria and the Hejaz. The Ottomans attempted to end the Mamlūk domination by sending an army to Egypt in 1786. Initially successful, this attempt failed and the troops were withdrawn a year later. A Mamlūk duumvirate was re-established, which lasted until Napoleon invaded Egypt in 1798.

Expansion. During the 16th century, when their regime in Egypt was strongest, the Ottomans used Egypt as a base for expansion to the south. Like the Mamlūk rulers before them, they attempted to control the southern approaches to Egypt by instituting their authority in Nubia; this they achieved by annexing Nubia as far south as the Third Cataract. Elsewhere, they undertook to reassert Egyptian command of the Red Sea, which the Portuguese had begun to contest during the early 16th century. Ottoman fleets and troops captured Yemen and Aden (1536–46) and thus dominated the lower Red Sea; in 1557 they strengthened this position by setting up a colony on the Abyssinian coast at Mitsiwa (Massawa). In the 17th century these outposts began to lose their importance as Ottoman and Portuguese power began to decline and the Dutch took over the spice trade.

Culture. Given the political instability and the economic decline that had prevailed in Egypt since late Mamlūk times, it is not surprising that the culture of Ottoman Egypt lacked vitality. Perhaps the most telling example of intellectual quiescence was the dramatic decline in the quantity of historical works produced in Egypt. As already noted, the Mamlūk period is renowned for the number and quality of its historians, partly because the amirs patronized court historians; by contrast, in almost three centuries of Ottoman rule Egypt produced only one historian worthy of note, al-Jabartī (died 1825) famous for his observations on the French occupation. The Ottomans also fell short of the Mamlūks' achievement in architecture; there is no lack of public buildings erected under Ottoman patronage, but even the best of these are imitations of the Byzantine basilica, which had been adopted as the model for mosques.

Impact of political instability

Religious affairs. Like all previous Muslim governments, the Ottomans continued to employ Copts in the financial offices of the bureaucracy. The Ottomans allowed the caliphate, so assiduously preserved in its nominal form by the Mamlūks, to lapse. At first the caliph was installed in Istanbul by Selim the Grim. Later the caliph—the last of the ʿAbbāsid line—returned to Egypt, where he died in the reign of Süleyman. The claim that the caliph had transferred his authority to the Ottoman sultan is an 18th-century invention. (D.P.L.)

V. Egypt since c. 1800

FROM THE FRENCH TO THE BRITISH OCCUPATION (1798–1882)

The French occupation and its consequences (1798–1805). Although several projects for a French occupation of Egypt had been advanced in the 17th and 18th centuries, the purpose of the expedition that sailed under Napoleon Bonaparte from Toulon in May 1798 was specifically connected with the war against Britain (see FRENCH REVOLUTIONARY AND NAPOLEONIC WARS). Bonaparte had discounted the feasibility of an invasion of England but hoped by occupying Egypt to damage British trade, to threaten India, and to obtain assets for bargaining in any future peace settlement. Meanwhile, as a colony under the benevolent and progressive administration of Revolutionary France, Egypt would be regenerated and regain its ancient prosperity. The military and naval forces were therefore accompanied by a commission of scholars and scientists to investigate and report upon the past and present condition of the country.

Eluding the British Mediterranean fleet under Lord Nelson, the French landed at Abū Qīr (Aboukir) Bay on July 1 and took Alexandria the next day. In an Arabic proclamation, Bonaparte assured the Egyptians that he came as a friend to Islām and the Ottoman sultan, to punish the usurping Mamlūks and to liberate the people. From Alexandria the French advanced on Cairo, defeating Murad Bey at Shubrākhīt (July 13), and again decisively at Imbābah, opposite Cairo in the so-called Battle of the Pyramids on July 21. Murad fled to Upper Egypt, while his colleague, Ibrāhīm Bey, together with the Ottoman viceroy, made his way to Syria.

Battle of the Pyramids

After entering Cairo (July 25), Bonaparte sought to conciliate the population, especially the religious leaders (*'ulamā'*), by demonstrating his sympathy with Islām and by establishing councils (divans) as a means of consulting Egyptian opinion. The destruction of the French fleet at Abū Qīr by Nelson in the so-called Battle of the Nile on August 1 virtually cut Bonaparte's communications and made it necessary for him to consolidate his rule and to make the expeditionary force as self-sufficient as possible. The savants, organized in the Institut d'Égypte, played their part in this. Meanwhile, Egyptian resentment at alien rule, administrative innovations, and the growing fiscal burden of military occupation, was exacerbated when the Ottoman sultan, Selim III (1789–1807), declared war on France, on September 11. An unforeseen revolt in Cairo, on October 21, was suppressed after an artillery bombardment that ended any hopes of cordial Franco-Egyptian coexistence.

Ottoman Syria, dominated by Ahmed Jezzar Paşa, the governor of Acre, was the base from which French-occupied Egypt might most easily be threatened, and Bonaparte resolved to deny it to his enemies. His invasion force crossed the frontier in February 1799 but failed to take Acre after a protracted siege (March 19–May 20), and Bonaparte evacuated Syrian territory. A seaborne Ottoman invading force landed at Abū Qīr in July but failed to maintain its bridgehead. At this point Bonaparte resolved to return to France and succeeded in slipping away on August 22, past the British fleet.

His successor as general in chief, Jean-Baptiste Kléber, viewed the situation of the expeditionary force with pessimism and, like many of the soldiers, wished to return to the theatre of war in Europe. He therefore entered into negotiations with the Ottomans and by the Convention of al-'Arīsh (January 24, 1800) agreed to evacuate Egypt. Sir Sydney Smith, the British naval commander in the eastern Mediterranean, sponsored the convention, but in this he had exceeded his powers and was instructed by his superior officer, Admiral Lord Keith, to require the French to surrender as prisoners of war. Although the Ottoman reoccupation was well under way, Kléber and the French determined on resistance and defeated the Turkish forces at the Battle of Heliopolis (March 20). A second revolt of Cairo, fomented by Ottoman fugitives, took about a month to suppress; but French authority had been restored when Kléber was assassinated by a Syrian Muslim, Sulaymān al-Ḥalabī, on June 14.

His successor, 'Abd Allāh Jacques Menou, a French officer (and former nobleman) who had turned Muslim, was determined to maintain the occupation and administered at first a tolerably settled country, although he lacked the prestige of his two predecessors. In 1801 a threefold invasion of Egypt began. British troops were landed at Abū Qīr in March, while the Ottomans advanced from Syria. Shortly afterward, British Indian forces were landed at Quṣayr on the Red Sea coast. The French garrison in Cairo capitulated in June, and Menou himself at Alexandria in September.

Surrender of the French

The brief episode of the French occupation was to be significant for Egypt in several ways. The arrival of a European army accompanied by scholars and scientists appropriately inaugurated the impact of the West, which was to be felt increasingly in the next 150 years. Egypt, protected for five centuries by the Mamlūk and Ottoman sultanates, was no longer immune from European attack: it had become an object of the contending policies of France and Britain, a part of the "Eastern Question." Bonaparte's savants had little success in interpreting Western culture to the traditionalist *'ulamā'* of Cairo; their achievement was rather to unveil Egypt to Europe. They uncovered the celebrated Rosetta Stone, which held a trilingual inscription making it possible to decipher hieroglyphics and which thus laid the foundation of modern Egyptology. Their reports and monographs were collected in the monumental *Description de l'Égypte* (1809–13, Paris).

Of more immediate consequence for Egypt was the effect of the French occupation upon internal politics. The Mamlūk ascendancy was fatally weakened. Murad Bey, who had made his peace with the French, died shortly before their capitulation in 1801; and Ibrāhīm Bey, who returned to Egypt with the Ottomans, had henceforward little power. The new Mamlūk leaders, 'Uthmān Bey al-Bardīsī and Muḥammad Bey al-Alfī, former retainers of Murad, headed rival factions and had in any case to reckon with the British and Ottoman occupation forces. In March 1803 the British were evacuated in accordance with the Peace of Amiens. But the Ottomans, determined to reassert their control over Egypt, remained, establishing their power through a viceroy and an occupying army, in which the most effective fighting force was an Albanian contingent. The Albanians, however, acted as an independent party and in May 1803 mutinied and installed their own leader as acting viceroy. When he was assassinated shortly afterward, the command of the Albanians passed to his lieutenant, Muḥammad 'Alī (born 1769), who, during the ensuing two years, cautiously strengthened his own position at the expense both of the Mamlūks and the Ottomans.

Ottomans remain in Egypt

Muḥammad 'Alī and his successors (1805–82). In May 1805 a revolt broke out in Cairo against the Ottoman viceroy, Khūrshīd Pasha. The *'ulamā'* invested Muḥammad 'Alī as viceroy. For some weeks there was street fighting and Khūrshīd was beseiged in the Citadel. In July Sultan Selim III confirmed Muḥammad 'Alī in office and the revolt ended.

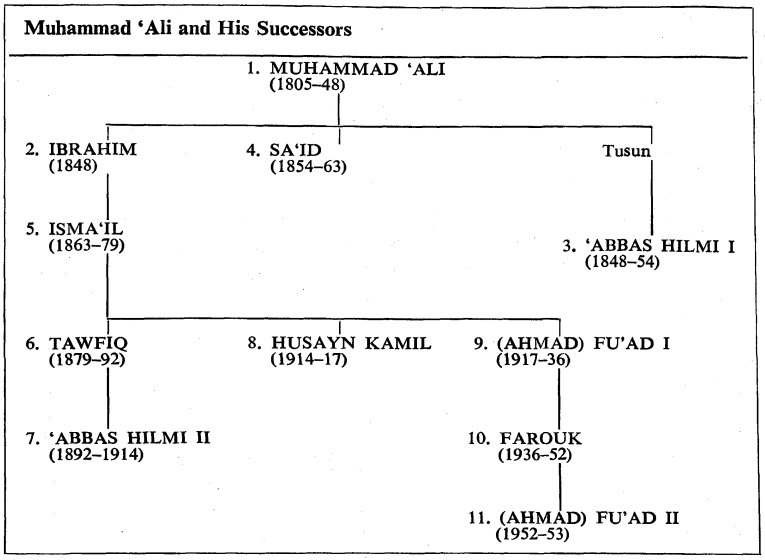
Muḥammad 'Alī's viceroyalty was marked by a series of military successes, some of which were attended by political failures that frustrated his wider aims. After the renewal of war between Britain and Napoleonic France in 1803, Egypt again became an area of strategic significance. A British expedition occupied Alexandria in 1807 but failed to capture Rosetta and, after a defeat at the hands of Muḥammad 'Alī's forces, was allowed to withdraw.

Military expansion. In Arabia, the domination of Mecca and Medina by fanatical Wahhābī Muslims was a serious embarrassment to the Ottoman sultan, who was the titular overlord of the Arabian territory of the Hejaz and the leading Muslim sovereign. At the invitation of Sultan Mahmud II (1808–39), Muḥammad 'Alī sent an expedition to Arabia that between 1811 and 1813 expelled the Wahhābīs from the Hejaz. In a further campaign (1816–18), Ibrāhīm Pasha, the viceroy's eldest son, defeated the Wahhābīs in their homeland of Najd, and brought central Arabia within Egyptian control. In 1820–21 Muḥammad 'Alī sent an expedition up the Nile and conquered much of what is now the northern Sudan. By so doing, he made himself master of one of the principal channels of the slave trade and began an African empire that was to be expanded under his successors.

After the outbreak of the Greek insurrection against Ottoman rule, Muḥammad 'Alī, at Sultan Mahmud II's request, suppressed the Cretan revolt in 1822. In 1825 Ibrāhīm began a victorious campaign in the Morea in southern Greece, where his military success provoked intervention by the European powers and brought on the destruction of the Ottoman and Egyptian fleets at the Battle of Navarino in October 1827. The Morea was evacuated the following year.

In 1831 Muḥammad 'Alī embarked upon the invasion of Syria. His pretext was a quarrel with the governor of Acre, but deeper considerations were involved, particularly the growing strength of the Sultan, which might threaten his own autonomy. Syria, moreover, was strategically important; and its products, especially timber,

Muḥammad 'Alī's invasion of Syria



usefully complemented the Egyptian economy. The Ottoman army was defeated near Konya in Anatolia (December 1832), and in 1833 the Sultan ceded the Syrian provinces to Muḥammad ʿAlī.

In 1839 Ottoman forces re-entered Syria but were defeated by Ibrāhīm at the Battle of Nizip (Nezib). A fortnight later Mahmud II died, and the Ottoman Empire seemed on the verge of dissolution; it was saved only by European intervention. In 1840 Ibrāhīm was compelled to evacuate Syria. Muḥammad ʿAlī's Arabian empire (which since 1833 had extended into the Yemen) crumbled at the same time. Although in June 1841 the new sultan, Abdülmecid I (1839–61), conferred on the family of Muḥammad ʿAlī the hereditary rule of Egypt, the viceroy's powers were declining. Because of his growing senility, Ibrāhīm succeeded him (July 1848) but died in the following November. The next viceroy was ʿAbbās I, the eldest grandson of Muḥammad ʿAlī. The old viceroy himself died in 1849.

Administrative changes. Muḥammad ʿAlī's military exploits would not have been possible but for radical changes in the administration of Egypt itself. Muḥammad ʿAlī was a pragmatic statesman whose principal object was to secure himself and his family in the unchallenged possession of Egypt. His immediate problem on his accession was to deal with the Mamlūks, who still dominated much of the country, and the *ʿulamāʾ*, who had helped him to power. The strength of these two groups rested largely on their control of the agricultural land of Egypt and the revenues arising therefrom. Gradually, between 1805 and 1815, Muḥammad ʿAlī eroded the system of tax farming that had diverted most of the revenues to the Mamlūks and other notables, imposed the direct levy of taxes, expropriated the landholders, and carried out a new tax survey. In 1809 he worsted the *ʿulamāʾ*, and in 1811 he massacred many of the Mamlūk leaders in Cairo, while Ibrāhīm expelled their survivors from Upper Egypt.

Restructuring of taxation

Muḥammad ʿAlī thus became effectively the sole landholder, with a monopoly over trade in crops, in Egypt, although later in his reign he made considerable grants of land to his family and dependents. The monopoly system was extended in due course from primary materials to manufactures, with the establishment of state control over the textile industry. Muḥammad ʿAlī's ambitious hopes of promoting an industrial revolution in Egypt were not realized, fundamentally because of the lack of available sources of power. The monopolies were resented by European merchants in Egypt and clashed with the economic doctrine of free trade upheld by the British government. Although a free-trade convention between Britain and the Ottoman Empire in 1838 (the Convention of Balta Liman) was technically binding on Egypt, Muḥammad ʿAlī succeeded in evading its application up to and even after the reversal of his fortunes in 1840–41.

Reform of the military

The old-style military forces (including the Albanians) on whom Muḥammad ʿAlī relied against his internal opponents and who conquered the Hejaz, Najd, and the northern Sudan, were heterogeneous and unruly. An attempt to introduce Western methods of training in 1815 provoked a mutiny. Muḥammad ʿAlī then decided to form an army of slave-troops dependent wholly upon himself and trained by European instructors. The conquest of the Sudan was intended to provide the recruits. But the slaves, encamped at Aswān, died wholesale, and, for the mass of his troops Muḥammad ʿAlī had to look elsewhere. In 1823 he took the momentous step of conscripting Egyptian peasants for the rank and file of his "new model army." On the other hand, the officers were mostly Turkish-speaking Ottomans, while the director of the whole enterprise, Sulaymān Pasha (Colonel Sève), was a former French officer. The conscription was brutally administered and military life harsh. There were several ineffective peasant revolts, while flight to the towns and (before 1831) to Syria produced rural depopulation and a decline in cultivation.

As reorganization proceeded, the viceroy gradually built a new administrative structure. While institutions were created and discarded according to his changing needs, Muḥammad ʿAlī depended essentially upon the members of his own family, particularly Ibrāhīm, and loyal servants, such as his Armenian confidant Boghos Bey. Characteristic of his governmental system were councils of officials, convened to deliberate on public business, and administrative departments (divans) that bore some resemblance to the ministries of European governments. In local administration, Muḥammad ʿAlī established a highly centralized system with a clear chain of command from Cairo through the provincial governors, down to the village headmen. Initiative was not encouraged, but firm control had taken the place of anarchy.

These changes necessitated the training of officers and officials in the new Europeanized ways of working; and this in turn resulted in the creation of a range of educational institutions alongside the traditional Muslim schools that prepared the *ʿulamāʾ*. Much of the foundation work was done by expatriates, while missions of Egyptian students were sent to Europe, especially to Paris. One of these, Rifāʿa Rāfiʿ aṭ-Ṭahṭāwī (1801–73), played the leading part in inaugurating the translation of European works into Arabic and so was a pioneer both in the interpretation of European culture to Egypt and in the renaissance of literary Arabic. The establishment of a government printing press (*c.* 1822) made possible the wide dissemination of the new books.

ʿAbbās I and Saʿīd, 1848–63. The reign of ʿAbbās I (1848–54) indicates how precarious was the advance of westernization in Egypt. The effort had already been relaxed in the last decade of Muḥammad ʿAlī's rule, and ʿAbbās showed himself to be a traditionalist. It was typ-

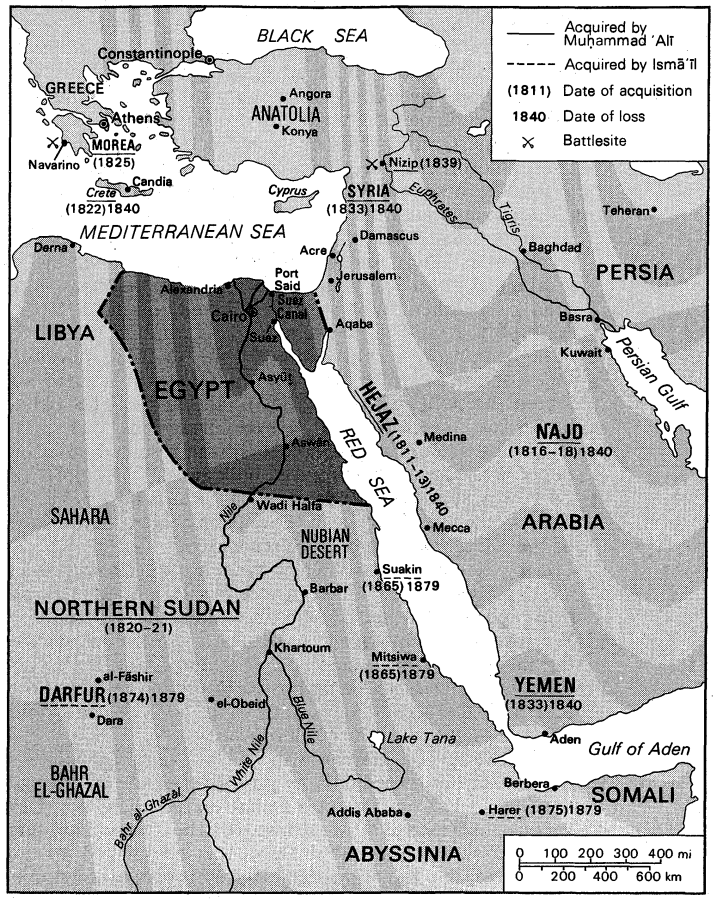
ical of his policy that he closed the school of languages and the translation bureau and sent their director, aṭ-Ṭahṭāwī, to virtual exile in the Sudan. The French, who had played so large a part in Muḥammad ʿAlī's reforms, fell into disfavour, and for diplomatic support ʿAbbās turned to their British rivals, whose support was needed against the Ottomans. Although initially ʿAbbās was ostentatiously loyal to the Sultan, he resented an attempt made at this time to curtail his autonomy. The British, for their part, had their communications with India facilitated by the grant of a concession to build a railway from Alexandria to Cairo; the line was completed between 1851 and 1856 and was extended to Suez two years later. Saʿīd (1854–63), who succeeded on ʿAbbās' mysterious and violent death, inaugurated another reversal of policy. While he lacked Muḥammad ʿAlī's energy and ability, he was not unsympathetic to the westernizers. To his French friend Ferdinand de Lesseps he granted in 1854 a concession for the cutting of a canal across the isthmus of Suez. This embroiled him both with the Sultan, whose prerogative had been encroached upon, and the British, whose overland railway route was threatened by the project; a deadlock lasted throughout his reign.

The concession to Ferdinand de Lesseps

Ismāʿīl, 1863–79. Ismāʿīl, the son of Ibrāhīm Pasha, who succeeded on the death of Saʿīd, displayed some of his grandfather's dynamic energy and enthusiasm for modernization. He lacked caution, however, and his reign ended in catastrophe. From his predecessors he inherited a precarious economy and a burden of debt. The American Civil War (1861–65) produced a boom in Egyptian long-staple cotton. This had been introduced and developed in Muḥammad ʿAlī's time, but its production had languished until the interruption of supplies of American cotton caused a fourfold increase in price during the war years. When peace returned, prices collapsed with disastrous consequences for the Egyptian economy. In the management of his finances, Ismāʿīl was both extravagant and unwise and laid himself open to unscrupulous exploitation. Ismāʿīl was committed to the Suez Canal project, but he modified the grant in two important respects: by withdrawing the cession of a strip of land from the Nile to the Suez isthmus, along which a fresh water canal was to be constructed; and by refusing to provide unlimited peasant labour for the project. The matter was submitted to arbitration; an indemnity of over £3,000,000 was imposed on Ismāʿīl, who also agreed to pay for a large block of shares put by de Lesseps to Saʿīd's account. French pressure on the Sultan succeeded at last in overcoming resistance to the canal project at Istanbul, and a firman (decree from the sultan) authorizing its construction was granted in March 1866. Work had in fact already been going on for seven years, and in November 1869 the Suez Canal was opened to shipping by the empress Eugénie, the wife of Napoleon III of France. The incident symbolized the political and cultural orientation of Egypt in the middle decades of the 19th century.

Ismāʿīl, in other ways, presented himself as the ruler of a new and important state. Although his relations with his suzerain, Sultan Abdülaziz (1861–76), were normally friendly, he was no less anxious than his predecessors to secure the autonomy of his dynasty. In 1866 he obtained a firman establishing the succession by primogeniture in his own line—abandoning the contemporary Ottoman rule of succession by the eldest male. A year later a firman conferred the special title of khedive, which had in fact been used unofficially since Muḥammad ʿAlī's time and which distinguished the viceroy of Egypt from other Ottoman governors. A period of strained relations developed between the Khedive and the Sultan arising from Ismāʿīl's implied pretensions to sovereignty at the time of the opening of the Suez Canal, but the two were later reconciled; a firman reconfirmed the Khedive's privileges in 1873. These concessions by the Sultan, however, cost Ismāʿīl heavy expenditure and an increase in the annual Egyptian tribute and formed another factor in the growth of Ismāʿīl's indebtedness.

Ismāʿīl had inherited an African empire in the northern



Expansion of Egypt under Muhammad ʿAlī and Ismāʿīl.

Sudan. Since the middle of the century, in consequence of the abolition of the monopolies, merchants had penetrated south and southwest, up the White Nile and the Baḥr al-Ghazāl, in search of ivory. An ancillary slave trade had developed that was repugnant to the European conscience. Humanitarian and expansionist motives thus coincided to persuade Ismāʿīl to extend Egyptian rule into these remoter regions. He made considerable use of expatriates, notably the Englishmen Sir Samuel Baker and Charles George ("Chinese") Gordon, who extended the Khedive's nominal authority to the African Great Lakes. Another series of events led to the conquest in 1874 of the sultanate of Darfur in the west. The Khedive also wished to make Egypt the dominant power in the Red Sea region. The Sultan granted him the old Ottoman ports of Suakin and Mitsiwa in 1865. Egyptian control was established on the Somali coast, and in 1875 Harar was captured. Attempts to invade Abyssinia in 1875 and 1876 were, however, unsuccessful and marked the limits of Ismāʿīl's imperial expansion.

Penetration southward

Like other parts of the Ottoman Empire, Egypt was bound by the Capitulations—a system of privileges derived from ancient treaties with former sultans. Under the Capitulations, European and American residents in Egypt were exempt from local taxation and were subject only to their own consular courts. By patient negotiations over several years, Nūbār Pasa, Ismāʿīl's Armenian minister, succeeded in establishing the Mixed Courts in 1875. These had jurisdiction in cases involving Egyptians and foreigners, or foreigners of different nationalities, and had both foreign and Egyptian judges, who administered codes based on French law.

By this time the social consequences of the agrarian and political changes inaugurated by Muḥammad ʿAlī were clearly appearing. The Khedive and his family were the principal landholders in Egypt, possessing extensive personal estates quite apart from the state lands. Around the khedivial family was a parvenu aristocracy that held the principal civil and military offices. Many of its mem-

bers were also great landowners; most of them were Turkish or Circassian by origin. Although the condition of the peasantry had been adversely affected by military conscription, by corvées for public works (including large-scale demands for labour on the railways and the Suez Canal), and by ill-considered economic and industrial experiments, the rights of cultivators on their land gradually increased. The richer peasants, from whom the village headmen were recruited, in particular increased in importance. When Ismāʿīl set up a consultative council (November 1866), the Assembly of Delegates, chosen by indirect election, the great majority of its members were village headmen. While Ismāʿīl did not intend that the Assembly should limit his power, its establishment and composition were indications of the political development of the Egyptians in 60 years. Conscription had affected the political significance of the army. The ascendancy of the entrenched Turco-Circassians was challenged by native Egyptian officers, who resented the privileged position of their foreign colleagues. The defeat of the Circassian commander in chief, Rātib Pasha, by the Abyssinians in 1876 was a blow from which the prestige of the old officer group never recovered.

Creation of the Assembly of Delegates

In the Assembly and the army, and among the westernized intelligentsia, politically conscious individuals and groups began to emerge who drew their ideas from both Western and Islāmic sources. Their organization was for the most part small scale and ephemeral, and their outlook was subversive, being hostile to the autocracy of the Khedive, the ascendancy of the Turco-Circassians, and the pervasive power of the Europeans.

Political tension increased in the last years of Ismāʿīl's reign. Various expedients to postpone bankruptcy (*e.g.*, the sale in 1875 of his Suez Canal shares to Britain) had failed, and in 1876 the Caisse de la Dette Publique (Commission of the Public Debt) was established for the service of the Egyptian debt. Its members were nominated by France, Britain, Austria, and Italy. In the same year, Egyptian revenue and expenditure were placed under the supervision of a British and a French controller (the Dual Control). After an international enquiry in 1878, Ismāʿīl accepted the principle of ministerial responsibility for government and authorized the formation of an international ministry under Nūbār. Ismāʿīl, however, was not prepared to yield his autocracy tamely. In 1879 he profited from an army demonstration against the European ministers to dismiss Nūbār, and he worked in alliance with the Assembly of Delegates to destroy international control over Egypt. By this time, however, his standing outside Egypt had been lost; and in June 1879, Sultan Abdülhamid II (1876–1908), at the instigation of France and Britain, deposed him in favour of his son, Tawfīq.

Renewed European intervention, 1879–82. European domination was immediately reasserted. The Dual Control was revived, the British controller being Evelyn Baring, later Lord Cromer. By the Law of Liquidation (July 1880), the annual revenues were divided into two approximately equal portions, one of which was assigned to the Caisse de la Dette. The Assembly of Delegates was dissolved. The forces of resistance that Ismāʿīl had stimulated were not, however, allayed by these means. There had already come into existence a nationalist group within the Assembly, prominent among whom was Sharīf (Chérif) Pasha, prime minister from April to August 1879. In the army, a group of Egyptian officers, whose leader was ʿUrābī (Arabi) Pasha, was disaffected from the Khedive and resentful of European control of Egypt. By 1881 these two groups had allied to form the National Party, al-Ḥizb al-Waṭanī.

Open tension appeared with a petition drawn up in January 1881 by ʿUrābī and two of his colleagues against the war minister, Rifqī Pasha, a Circassian. They were arrested and court-martialled but released by mutineers. Tawfīq capitulated, dismissed Rifqī, and appointed Bārūdī Pasha, one of ʿUrābī's friends, as war minister. But the ʿUrābists still felt themselves endangered; a military demonstration in Cairo in September 1881 compelled Tawfīq to appoint a new ministry under Sharīf and to convoke the Assembly. But the alliance between the military group and Sharīf was uneasy.

Meanwhile, the European powers were becoming alarmed. An Anglo-French note sent in January 1882 with the intention of strengthening the Khedive against his opponents had the contrary effect. The Assembly of Delegates swung toward the ʿUrābists. Sharīf resigned and Bārūdī became prime minister with ʿUrābī as war minister. Rioting ensued on June 11 after British and French naval forces had been sent to Alexandria. From this point Britain took the initiative. The French refused participation in a bombardment of Alexandria (July 11), while an international conference held at Istanbul was boycotted by the Turks and produced no solution of the problem. The British government finally resolved on intervention and sent an expeditionary force to the Suez Canal. The ʿUrābists were rapidly defeated at Tall al-Kabīr (September 13, 1882), and Cairo was occupied the next day.

European intervention

THE PERIOD OF BRITISH DOMINATION (1882–1952)

The British occupation and the Protectorate (1882–1922). The British occupation marked the culmination of developments that had been at work since 1798: the de facto separation of Egypt from the Ottoman Empire, the attempt of European powers to influence or control the country, and the rivalry of France and Britain for ascendancy in the country. Through the last minute withdrawal of the French, the British had secured the sole domination of Egypt. W.E. Gladstone's Liberal government was, however, reluctant to prolong the occupation or to establish formal political control, which it feared would antagonize both the Sultan and the other European powers; but the British were unwilling to evacuate Egypt without securing their strategic interests, and this never seemed possible without maintaining a military presence there.

An incident at the outset of the occupation was significant of future tensions. On British insistence, the Khedive's government was obliged to place ʿUrābī and his associates on public trial and to commute the resulting death sentences to exile. Tawfīq's prestige, slight enough at his accession, and diminished in the three years before the occupation, was still further undermined by this intervention of the British government. Meanwhile, Lord Dufferin, the British ambassador in Istanbul, visited Egypt and prepared a report on measures to be taken for the reconstruction of the administrative system. The projects of reform that he envisaged would necessitate an indefinite continuation of the occupation. The implications of this for British policy were slowly and reluctantly accepted by the ministry in London, under pressure from its representative in Cairo, the British agent and consul general, Sir Evelyn Baring, who in 1891 became Lord Cromer.

Public trial of ʿUrābī

Two principal problems confronted the occupying power: first, the acquisition of some degree of international recognition for its special but ambiguous position in Egypt; secondly, a definition of its relationship to the khedivial government, which formed the official administration of the country. The main European opponents of recognition of the British position were the French, who resented the abolition of the Dual Control (December 1882). The Caisse de la Dette remained in existence and until 1904 the British had to tread warily in order to circumvent French opposition in this institution. In the early years of the occupation, when Egyptian finances were in disarray, French hostility was a serious problem, but from 1889 onward there was a budget surplus and consequently greater freedom of action for the Egyptian government. A moderate degree of international agreement over Egypt was attained by the Convention of London (1885), which secured an international loan for the Egyptian government and added two further members (nominated by Germany and Russia) to the Caisse de la Dette. In 1888 the Convention of Constantinople (Istanbul) provided that the Suez Canal should always be open in war and peace alike. This was, however, a statement of principle rather than fact; without British cooperation it remained a dead letter.

Cromer's personal rule

In matters concerning the international status of Egypt, the decisions were taken in London, but where the internal administration of the country was concerned, Cromer's opinions were usually conclusive. Although throughout the occupation the facade of khedivial government was retained, British advisers attached to the various ministries were more influential than their ministers, while Cromer himself steadily increased his control over the whole administrative machine. Tawfīq himself gave little trouble: he owed his throne to the British and was unpopular with the Egyptians. His prime ministers were more tenacious of their rights. Sharīf Pasha, prime minister at the beginning of the occupation, opposed the proposal, made by the British government through Baring, that the Egyptians should abandon the Sudan, where the Mahdī was conducting a successful revolutionary war (see NILOTIC SUDAN, HISTORY OF THE). In January 1884 Baring, who had obtained from London a declaration that British "advice" was mandatory on Egyptian ministers and governors, obtained Sharīf's resignation. He was succeeded by Nubar Pasha, who resented the power of the British advisers, especially in the Ministry of the Interior. In 1888 Nubar challenged Baring and was dismissed. Riyāḍ (Riaz) Pasha, another elder statesman, who succeeded as prime minister, was a man of determined character—he had resigned from Sharīf's ministry in 1882 over ʿUrābī's trial, and he was as unwilling as Nubar to allow further British invasion of the internal administration. A clash developed over the Ministry of Justice, and Riyāḍ resigned in 1891. Thereafter, until November 1908, with a break in 1893–95, the prime minister was Muṣṭafā Fahmī Pasha, who showed himself an obedient instrument of Cromer.

ʿAbbās Ḥilmī II, 1892–1914. The death of Tawfīq and the accession of his 17-year-old son, ʿAbbās Ḥilmī II, in 1892 marked the beginning of a new phase of opposition to the occupation. The new khedive was not content to accept Cromer's tutelage, while the British agent resented the attempts of one so much his junior to play a serious role in Egyptian politics. ʿAbbās dismissed Muṣṭafā Fahmī in January 1893 and tried to appoint his own nominee as prime minister. Cromer, backed by the British government, frustrated his endeavours, and Fahmī returned to office in November 1895. ʿAbbās provoked another crisis in January 1894 by public criticism of the British officers who had reorganized the Egyptian army after the occupation, and especially H.H. Kitchener, the sirdar (commander in chief). Once again Cromer intervened, and ʿAbbās was compelled to make humiliating amends.

Other considerations apart, the behaviour of ʿAbbās in the early years of his reign indicated the emergence to political consciousness and activity of a new generation who had only been children when the occupation began.

Muṣṭafā Kāmil

One of ʿAbbās' contemporaries was the lawyer and journalist Muṣṭafā Kāmil (1874–1908), who had studied in France and then had entered a circle of anglophobe writers and politicians. On returning to Egypt in 1894 he had reached an understanding with the Khedive on the basis of their common detestation of the British occupation. By his speeches and writings (in 1900 he founded his own newspaper, *al-Liwāʾ*) he endeavoured to create a spirit of Egyptian patriotism that would appeal to all classes and communities and rally the nation around the Khedive. A boost was given to nationalism by the campaigns for the reconquest of the Sudan (1896–98) and still more by the Condominium Agreement of 1899, which nominally gave Egypt and Britain joint responsibility for the administration of the reconquered territory but in effect made the Sudan a British possession. This was humiliating for the Egyptians, who had ruled there in the past and had taken a major part in its reconquest.

A final episode in the reconquest of the Sudan, the confrontation of British and French at Fashoda on the White Nile in 1898, was followed by the reconciliation of the two powers in the Entente Cordiale (1904), which in effect gave Britain a free hand in Egypt. This was a blow to the hopes of Muṣṭafā Kāmil and to his alliance with the Khedive, who showed himself more willing to cooperate with Cromer. Muṣṭafā Kāmil now turned to Sultan Abdülhamid. When a dispute (the Tābah Incident, 1906) arose between the Ottomans and the occupying power over the Sinai Peninsula, Muṣṭafā Kāmil sought to rally Egyptian nationalist opinion in favour of the Sultan. In this he was not wholly successful since moderates and Christians were disturbed by the pan-Islāmic tone of his propaganda.

British domination in Egypt and Cromer's personal ascendancy never seemed more secure than in the period following the Entente Cordiale. But the "veiled protectorate" had hidden weaknesses. Cromer was both out of touch and out of sympathy with the new generation of Egyptians. The occupation had become to all intents and purposes permanent, and the consequent growth of the British official establishment created frustration among educated Egyptians, whose resentment was increasingly articulate. The British, however, saw themselves as the benefactors of the Egyptian peasantry, whom they had delivered from the corvée and the lash.

The Dinshawāy Incident

The Dinshawāy Incident showed them in another light. In June 1906 a fracas between villagers at Dinshawāy and a party of British officers out pigeon shooting resulted in the death of a British officer. The special tribunal set up to try the matter imposed exemplary and brutal sentences on the villagers. In the bitter aftermath of this affair, Cromer retired in May 1907.

The breakdown of the understanding between the Khedive and Muṣṭafā Kāmil, and the latter's handling of the Tābah Incident, stimulated the formation of new political groupings. Some moderate nationalists founded the newspaper *al-Jarīdah* in March 1907 and organized themselves as the Umma (Nation) Party in September. In the following month, Muṣṭafā Kāmil produced a program for his own followers, who revived the old name of the Nationalist Party. The Khedive, for his part, found a new ally in Shaykh ʿAlī Yūsuf, who since 1889 had been publishing a newspaper called *al-Muʾayyad* and now organized the so-called Constitutional Reform Party. These "parties" were all outgrowths of political journalism, were centred around single individuals or small groups, and were lacking in both a coherent ideology and a mass appeal. Their existence was precarious: the Nationalist Party never recovered from Muṣṭafā Kāmil's death in February 1908, while the Constitutional Reform Party expired with its founder in 1913.

Sir Eldon Gorst, who succeeded Cromer, had served in Egypt from 1886 to 1904 and brought a fresh mind to bear on the problems of the occupation. He obtained an understanding with the Khedive and endeavoured to diminish the growing power and numbers of the British establishment. At the same time he tried to give more effective authority to Egyptian political institutions. Muṣṭafā Fahmī's long premiership ended and he was followed by a Copt, Buṭrus Ghālī Pasha. When Gorst died prematurely in July 1911, he had attained only limited success. Many British officials resented his policies, which at the same time failed to conciliate the nationalists. A project for the extension of the Suez Canal Company's 99-year concession by 40 years was thrown out by the General Assembly (a quasi-parliamentary body, set up in 1883), while Buṭrus Ghālī, who had advocated it, was assassinated a few days later by a Muslim extremist. The appointment of Lord Kitchener to succeed Gorst portended the end of conciliation of the Khedive. But Kitchener, although autocratic, was not wholly illiberal; his attempts to limit the power and influence of ʿAbbās Ḥilmī served the interests of the nationalists. The Organic Law of 1913 created a new and more powerful Legislative Assembly that served as a training ground for the nationalist leaders of the postwar period. At the same time, the peasants were helped by the improvement of agriculture and by legal protection of their holdings from seizure for debt.

World War I and independence. In November 1914 Britain declared war on the Ottoman Empire and in December proclaimed a protectorate over Egypt, deposed ʿAbbās, and appointed his uncle, Ḥusayn Kāmil, with the title of sultan. Kitchener was succeeded by Sir Henry

MacMahon, and he by Sir Reginald Wingate, both with the title of high commissioner. Although Egypt was not required to provide troops, the people, and particularly the peasantry, suffered from the effects of war. The declaration of martial law and the suspension of the Legislative Assembly curbed the activities of middle class nationalists. In October 1917 the amenable Ḥusayn Kāmil died, and was succeeded as sultan by his ambitious brother, Aḥmad Fu'ād.

Attempt of Zaghlūl to secure Egypt's independence

On November 13, 1918, two days after the Armistice, Wingate was visited by three Egyptian politicians headed by Sa'd Zaghlūl Pasha. Zaghlūl had held ministerial office, had sat in the Legislative Assembly before the war, and was coming into prominence as a nationalist spokesman. He demanded autonomy for Egypt and announced his intention of leading a delegation (Arabic *wafd*) to state his case in England. The British government's refusal to accept a delegation, followed by the arrest of Zaghlūl, produced a widespread revolt in Egypt; and Lord Allenby, the victor over the Turks in Palestine, was sent out as special high commissioner. Although regarded in Britain as a strong man who would save the situation, he insisted on concessions to the nationalists in the hopes of reaching a settlement. Zaghlūl was released, and the Wafd, now a country-wide organization, dominated Egyptian politics. The Milner Commission (1919–20), sent to report on the establishment of constitutional government under the protectorate, was boycotted, but Milner subsequently had private talks with Zaghlūl in London. Finally, hoping to outmanoeuvre Zaghlūl and to build up a group of pro-British politicians in Egypt, Allenby pressed his government to promise independence without previously securing British interests by a treaty. The declaration of independence (February 28, 1922) ended the protectorate but, pending negotiations, reserved four matters to the discretion of the British government: the security of imperial communications, defense, the protection of foreign interests and of minorities, and the Sudan. On March 15 the Sultan became King Fu'ād I of Egypt.

The Kingdom of Egypt (1922–52). The new kingdom was in form a constitutional monarchy. The constitution, based on that of Belgium and promulgated in April 1923, defined the King's executive powers and established a bicameral legislature. An electoral law provided for universal manhood suffrage and the indirect election of deputies to the lower house: the Senate was half elected and half appointed. But Egyptian constitutionalism was as illusory as Egyptian independence. A political struggle was continually waged among three opportunist contestants—the King, the Wafd, and the British.

The Wafd and the King. Fu'ād shared the autocratic tendencies of his predecessors. He was never popular and felt insecure, particularly in the early years of his rule, since his accession excluded claimants with greater legitimacy. Hence he was prepared to intrigue with the nationalists or with the British to secure his position and powers. The Wafd, with its mass following, elaborate organization, and (until his death in 1927) charismatic leader in Zaghlūl, was the only truly national party in Egypt. Ideologically, it stood for national independence against British domination and for constitutional government against royal autocracy. In practice—and increasingly as time went on—its leaders were prepared to make deals with the British or the King to obtain or retain power. Personal and political rivalries led to the formation of splinter parties, the first of which, the Liberal Constitutionalist Party, broke off as early as 1922. The primary aim of the British government, represented by its high commissioner (after 1936, its ambassador), was to secure imperial interests, especially the control of communications through the Suez Canal. The need for a treaty to safeguard these interests led Britain on more than one occasion to conciliate nationalist feeling by supporting the Wafd against the King.

The repetitive course of events from 1923 to the outbreak of the World War II reflects the interplay of these three forces. The first general election, in January 1924, gave the Wafd a majority, and Zaghlūl became prime minister. His few months of office were marked by unsuccessful treaty discussions with the British prime minister Ramsay MacDonald and by tension with the King. When in November 1924 Sir Lee Stack, the sirdar and governor-general of the Sudan, was assassinated in Cairo, Allenby immediately presented an ultimatum that, although later modified by the British government, caused Zaghlūl to resign. The general election of March 1925, in spite of manipulation by the new government, left the Wafd still the strongest party, but the Parliament no sooner met than it was dissolved. For over a year Egypt was governed by decree. The third general election, in May 1926, again gave the Wafd a majority. The British frowned on a return of Zaghlūl to the premiership, which went to the Liberal Constitutionalist, 'Adlī Yegen (Yakan), while Zaghlūl held the presidency of the Chamber of Deputies until his death in the following year. Once again tension developed between the parliament and the King, and in April 1927 'Adlī resigned, to be succeeded by another Liberal Constitutionalist, 'Abd al-Khāliq Tharwat (Sarwat) Pasha, who negotiated a draft treaty with Sir Austen Chamberlain, the British foreign secretary. The draft treaty, however, failed to win the approval of the Wafd. Tharwat resigned (March 1928), and Muṣṭafā an-Naḥḥās (Nahas) Pasha, Zaghlūl's successor, became prime minister. But he was unacceptable both to the British and the King, who dismissed him in June and dissolved the Parliament in July. In effect, the constitution was suspended, and Egypt was again governed by decree under a Liberal Constitutionalist premier, Muḥammad Maḥmūd Pasha.

An-Naḥḥās prime minister

Negotiations between Maḥmūd and Arthur Henderson, the British foreign secretary, resulted in agreement on draft treaty proposals in June 1929; but since Maḥmūd was unable to overcome Wafdist opposition, British influence was thrown behind a return to constitutional government, hoping that a freely elected Parliament would approve the proposals. In the fourth general election (December 1929), the Wafd won a majority, and an-Naḥḥās again became prime minister. Resumed negotiations with Henderson broke down over the problem of the Sudan, from which the Egyptians had been virtually excluded since 1924. An-Naḥḥās also clashed with the King, whose influence he sought to curtail. He resigned in June 1930, and Fu'ād appointed Ismā'īl Ṣidqī (Sidki) Pasha to the premiership. The constitution of 1923 was abrogated, and another promulgated by royal decree. This, with its accompanying electoral law, strengthened the King's power. By this and other measures, Ṣidqī sought to break the power of the Wafd, which boycotted the general election of June 1931. The strong government of Ṣidqī lasted until September 1933, when he was dismissed by the King. Thereafter, for over a year, palace-appointed governments ruled Egypt.

But Fu'ād, whose health was failing, could not hold out indefinitely against the internal pressure of the Wafd and the external pressure of Britain, which was becoming increasingly anxious for a treaty with Egypt as relations with Italy deteriorated over the Abyssinian question. In April 1935 the constitution of 1923 was restored, and a general election in May 1936 gave the Wafd a majority once more. Fu'ād had died in the previous month and was succeeded by his son Farouk (Fārūq), still a minor. An-Naḥḥās became prime minister for the third time. In the changed national and international situation, agreement was quickly reached with Britain, and a treaty of mutual defense and alliance was signed in August 1936. At the conference of Montreux, held in the following year, Egypt, with the backing of Britain, obtained the immediate abolition of the Capitulations and the extinction of the Mixed Courts after 12 years. In 1937 also, Egypt became a member of the League of Nations.

Treaty with Britain

An-Naḥḥās was now at the height of his power but was soon to be overthrown. In July 1937 King Farouk came of age and assumed his full royal powers. He was both popular and ambitious to rule, and tension rapidly developed between him and his prime minister. A split developed in the Wafd: Maḥmūd Fahmī an-Nuqrāshī (Nokrashy) Pasha and Aḥmad Māhir (Maher) Pasha were expelled and formed the Sa'dist Group. The Wafdist youth

movement, the Blueshirts, was opposed by the Greenshirts of Young Egypt, a fascist organization. In December 1937 Farouk dismissed an-Naḥḥās. In the ensuing general election, the Wafd won only 12 seats.

World War II and its aftermath. Although at the outbreak of World War II, in September 1939, Egypt provided facilities for the British war effort, few Egyptians were enthusiastic supporters of Britain, and many expected its defeat. In 1940 the British brought pressure on the King to dismiss his prime minister, ʿAlī Māhir, and to appoint a more cooperative government. When, early in 1942, German forces prepared to invade Egypt, British intervention compelled King Farouk to accept an-Naḥḥās as prime minister. The Wafd, its power confirmed by overwhelming success in the general election of 1942, cooperated with Britain. Nevertheless, the intervention in February 1942 had disastrous consequences. It confirmed Farouk's hostility to both the British and an-Naḥḥās and tarnished the Wafd's pretensions as the standard-bearer of Egyptian nationalism. The Wafd was damaged also by internal rivalries.

An-Naḥḥās was dismissed by the King in October 1944. His successor, Aḥmad Māhir, was acceptable to the British, but he was assassinated in February 1945 at the moment of Egypt's declaration of war on Germany and Japan. He was succeeded by a Saʿdist, an-Nuqrāshī Pasha.

At the end of World War II, Egypt was in a thoroughly unstable condition. The Wafd declined, and its political opponents took up the nationalist demand for a revision of the treaty of 1936—in particular for the complete evacuation of British troops from Egypt and the ending of British control in the Sudan. Politics were passing into the hands of radicals. The Muslim Brotherhood, founded in 1928, developed from an orthodox Islāmic reformist movement into a militant mass organization. Demonstrations in Cairo became increasingly frequent and violent. The pressure rendered it impossible for any Egyptian government to attempt a settlement of its two main external problems: treaty revision with Britain, and Palestine. Negotiations with Britain, undertaken by an-Nuqrāshī and (after February 1946) by his successor, Ṣidqī, broke down over the British refusal to prejudice the possible independence of the Sudan. Although Egypt referred the dispute to the United Nations in July 1947, the deadlock continued.

Anti-British demonstrations

Until the interwar period neither the Egyptian public nor the politicians had shown much interest in Arab affairs generally; Egyptian nationalism had developed as an indigenous response to local conditions. After 1936, however, Egypt became involved in the Palestine problem, and in 1943–44 it played a leading part in the formation of the Arab League. After World War II, Egypt became increasingly committed to the Arab cause in Palestine, but its unexpected and crushing defeat in the first Arab–Israeli War (1948–49) contributed to disillusionment and political instability. The Muslim Brotherhood increased its terrorist activities. An-Nuqrāshī, again prime minister, endeavoured to suppress the organization and was assassinated in December 1948. The brotherhood's leader was murdered two months later.

A general election in January 1950 gave the Wafd a majority, and an-Naḥḥās again formed a government. Failing to reach agreement with Britain, in October 1951 he abrogated both the 1936 treaty and the Condominium Agreement. Anti-British demonstrations were followed by guerrilla warfare against the British garrison in the Canal Zone. British military action in Ismailia was followed on January 26, 1952, by the burning of Cairo by demonstrators. An-Naḥḥās was dismissed, and there were four prime ministers in the ensuing six months.

THE REVOLUTION AND THE REPUBLIC

The Wafd and the old political parties were thoroughly discredited, the King hardly less so. Since late 1949 a number of young officers had been organizing a conspiratorial movement. The King became aware of the challenge at the end of 1951, and when he tried to regain control of the situation by appointing his brother-in-law as war minister, the Free Officers (as they were now known) carried out an armed coup in July 1952. Farouk was forced to abdicate and leave Egypt, his infant son being proclaimed king as Fu'ād II with a council of regency.

Abdication of Farouk

Struggle for power. For the next 12 months the chief Free Officers, organized as the Revolution Command Council (RCC), groped their way toward the establishment of a new regime. Their leader, Col. Gamal Abdel Nasser, was at first less prominent than an older officer, Gen. Mohammad Naguib. In September 1952, Naguib took over from ʿAlī Māhir, the last civilian prime minister, and two days later an Agrarian Reform Law was decreed. This was not only a token measure of radical social legislation but also an attack on the landholders who were closely associated with the old regime. Meanwhile, both the Muslim Brotherhood and the old political parties were active. In December 1952 the constitution of 1923 was abolished, and a month later the parties were dissolved and their funds seized. To build up mass support, an organization headed by Nasser and styled the Liberation Rally was set on foot. Finally, in June 1953, Egypt was formally declared a republic, Naguib combining the posts of president and prime minister, with Nasser as his deputy and minister of the interior. Publicized trials of former politicians and the propagandist activities of a Ministry of National Guidance were used to strengthen the new regime.

The Muslim Brotherhood had hoped to profit from the downfall of the monarchy and of the old parties to acquire power, and it was a serious challenge to the RCC. An attempt by a Muslim Brother to assassinate Nasser in October 1954 was followed by rigorous measures against the movement, and its power was at last broken. In the meantime, a division had appeared in the RCC itself, between a faction associated with Naguib and another headed by Nasser. An attempt to remove Naguib in February 1954 proved premature, but Nasser finally became prime minister in April. Naguib was still president, but he was dismissed and held under house arrest in November. The recognition of Nasser's supremacy came with the promulgation of a new constitution in January 1956, establishing a presidential system of government. In the following June a plebiscite approved the constitution and formally elected Nasser as president.

Ascendancy of Nasser

The abolition of the old regime and the creation of a new political establishment in Egypt between 1952 and 1956 was accompanied by analogous developments in foreign affairs. Coming to power without previous commitments, and with the means to dominate the manifestation of public opinion imprudently encouraged by its predecessors, the military government was able rapidly to reach a settlement with Britain of the two problems of the Sudan and the Canal Zone. An agreement signed in 1953 established a transitional period of self-government for the Sudan, and it became an independent republic in January 1956. Its subsequent relations with Egypt have varied in cordiality. A second Anglo-Egyptian agreement, in 1954, terminated the 1936 treaty and provided for the British evacuation of the Canal Zone. In June of the following year, the last British troops left Egypt.

Egypt and the U.A.R. Signs were already appearing of a rift between Egypt and the West. The negotiation by Britain of the Baghdad Pact with Pakistan, Iran, Turkey, and Iraq was seen by the military government as an attempt to isolate Egypt. In reaction, Nasser turned in 1955 to association with the Afro-Asian states in the Bandung Conference and obtained arms from Czechoslovakia. Matters came to a head in the following year, when, in reply to the denial of funds previously promised by the U.S. and Britain for the construction of a high dam at Aswān, Nasser announced the nationalization of the Suez Canal and expropriated the company. The ensuing crisis culminated in hostilities against Egypt by Israel, threatened by Egypt's emancipation from British control; by France, alarmed at Egyptian aid to the revolt in Algeria; and by Britain, resentful of the policies pursued by the military government. Pressure on Britain and France by the U.S. and the Soviet Union ended the brief Suez War (October 29–November 7, 1956), but a casualty was the

Suez War

Anglo-Egyptian agreement on the Canal Zone, abrogated by Egypt early in 1957.

The unsuccessful Franco-British intervention increased Nasser's prestige in Egypt and the Arab world generally, and from that date he confidently pursued for a decade a more radical line in both domestic and foreign policy. His ambitious attempts to acquire a dominating position in the Islāmic world were disappointing. Early in 1958 Syria combined with Egypt to form the United Arab Republic (U.A.R.), but it was a reluctant marriage of convenience and was dissolved in September 1961, with Egypt retaining the new designation. After the secession of Syria, an elected National Congress the next spring approved a National Charter, which set up still another organization to associate the masses with the government, the Arab Socialist Union (ASU). In March 1964 a new National Assembly was elected, and a new constitution (again placing supreme power in the hands of the president) was proclaimed by Nasser.

Throughout this period, the potential military danger from Israel was an important factor in the calculations of the U.A.R. government. Fear of Israel was a motive in turning the U.A.R. toward the Soviet bloc and led to a series of initiatives for cooperation among the Arab states, which, however, had little practical outcome. For 10 years peace on the borders of Israel was precariously maintained by the presence of observers from the United Nations. When, in June 1967, they were withdrawn from the U.A.R. frontier on Nasser's demand, the Israeli government launched a pre-emptive strike on the U.A.R. and Jordan. When the Six-Day War ended, Israeli forces occupied the whole of the Sinai Peninsula. The Suez Canal ceased to be a strategic and commercial artery and became a fortified ditch between the two belligerents. Small-scale military activity continued during the following years, with the U.A.R. becoming steadily more dependent upon the Soviet Union to safeguard its cities and air space against Israeli attack. In the summer of 1970, pressure by the U.S. and the Soviet Union on Israel and the U.A.R., respectively, brought both sides to initiate negotiations through the UN. The talks, however, broke down; and in September a new crisis appeared in Jordan, where operations of King Hussein's forces against the Palestinian guerrillas and refugees resulted in civil war. Nasser's mediation secured an agreement between Jordan and the Palestine Liberation Organization (PLO), but Nasser died a few hours later (September 28, 1970).

Six-day Arab–Israeli war

(P.M.Ho.)

Sadat succeeds Nasser. Nasser was succeeded as president by his friend and colleague Anwar as-Sadat. After nearly 18 years of strong leadership by one man, Sadat had to shift policies and governments frequently. This period was marked by a gradual lessening of restrictions, including the reformation of the ASU and a new constitution in 1971 and the abolition of censorship in 1974; this was not accomplished, however, without political opposition and street demonstrations.

Sadat completed negotiations with Libya and Syria (Syria replaced The Sudan as the originally planned third partner) in the federation of the Arab Republic of Egypt in September 1971. It was agreed that Cairo would serve as the capital and Sadat as the head of a three-man Presidential Council. Relations among the partners, however, and especially between Egypt and Libya, continued under frequent strain, and by 1974 the federation virtually ceased to exist.

Although Pres. N.V. Podgorny of the Soviet Union had joined Sadat in the opening of the Aswān High Dam in January 1971, and the Soviet Union had agreed to finance the resulting rural electrification program, Egyptian–Soviet relations deteriorated to the extent that in July 1972 Sadat announced the withdrawal of all Soviet military advisers and personnel. In 1973 Sadat made progress in his relations with West Germany, Jordan, and Saudi Arabia; but in the period of détente between the U.S. and the Soviet Union the "no peace, no war" policy with Israel appeared likely to perpetuate the Israeli occupation of Egyptian and Syrian territories.

Egyptian–Soviet relations

Therefore, after frequent, but unfulfilled, warnings, on October 6, 1973, Egypt and Syria attacked Israel on both fronts. They met with initial success, but by October 23 the Egyptian Third Army was surrounded on the west bank of the Suez Canal, and the Israelis had driven the Syrians to within 20 miles of Damascus. UN peacekeeping forces were brought in, and beginning October 28 Egyptians and Israelis negotiated directly for the first time since 1956. A Middle East peace conference began in December at Geneva, and Henry A. Kissinger, the U.S. secretary of state, became the leading arbitrator; agreements were reached on January 18, 1974, on the Suez front and May 29 on the Syrian–Israeli border. The resulting disengagement changed little except that the canal was cleared of mines and sunken ships and was reopened on June 5, 1975, the eighth anniversary of its closing.

Tensions between Egypt and the Arab world, on one side, and Israel, on the other, continued, with the question of the Palestinian refugees a major issue. Sadat played an important role in 1974 in persuading King Hussein of Jordan to recognize the PLO as representing all Palestinians.

Kissinger continued to serve as a mediator, resulting in a pact signed in Geneva in September 1975, with Egypt renouncing the use of force and guaranteeing Israel merchant-shipping rights to the Suez Canal in return for a further Israeli withdrawal in the Sinai.

In 1976 the Egyptian economy was plagued by a large trade deficit and foreign debt and an annual rate of inflation of about 30 percent. Despite these difficulties, Sadat was re-elected on September 16 and the ASU won more than two-thirds of the seats in parliamentary elections. Economic unrest culminated in major rioting on January 18–19, 1977, in Cairo, Alexandria, and other Delta cities. Troops were called in to quell the disturbances, and 1,600 persons were arrested. The government then attempted to tighten its control of political activity, and on June 29 a law was passed to reform the country's political structure.

Election of 1976

Sadat initiated a move to bring peace to the Middle East when he spoke before the Israeli Knesset in Jerusalem on November 20, 1977. This was followed by his announcement on November 27 of a conference to be held in Cairo to clear away obstacles to the resumption of the Geneva conference. These moves were strongly opposed by Algeria, Libya, Syria, Iraq, Yemen (Aden), the PLO, and the Soviet Union but were supported by Israel and the United States. The conference opened in Cairo on December 14, with Egypt and Israel participating, and the Israeli prime minister, Menahem Begin, met with Sadat in Ismailia on December 25–26. Negotiations continued into 1978, with Sadat maintaining that resolution of the claims of the Palestinians was essential to any total peace settlement.

(Ed.)

BIBLIOGRAPHY.

Ancient Egypt: *The Cambridge Ancient History,* 3rd ed., vol. 1 and 2, each in 2 pt. (1970–71), an up-to-date account, newly written by a team of international scholars—for Egypt it takes the story down to the end of the 21st dynasty; ETIENNE DRIOTON and JACQUES VANDIER, *L'Égypte,* 4th ed. (1962), a student's history, with excellent detailed notes on particular problems; SIR ALAN H. GARDINER, *Egypt of the Pharaohs* (1961), a very personal history, notable for the use made of Egyptian texts, by one of the greatest modern scholars; WILLIAM C. HAYES, *The Scepter of Egypt,* 2 pt. (1953–59), a cultural history of Egypt down to the end of the 20th dynasty, based on the collections in the Metropolitan Museum of Art, New York, by the leading U.S. scholar of the postwar period; WILLIAM C. HAYES, *Most Ancient Egypt* (1965), a fragment of a general history of Egypt that deals brilliantly with the earliest periods down to Badarian times; HERMANN KEES, *Das alte Ägypten,* 2nd ed. (1958; Eng. trans., *Ancient Egypt,* 1961), a cultural topography of ancient Egypt in which all of the great cities and districts are treated (a mine of information); W.S. SMITH, *The Art and Architecture of Ancient Egypt* (1958), the best general account of Egyptian art available, especially good for the Old and Middle Kingdoms.

Egypt from the 18th dynasty to 332 BC: Basic general histories with chapters on the 18th dynasty until 332 BC are SIR ALAN H. GARDINER, *Egypt of the Pharaohs* (1961); ETIENNE DRIOTON and JACQUES VANDIER, *L'Égypte,* 4th ed. (1962), containing extensive bibliography; and WOLFGANG HELCK (ed.), *Geschichte des Alten Ägypten* (1968). In JOHN A. WILSON,

The Burden of Egypt (1951), the discussion of this period is more selective and interpretative. The history and culture of the New Kingdom receive excellent treatment in GEORGE STEINDORFF and KEITH C. SEELE, *When Egypt Ruled the East*, 2nd ed. (1957); and WILLIAM C. HAYES, *The Scepter of Egypt*, pt. 2 (1959). *The Cambridge Ancient History*, rev. ed., vol. 2, ch. 8, 9, 19, 23, and 35 issued in fascicles (1962–66), are detailed discussions by various scholars on certain phases of New Kingdom history. Indispensable for foreign relations during the New Kingdom is WOLFGANG HELCK, *Die Beziehungen Ägyptens zu Vorderasien im 3. und 2. Jahrtausend v. Chr.* (1962). The rise of the New Kingdom is admirably treated in JURGEN VON BECKERATH, *Untersuchungen zur politischen Geschichte der zweiten Zwischenzeit in Ägypten*, ch. 5 (1964). DONALD B. REDFORD, *History and Chronology of the Eighteenth Dynasty of Egypt: Seven Studies* (1967), includes an excellent chapter re-evaluating Hatshepsut. An informative account of the height of the empire is ELIZABETH RIEFSTAHL, *Thebes in the Time of Amunhotep III* (1964). Both CYRIL ALDRED, *Akhenaten, Pharaoh of Egypt* (1968); and F.J. GILES, *Ikhnaton: Legend and History* (1970), contain interesting interpretations of the controversial Amarna period, though a long co-regency period is assumed. A comprehensive account of Tutankhamen and his tomb is CHRISTIANE DESROCHES-NOBLECOURT, *Vie et mort d'un pharaon: Toutankhamon* (1963; Eng. trans., *Tutankhamen: Life and Death of a Pharaoh*, 1963). Since the dated discussion of JURGEN VON BECKERATH, *Tanis und Theben* (1951), no comprehensive work on the Ramesside period has appeared. EDWARD F. WENTE, *Late Ramesside Letters* (1967), deals with material from the very end of the period. A good general work treating Egyptian society and culture is CYRIL ALDRED, *The Egyptians* (1961). WOLFGANG HELCK, *Zur Verwaltung des mittleren und neuen Reichs* (1958), is basic for the administration and civil service, while HERMANN KEES, *Das Priestertum im ägyptischen Staat, vom neuen Reiche bis zur Spätzeit* (1953), is a scholarly analysis of the priesthoods, more popularly treated in SERGE SAUNERON, *Les Prêtres de l'ancienne Égypte* (1957; Eng. trans., *The Priests of Ancient Egypt*, 1960). For the period of the 21st to 25th dynasties, HERMANN KEES, *Die Hohenpriester des Amun von Karnak von Herihor bis zum Ende der Äthiopenzeit* (1964), is important. The best work on the Saite 26th dynasty until Alexander is F.K. KIENITZ, *Die politische Geschichte Ägyptens vom 7. bis zum 4. Jahrhundert vor der Zeitwende* (1953), with Egyptian and classical sources.

Hellenistic and Roman Egypt: The most important works on this period include: H.I. BELL, *Egypt, from Alexander the Great to the Arab Conquest* (1948); E.R. BEVAN, *A History of Egypt Under the Ptolemaic Dynasty* (1927); AUGUSTE BOUCHE-LECLERCQ, *Histoire des Lagides*, 4 vol. (1903–07); N. LEWIS, "Greco-Roman Egypt, Fact or Fiction?," *Proceedings of the Twelfth International Congress of Papyrology*, American Studies in Papyrology VII, pp. 3–14 (1970); M.I. ROSTOVTZEFF, *The Social and Economic History of the Hellenistic World*, 3 vol. (1941); *The Social and Economic History of the Roman Empire*, 2nd ed., 2 vol. (1957); C.B. WELLES, *Alexander and the Hellenistic World* (1970); J.G. MILNE, *A History of Egypt Under Roman Rule*, 3rd ed. rev. (1924); A.H.M. JONES, *The Later Roman Empire, 284–602: A Social, Economic, and Administrative Survey*, 2 vol. (1964); CLAIRE PREAUX, *L'Économie royale des Lagides* (1939); J. MODRZEJEWSKI, "La Règle de droit dans l'Égypte ptolémaïque," *Essays in Honor of C. Bradford Welles*, American Studies in Papyrology I, pp. 125–173 (1966); "La Règle de droit dans l'Égypte romaine," *Proceedings of the Twelfth International Congress of Papyrology*, American Studies in Papyrology VII, pp. 317–377 (1970); LUDWIG MITTEIS and U. WILCKEN, *Grundzüge und Chrestomathie der Papyruskunde* (1912); M.I. ROSTOVTZEFF, *A Large Estate in Egypt in the Third Century B.C.* (1922); and S.L. WALLACE, *Taxation in Egypt From Augustus to Diocletian* (1938).

Egypt from c. 630–c. 1800: There are two standard works that survey medieval Egyptian history as a whole: STANLEY LANE-POOLE, *A History of Egypt in the Middle Ages* (1901); and GASTON WIET, *L'Égypte arabe de la conquête arabe à la conquête ottomane, 642–1517 de l'ère chrétienne*, vol. 4 of *Histoire de la nation égyptienne*, ed. by GABRIEL HANOTAUX (1938). Each of these is outdated in many respects, but each has the virtue of presenting an accurate summary of the political history of the period, based on primary Arabic sources; also, both are strong on Egyptian architecture as an insight into political, social, and economic history. For the economic history, SUBHI LABIB, *Handelsgeschichte Ägyptens im Spätmittelalter, 1171–1517* (1965), is a pioneering work; Labib has summarized this book in English in the form of an article: "Egyptian Commercial Policy in the Middle Ages," in *Studies in the Economic History of the Middle East*, ed. by M.A. COOK, pp. 63–77 (1970). Labib's book should be supplemented by another article: ROBERT LOPEZ, HARRY MISKIMIN, and ABRAHAM UDOVITCH, "England to Egypt, 1350–1500: Long-Term Trends and Long-Distance Trade," *Studies in the Economic History of the Middle East*, pp. 93–128 (1970). AZIZ S. ATIYA, *A History of Eastern Christianity* (1968), is authoritative for Coptic history. For the beginnings of Muslim Egypt, ALFRED J. BUTLER, *The Arab Conquest of Egypt and the Last Thirty Years of the Roman Dominion* (1902), has not yet been superseded; however, it should be supplemented by FRANCESCO GABRIELI, *Maometto e le grandi conquiste arabe* (1967; Eng. trans., *Muhammad and the Conquests of Islam*, 1968), for the conquest; and by DANIEL C. DENNETT, *Conversion and the Poll Tax in Early Islam* (1950), for Muslim tax policy in Egypt. For the Ṭūlūnids, there is only ZAKY MOHAMED HASAN, *Les Tulunides* (1933). MARIUS CANARD, *Les Institutions des Fâtimides en Égypte* (1957), is good for Fāṭimid history, but Fāṭimid studies have been recently transformed by S.D. GOITEIN with the publication of his *Mediterranean Society: The Jewish Communities of the Arab World as Portrayed in the Documents of the Cairo Geniza*, vol. 1, *Economic Foundations* (1967). Three articles by H.A.R. GIBB are definitive for Egypt under the Ayyūbids and during the Crusades, all published in *A History of the Crusades*, ed. by KENNETH M. SETTON, 2nd ed., 2 vol. (1969): "The Caliphate and the Arab States," 1:81–98; "The Rise of Saladin," 1:563–89; "The Aiyūbids," 2:693–714. For Mamlūk and Ottoman Egypt, see the bibliography for the article MAMLUKS; this should be supplemented by F.R.C. BAGLEY, "Egypt and the Eastern Arab Countries in the First Three Centuries of the Ottoman Period," in *The Muslim World*, vol. 3, ed. by F.R.C. BAGLEY, pp. 50–96 (1969).

Egypt since 1800: J.M. AHMED, *The Intellectual Origins of Egyptian Nationalism* (1960), particularly concerned with the nationalists of the period from 1892 to 1914, by a Sudanese scholar; GABRIEL BAER, *A History of Landownership in Modern Egypt 1800–1950* (1962), essential to the understanding of the social and economic history of modern Egypt; P.M. HOLT, *Egypt and the Fertile Crescent, 1516–1922* (1966), outlines the political history of modern Egypt in relation to that of neighbouring Arab lands; P.M. HOLT (ed.), *Political and Social Change in Modern Egypt* (1968), a collection of papers by various scholars on aspects of political, social, and economic history; J.C. HUREWITZ, *Middle East Politics: The Military Dimension* (1969), a general study of the phenomenon of military rule in the Middle East, with a section on Egypt since the revolution of 1952; JACOB M. LANDAU, *Parliaments and Parties in Egypt* (1953), an important investigation of constitutional and political developments under the khediviate and monarchy; HELEN ANNE B. RIVLIN, *The Agricultural Policies of Muḥammad ʿAlī in Egypt* (1961), covers a wider range of topics than the title suggests—*e.g.*, the administrative, technological, and military developments under Muḥammad ʿAlī; R.L. TIGNOR, *Modernization and British Colonial Rule in Egypt, 1882–1914* (1966), an examination of the British system of government in Egypt; P.J. VATIKIOTIS, *The Modern History of Egypt* (1969), a comprehensive and detailed study of particular value for recent times.

(T.G.H.J./E.F.W./A.E.S./D.P.L./P.M.Ho.)

Egyptian Law

Egyptian law is the law that evolved from the beginnings of Nile civilization in the 4th millennium BC down to the era of Roman occupation, from 30 BC. Thus, Egyptian law has a history of some three millennia—longer than that of any other civilization. Its earliest traces date from the time of the unification of Upper and Lower Egypt (roughly corresponding to the Nile Valley and the Delta) under King Menes, about 3100 BC. He stood first in the long line of pharaohs constituting 31 dynasties, the most important of which are traditionally grouped into the Old Kingdom (*c.* 2686–*c.* 2160 BC), the Middle Kingdom (*c.* 2040–1786), and the New Kingdom (1567–1085). Between these high moments there were periods of decline in political institutions and in both public and private law. In the period following the New Kingdom (or Empire), the central authority weakened, and the country split up into small states, with a theocratic regime in Upper Egypt. Egypt was successively conquered by Ethiopians, Assyrians, Persians, and Alexander the Great. When one of Alexander's Macedonian generals, Ptolemy, seized the Egyptian throne in 305 BC, there was inaugurated a period of Hellenistic influence, followed by a period of Roman influence, with certain aspects of Egyptian law slowly merging with Greek and

Roman imperial law. Evidence from the late Roman period (3rd to 5th centuries AD) shows, however, that, even though Egyptian law was submerged for many centuries in the great cities, it had been retained, along with the Egyptian language, in the country as a whole.

The administration of ancient Egyptian law was closely related to the institution of kingship and the concept of a divine pharaoh. The pharaoh's will was supreme; government was by royal decree. The complexity and all-embracing nature of the king's official duties as head of all branches of administration, however, forced him, at an early period, to delegate his powers. From the period of the Old Kingdom these delegated powers were largely in the hands of province governors. But the most powerful man in the state, next to the pharaoh himself, had become the vizier, who oversaw virtually every branch of state administration, domestic and foreign. In the area of law, he tried cases and appointed lower magistrates. In his office were filed records of every department of government, private wills and contracts, and royal decrees.

Sources of Egyptian law

Although, according to Greek historians, several pharaohs had reputations as lawgivers, no Egyptian code of laws has been preserved. Before the Late Period, only traces of law from writings on papyri and stone exist, together with remnants of such items as accounts, deeds, and contracts. Excerpts from acts that have been saved are imprecise, because the Egyptians reproduced only the essentials, and the essentials are at times so brief as to inspire considerable argument among modern historians. With the advent of the demotic language in the 7th century BC, documents became more numerous, because many transactions that had previously been accomplished orally now required written deeds. Yet there still is no code of law, but it may perhaps be possible in the future to reconstruct such a code from the hundreds of legal documents discovered, many of which remain to be published. From the information that is available, it is possible to know some elements of the legal relationships that existed among the people of the Nile Valley during their long history.

SUBJECTS AND INSTITUTIONS OF EGYPTIAN LAW

The law of persons. *Status.* In Egypt both public and private law emphasized the importance of the individual, and, as a result, women as well as men had full rights under the law. A woman could own and alienate property, be a party to a contract, file suit, make a will, and bear witness without the assistance of a father or husband. These rights were not exclusively limited to the wealthier classes. Workers, too, had certain rights of possession, alienation, suit, witness, and contract. Even slaves, although they could be purchased as merchandise, had a measure of civil status, since they could possess property and dispose of it as they wished.

Marriage. In Egypt, marriage was monogamous, although exceptions were made for the king and princes during feudal times. Each party had definite rights within the marriage, so it is likely that the union was brought about without restrictions being imposed by the respective families. The marriage contract, on the other hand, was very strict. The couple set up certain property as conjugal property, in which the man had an interest in two-thirds, the woman in the remainder. The man administered this property and saw that acquisitions were distributed in accordance with prescribed proportions. In addition to this, each party was able to possess his or her own property.

Having children was the purpose of marriage, and those who did not have any were advised to adopt. In a particular case, three children were born to a female slave, the father apparently being the master of the household. Since he had no children within his marriage, his wife adopted the children born to the slave, after his death. From the moment of adoption they were automatically free, without having to submit to any emancipation procedures (the same was true when a slave married a free woman). In addition, the children became her heirs.

It is necessary to point out that the slave who bore the children had been purchased during the marriage and constituted an acquisition of the joint estate. The wife was able to adopt the children only because she had been made universal heir by her husband; otherwise, she would have been able to have only one-third of the slaves and one-third of the offspring at her disposal, with the property representing the husband's two-thirds passing to his nearest blood relative.

Divorce. There is no evidence that divorce occurred before the Late Period (1085–332 BC). When its occurrence is recorded, it appears to have been accomplished with a minimum of discomfort to the parties and with considerable respect for the rights of each. The Egyptian method of divorce apparently influenced the practices of diverse peoples living under Egyptian rule. Aramaic texts from the city of Elephantine in the 5th century BC show that members of the Jewish colony gave up divorce by repudiation in order to practice that pronounced by judicial decision.

Family and succession. It is not known at what age children reached legal majority, but paternal authority over them then ceased to exist. In certain periods this authority, as well as the family property, passed on to the eldest son; but at other times masculine primogeniture was not followed, and the inheritance was divided equally among the children. At the very end of the 18th dynasty (1567–1320 BC), because of the individualistic pressures of the time, the vizier was forced to allow the division of a military benefice among the various members of a family, even though the benefice had been transmitted through one member since the beginning of the dynasty. Even when masculine primogeniture predominated, there is evidence that the remaining children as well as the spouse had some guarantee of receiving a share in the estate. The wife could become universal heir only if there were no children, although she was always permitted at least a partial legacy. The property of the wife passed to her children at the moment of her death, with the exception of that part of the property acquired in common, which was to be maintained for the enjoyment of the surviving spouse.

Egyptians were able to will part of their belongings, with the surplus being governed by the law of succession. Men and women could depart from this law by drawing up an act called an *imyt-per* or enregistered document. There is, for example, a record of a woman disinheriting some of her children; and there is a record of a man favouring one of his daughters by explicitly upholding her rights over the part of his property that should have gone to the rest of his children.

The house document

Children, as coheirs of their parents, were responsible for the cost of burying them, even if they were not charged with the actual act. For this reason, most people set up perpetual foundations to pay for these expenses and those associated with maintaining the tomb after death. Often members of the family would be appointed "funerary priests" with the responsibility for tomb maintenance; they would have usufructuary rights (rights of use and enjoyment) in the products of the foundation but could not divide it among their children.

Property and contracts. Each person was able to dispose of his property as he wished, even to the extent of making such property inalienable, as with a perpetual foundation for the maintenance of one's tomb. Originally, transfer of property was accomplished orally before a local, or the central, bureau, and recorded by a scribe, and the earliest documents of transfer specified a definite price. The agreement took place without actual payment or delivery, but it is possible that at least one of the two operations was required in order to make the agreement perfect. Later, with the rise of the demotic language, different procedures were initiated; every buyer had to have in his hands a written deed that gave to him the right of ownership and another that ensured him the enjoyment of it.

In Egyptian law, property transfers were not separated from contractual agreements. Indeed, such things as the hiring out of slaves for a day's work were considered to be sale, and work was freely traded for cattle and other commodities. Any agreement created certain obligations,

and the law that was to be applied to them was determined by the wishes of the parties involved. Guarantees were offered against defects in the property and restrictions in its use and against defects in the law.

Court procedures. Egyptian judicial procedure was accusatory, and it was the responsibility of the plaintiff to bring the matter before the tribunal. If the court held the issue to be admissible, the defendant was cited to appear before the court. The two parties acted for themselves without the aid of professional advocates, but the scribes working for the judicial administration were able to furnish necessary information on procedures and other matters. After both parties spoke, documentary information was presented, then both parties spoke again; debate ended at this point unless there was a call for witnesses. The decision was made, as a rule, on the documents; if convincing ones were not available, then the judge ruled on previous testimony. The judgment settled the issue, explaining the grounds, the legal foundation for the decision, and the disposition of the case, along with the administrative procedures for the keeping of the written record of the case.

Criminal law and tort. A judicial hierarchy existed in the area of criminal justice. Certain lesser crimes could be handled by lower officials, whereas more severe ones were taken directly to the vizier; still others could be handled only by the king. An example of this hierarchy can be seen in records dealing with the workers at the Theban necropolis. The workers evidently had the right to group themselves into an assembly that had administrative and judicial powers. In one particular case they took some workers who had slandered their chief and sentenced them to be beaten with sticks. In the case of a woman who had stolen the tool of a worker, and things in a chapel, and then committed perjury, the workers could only conduct the investigation, because the case exceeded their competence. In very serious crimes, this procedure was repeated at a higher level, with the vizier directing the investigation and then turning the matter over to the king for judgment. There were even instances when the king would appoint a special commission to sit in judgment. One such case occurred when there was a plot against Ramses III (ruled 1198–1166 BC), who nominated a special commission to whom he gave full responsibility to deal with the matter. He told the commission members that they were to judge forcefully but should not depart from strict justice.

In general, punishment was severe; beatings and floggings were commonplace, as was mutilation. For more serious crimes, penal servitude and execution were resorted to.

CHARACTERISTICS AND IMPORTANCE OF THE EGYPTIAN LEGAL SYSTEM

Because there is no extant code and few commentaries, any theory of Egyptian law as such must be ascertained from those documents available. One such document, dating from the beginning of the second millennium BC, contains instructions from the king to his vizier, who is told to judge all suits in conformity with the law. Since the laws were general, the vizier had a certain amount of latitude and was able to employ a degree of equity. Still, it seems that he was accountable to the public and was in some cases to use previous judgments as sources for his decisions.

Ancient Egyptian law was remarkable in terms of its lines of application, since it had as its primary objective the promotion of basic human rights. There were, certainly, periods of regression, but these were balanced by others in which great forward strides were made. It is interesting that the work of Bocchoris (last king of the 24th dynasty, 8th century BC), who was deemed a great lawgiver, occurred during a period of social and economic stagnation when a large part of the population had the status of tenant on the great domains owned by the temples. Yet Bocchoris freed the individual, suppressed imprisonment for debt, prohibited the interest of a debt from doubling the capital, and fixed the rules of written contracts; he also initiated certain reforms dealing with the mobility of private property. All these changes were eminently important, but corresponded with the spirit of the Egyptian law in the high points of its history.

Furthermore, Egyptian law had a definite effect on Greek law during the Hellenistic period and later on Roman imperial law. Even before this time, Egyptian influence had been felt on Greek law, for in the early 6th century BC Solon, an early Athenian lawgiver, came to Egypt and was inspired by the legislative work of Bocchoris. The result was a decisive turn in Greek law, although the Greeks never emancipated women as had the Egyptians.

BIBLIOGRAPHY. J. PIRENNE and ARISTIDE THEODORIDES, *Droit égyptien, introduction bibliographique* (1966), contains a thorough bibliography relating to the sources of ancient Egyptian law and the authors who have dealt with the subject, with commentaries and evaluations. J. PIRENNE, *Histoire des institutions et du droit privé de l'ancienne Égypte*, 3 vol. (1932–35), is the major historical description of Egyptian law, although the same author's *Histoire de la civilisation de l'Égypte ancienne*, 3 vol. (1961–63), is a later work containing information on law. E. SEIDL has offered a systematic description of the subjects relating to the private law in *Einführung in die ägyptische Rechtsgeschichte bis zum Ende des Neuen Reiches*, 2nd ed. (1951); *Ägyptische Rechtsgeschichte der Saiten- und Perserzeit*, 2nd ed. (1968); and *Ptolemäische Rechtsgeschichte*, 2nd ed. (1962). Each year since 1957 in the *Revue Internationale des Droits de l'Antiquité*, ARISTIDE THEODORIDES has examined the documents of Egyptian law, with a view toward forming a "corpus" of the texts and a dictionary of technical terms. The juridical texts of the Old Kingdom have been brought together again, with translation, explanations, and commentary, by H. GOEDICKE in *Königliche Dokumente aus dem Alten Reich* (1967) and *Die privaten Rechtsinschriften aus dem Alten Reich* (1970). See also P.W. PESTMAN, *Marriage and Matrimonial Property in Ancient Egypt* (1961); and TYCHO MRSICH, *Untersuchungen zur Hausurkunde des Alten Reiches: Ein Beitrag zum altägyptischen Stiftungsrecht* (1968).

(A.E.T.)

Egyptian Religion

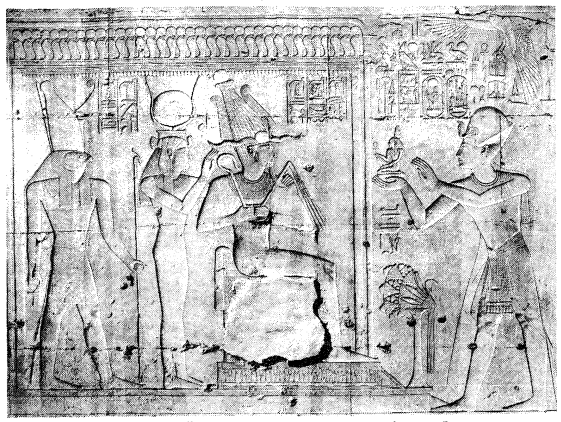
The term Egyptian religion here refers to the indigenous religion of ancient Egypt from the Late Neolithic Period to the first centuries AD, and includes both folk traditions and the court religion of the pharaohs.

NATURE AND SIGNIFICANCE

Egyptian religion shared two characteristics with Egyptian culture in general: longevity and a development that, on the whole, was undisturbed by foreign influence. All manifestations of cultural and social life were so deeply permeated with religious ideas that an understanding of Egyptian culture is impossible without an understanding of Egyptian religion, and vice versa. Egyptian religion had a widespread influence on the neighbouring cultures of its time and on the cultures of later eras. One of the most striking aspects of Egyptian religion was the immense number of gods in human, animal, and material forms, who represented both the powers of nature and abstract ideas. They were never grouped systematically, though tendencies toward systematization were manifest in local groupings of gods and mythical associations. Another characteristic was their lack of distinct personality; *i.e.*, one god could easily be equated with another on the basis of one common feature or embodied power. In addition, they were all immanent, and there was no contrast between the divine and the world. The divinity of the world (cosmos, nature, mankind, society) implied a view of life that extended beyond physical life. The state of death was regarded as part of the well-ordered life cared for by the gods. Various material and ritual guarantees for life after death aided in the realization of this idea.

The divine world order

Thus the world appeared as a large, permanent unity in spite of the multitude of gods and their cults. It was regulated by the rule of pharaoh, whose main task is to establish *maʿat* instead of disorder. *Maʿat* stands for all order in life. The translations "order," "justice," and "truth" indicate its diverse aspects. In social life the demands of ethics help to realize *maʿat*.



King Seti I offering a figure of the goddess Ma'at to Osiris, Isis, and Horus. Relief from the temple of King Seti I, Abydos, *c.* 1300 BC.

By courtesy of the Egyptian Exploration Society; photograph, Oriental Institute, Chicago

Characteristic of Egyptian religious thinking and national character was the toleration of ideas that, from the point of view of later cultures, contradict each other. This toleration was accompanied by the desire to retain ideas and forms once conceived, even when they had been replaced by new ones. As a consequence, the number of mythical, ritual, and theological creations increased in the course of time. It is difficult, therefore, to fix the time of origin and the primary life span of an idea that has become fixed in tradition.

The basic feature of religious feeling was optimism. Notions of an end of the world and skepticism rarely arose, as death was viewed as part of life and the realization of *ma'at* was always possible. Both magic and ritual acts were means to preserve the security of life, as they protected against danger and menace. Egyptian religion can therefore be classified among the typical ritual religions in which faith is less important than correct rites.

SOURCES AND THEIR INTERPRETATION

Tombs

Nonliterary. Tombs provide sources for the knowledge of Egyptian religion in prehistoric as well as later times. Their offering places indicate belief in an afterlife: the dead need offerings of food and drink. A false door enables communication with the world of the living.

Funerary gifts and paintings. Real objects or models provided the deceased with household goods, articles of daily use, and servants, as in life. In the late Old Kingdom after *c.* 2250 BC, and in the First Intermediate Period (*c.* 2160–*c.* 2040 BC), friezes of objects (*i.e.*, paintings of funerary gifts) were added. In this way, it was thought, the property of the dead could be preserved in case the tomb was plundered. From the early 4th dynasty (*c.* 2613–*c.* 2494 BC) on, tomb decorations (reliefs or paintings) occur. The main subjects are the so-called offering lists (lists of the food, drink, clothes, oils, and ointments that the dead needed) and representations of offering rituals and revivification ceremonies (*e.g.*, the "Opening of the Mouth," from the 18th dynasty—1567–1320 BC). In addition, there are the so-called scenes from life (agriculture, cattle breeding, crafts) and, occasionally, from the 6th dynasty (*c.* 2345–*c.* 2181 BC) on, representations of biographical events. Ritual scenes are not represented in tombs, and the king does not appear until the 18th dynasty. In the Old Kingdom (*c.* 2686–*c.* 2160 BC), the pyramid temples of the kings attest the existence of a royal funerary cult similar to that of private persons. From the 5th dynasty (*c.* 2494–*c.* 2345 BC) on they are decorated with scenes from the king's life.

Temples and images. The sun temples are limited to the 5th dynasty for the cult of the sun god, Re. Numerous temples for the gods date from the 18th dynasty to the time of the Roman emperors. The image of the god was kept in the holy of holies ("chamber of the cultic image"), where a daily offering ritual was performed. During festivals it was carried in procession before the public and taken to visit other temples. The ritual scenes decorating the temples of the New Kingdom (1567–*c.* 1085 BC) enable scholars to determine to a large extent the function of the different rooms.

Sculptures of gods in human form are found from prehistoric times on (*e.g.*, Min of Coptos). The different gods are not distinguished by personal features but by attributes or shape—animal, hybrid, or human. Cult implements—portable wooden statues of gods, censers, libation vessels, altars, offering tables—provide clues to religious practices.

Literary sources. *Pyramid Texts and Coffin Texts.* The oldest text collections are the Pyramid Texts inscribed in the subterranean rooms of the pyramids of the kings and the queens from the end of the 5th (*c.* 2345 BC) to the end of the 6th dynasty (*c.* 2181 BC). They contain spells for the offering and funerary rituals, magical texts, and texts pertaining to the king's existence in the hereafter. The Coffin Texts (found on coffins of the First Intermediate Period and the Middle Kingdom—*c.* 2040–1786 BC—and on several papyri) are a collection of spells for private individuals containing copious extracts from the Pyramid Texts, texts from the mortuary cult and the temple ritual, and purely mythological texts.

Book of the Dead. The latest collection is the Book of the Dead: spells written on papyrus and deposited with the dead in the tomb for their use in the hereafter (from the 18th dynasty to the Late Period). Although derived in part from the Coffin Texts, they retain little connection with rituals but are meant to provide the dead with power in the hereafter (which explains why there are so many spells for transformation of the dead into powerful beings). Best known is chapter 125, with the so-called negative confession and the psychostasia (weighing of the soul) in the judgment of the dead.

Psychostasia

Other texts. Further mortuary literature includes offering liturgies (already on a papyrus from the Old Kingdom), texts for the revivification ceremony ("Opening of the Mouth"), and guidebooks to the hereafter (from the 18th dynasty on: *Book of the Two Ways, Amduat,* and others). Ritual texts throwing light on the temple cult and hymns are preserved on temple walls and on papyri from the 18th dynasty on. Rare, however, are theological treatises like the "Memphite Theology," in which the creation of the world, its regulation by Ptah, and the supremacy of Memphis are confirmed theologically. Prayers are handed down in fixed formulas and, beginning in the 18th dynasty, in free form. Mythical subjects mainly are dealt with in ritual texts. Narrated myths form a part of the narrative literature, and the so-called tales are often built upon mythical material.

Later literary sources. The main source in the Old Testament is the story of Joseph in Genesis and Exodus. The Psalms and the apocryphal Wisdom of Solomon show Egyptian influence. Some things have been learned from the correspondence between the pharaohs and the rulers of Syria and Palestine preserved in Tell el-Amarna. The accounts of Greek and Roman historians and travellers are of particular value: Hecataeus of Miletus (6th century BC), Herodotus (5th century BC), Diodorus Siculus (1st century BC), Strabo, the geographer (1st century BC–1st century AD). Egyptian ideas in retouched form are to be found in Plutarch's *De Iside et Osiride* ("Concerning Isis and Osiris") and Apuleius' *Metamorphoses* (*The Golden Ass*). Other important sources are the late lexicographers such as Isidore of Seville (about AD 600) and Suidas (10th century AD).

RELIGIOUS BELIEFS

The gods. Egyptian gods have different forms and personify different powers. There are numerous animal forms: bull gods like Apis at Memphis, Mnevis at Heliopolis, Buchis at Hermonthis; rams like Khnum at Elephantine and the ram of Mendes; falcons like Horus, the god of the king; doglike animals such as the god of the cemeteries, Anubis, and Wepwawet, the "Opener-of-the-Ways"; vulture goddesses like Mut in Karnak and Nekhbet in Elkab. They may also have a human form or a

Table 1: Major Gods and Goddesses

	alternative names	areas of responsibility	principal sanctuary	relations to other deities	symbols and emblems
Amon	Amen, Ammon, Amun	king of the gods; patron deity of the pharaohs; identified with the sun god Re as Amon-Re	originally Khnum, later Thebes	husband of Mut; member of the Theban triad	male figure with a ram's head, or a ram wearing the triple crown
Anubis	Anpu	god of the dead, particularly the aspects of the funeral cult and the care of the dead	Cynopolis	son of Re, or son of Osiris by Nephthys	black jackal, or a male figure with the head of a jackal or dog
Aton	Aten	represented the sun at its zenith; later, for a short time was made the chief and only god (with the exception of the pharaoh)	Akhetaton (Tell el-Amarna)	a form of the sun god Re	a red solar disk with rays terminating in hands
Atum	Tem, Tum	creator of gods, men, and the divine order; founder of the Heliopolitan Ennead; represented the setting sun	Heliopolis	merged with the cult of Re, becoming Atum-Re	old, bearded man wearing the double crown and carrying the *ankh;* sometimes, a male figure with the head of an ichneumon
Bast	Bastet, Ubasti	goddess of music, dance; originally symbolized the fertilizing warmth of the sun; during the New Kingdom thought to be a lioness war goddess	Bubastis	wife of Ptah; member of the Memphite triad; sometimes identified with Sekhmet	cat, or cat-headed woman, holding in her hand either a sistrum or an aegis
Bes	Bisu	originally, protector of the royal house; later, popular god of recreation; also associated with childbirth; the enemy of noxious beasts; protector and bringer of peace to the dead	originated in the Sudan; later adopted throughout Egypt	husband either of Beset, his female counterpart, or of Taurt	dwarf with large head, goggle eyes, protruding tongue, bowlegs, and a bushy tail
Buto	Edjo, Udjo, Wadjet, Wadjit	tutelary goddess of Lower Egypt; defender of the king; personification of the sun's burning heat	Buto	later thought to be the nurse of the infant god Horus	female figure wearing a *uraeus* or the red crown of Lower Egypt; or, a *uraeus* twined around a papyrus stem or as part of the king's diadem
Geb	Keb, Seb	god of the earth; the physical support of the world	Heliopolis	husband and brother of Nut; son of Shu and Tefnut; father of the Osirian gods	a goose or a male figure with his head surmounted by a goose
Hathor	Athyr	originally a personification of the sky; at Dandarah, goddess of festivity and love; protectress of women; at Thebes patroness of the region of the dead; later identified locally with tree, lion, or cow goddesses	originally, Dandarah; later, Ombos, Edfu, and western Thebes	daughter of Nut and Re	cow, cow-headed goddess, or goddess with a human head adorned with horns or cow's ears and heavy tresses framing her face
Horus	Hor	originally the god of Lower Egypt; later identified with the reigning king and thus the opponent of Seth	Nekhen, Edfu, later all of Egypt	son of Osiris	falcon or falcon-headed male figure
Isis	Aset, Eset	queen of the gods; great mother goddess figure; retriever and embalmer of the body of Osiris; protector of the child Horus	originally, Sebennytos; later spread throughout Egypt, the Near East, and the classical world	wife of Osiris; mother of Horus	female figure with both a vulture headdress and a throne or disk flanked by cow's or ram's horns on her head
Ma'at	Mayet	goddess of law, truth, and justice; served as the balance in the scale used to weigh the heart of the deceased	a deified abstract *par excellence;* thus, she had no single cult centre	daughter of Re; wife of Thoth	female figure standing or sitting on her heels, wearing an ostrich feather on her head
Min		personification of the generative force in nature; god of fertility and harvest; protector of desert travellers and god of the road	Coptos, Akhmīm	son of Re, or of Shu	ithyphallic bearded man, his crown surmounted by two tall plumes, and holding a flail or thunderbolt in his right hand
Mut		a vulture goddess of Thebes; a great and mighty divine mother	Thebes	wife of Amon-Re	female figure whose head or headdress was in the form of a vulture or who wears a heavy wig surmounted by the double crown
Nekhbet	Nekhebet	tutelary goddess of Upper Egypt; protectress of childbirth	Elkab (Nekheb)	daughter of Re; wife of Khenti-Amentiu or of Hapi	woman or vulture wearing the white crown of Upper Egypt; often shown spreading wings above king
Nut	Neuth, Nuit	goddess of the sky, arch of the heavens; protectress of the dead	Memphis and the Delta region; also, Dandarah	wife and sister of Geb; mother of the Osirian gods	elongated female figure arched over Shu; often wearing a water-pot or pear-shaped vessel on her head
Osiris	Usire	originally a fertility god and giver of civilization; later, with Re, the supreme god of Egypt and ruler of the dead	Busiris, Abydos	son of Nut and Geb; father of Horus	represented as a dead king in mummy wrappings, the hands holding the crook and flail and wearing the *atef* crown on his head
Ptah	Phthah	god of fertility, creator of the universe, maker of things; the patron of craftsmen and fine arts	originally Memphis, later, all of Egypt	husband of Sekhmet; father of Nefertum; member of the Memphite triad	mummified man with a shaven head his hands holding the *djed* symbol and a sceptre
Re	Phra, Ra	personification of the sun at its zenith; king of the gods and father of mankind; protector of kings; chief state god	Heliopolis	father of Shu and Tefnut, or, alternatively, son of Geb and Nut	falcon-headed man crowned with the sun disk encircled by the *uraeus;* held the *ankh* and a sceptre
Sebek	Sebeq, Sobk, Suchos	originally, perhaps associated with fertility, death, and burial; later, a protector of reptiles and patron of kings; occasionally, a personification of evil	Crocodilopolis; later, Ombos, Thebes, and Lake Moeris	son of Neith; sometimes identified with Seth or with Re (as Sebek-Re)	crocodile or crocodile-headed man wearing on his head either a solar disk and *uraeus* or a pair of horns surmounted by a disk and a pair of plumes
Sekhmet	Sekhet	warlike solar goddess	Memphis and the Delta region	wife of Ptah; mother of Nefertum; member of the Memphite triad	lioness or woman with a lion's head, on which she wore the solar disk and *uraeus*
Seth	Set, Setekh, Setesh	partner and rival of Horus in the Osirian myths	Ombos	brother or son of Osiris	the "typhonian animal": a composite creature with a grey hound's body, slanting eyes, square-tipped ears, and long forked tail
Shu		god of light and air and supporter of the sky; later, personification of the divine intelligence	Dandarah, Edfu, Apollinopolis Magna, Memphis	son of Re; husband and brother of Tefnut	male figure with an ostrich feather on his head
Thoth	Djhowtey	moon god; later the god of reckoning and of learning in general; inventor of writing, scribe of the gods; lord of magic	Hermopolis Magna	protector of Isis during her pregnancy; healer of her son Horus' injury; husband of Ma'at	male figure with an ibis' head on which were placed the moon crescent and disk

hybrid form with an animal head. Powers of nature like Re, the sun; Nut, the sky; Geb, the earth; and Thoth, the moon (and therefore "the reckoner of time"), are mainly rendered in human form but can also assume animal form (*e.g.*, Thoth as ibis or baboon). Ptah of Memphis, the creator-god; Min of Coptos, the god of fertility; Khons of Thebes as moon god; and Osiris as chthonian (underworld) fertility god are mostly human-shaped. Amon-Re of Karnak, usually in human form, can also have the form of a ram or a Nile goose and is looked upon as "breath of life" too. In theological theories each god is attributed as many powers as possible, hence the wide variety of divine manifestations.

Ma'at, the central conception of order, can be called a goddess insofar as she is represented as such, has a temple in Karnak, and is called "daughter of Re." But actually she is an abstract idea apart from the world of the gods.

Cosmogonies

Myths. Cosmogonies and other myths illustrate the order of the world and the cooperation of the gods. One cosmogony presents the image of a primeval hill emerging from the primeval flood (Nu) and bearing the first life. The Heliopolitan cosmogony says that the primeval god, Atum, spat out or coughed up the first couple of gods, Shu and Tefnut (air and moisture), or created them by self-generation. Shu separated heaven (Nut) from earth (Geb). These five gods are thought of as three generations. Osiris, the chthonian fertility god, is joined by Isis, his wife, and Seth, his murderer, who is married to Nephthys, the "Hostess of the House." Horus, the king's god and the son and heir of Osiris, rules here on earth. The myth of Osiris, the dead god, who begets a son out of the underworld, is connected with the cult of the first kings at Abydos. There, the so-called Osiris mysteries, annually re-enacting the mysterious event, developed. To the Egyptian they mean the central point in his hope for life; to the Hellenists they symbolize the mysterious teachings of ancient Egypt. The connection of the gods of the Heliopolitan cosmogony with the gods of the Osiris family as "Children of Nut" makes up a family tree, reaching from the creator-god, Atum, to the ruling king, the "Living Horus."

Human nature and destiny. Man possesses spiritual power extending in part beyond his personal life.

Concepts of soul. First of all there is the *ka*, which, created with the person, can be defined as "vital force" and outlasts man. The bird-shaped *ba*-soul seems to be set free only after death. Gods too have *ba*- and *ka*-souls. The bird-shaped *akh*-soul is also preserved after death and can appear as a ghost or revenant. The heart is considered as the seat of the power of discrimination, and therefore the word "heart" also has the meaning "conscience." Image and name, which mainly play a part in magic, are to a certain extent regarded as independent.

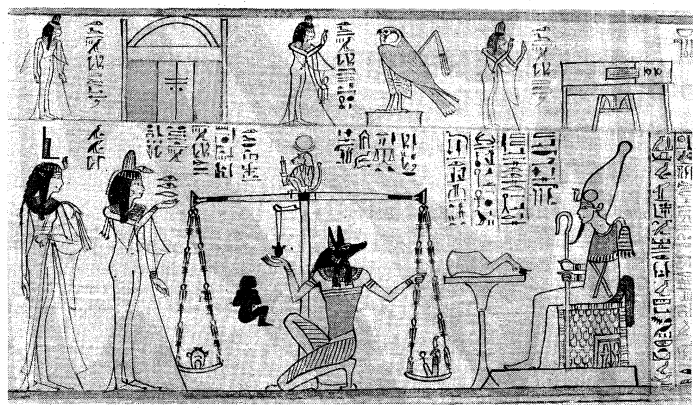
Ethics. Ethics is part of social life. Its rules, preserved in biographies and teachings, do not refer to an absolute good or evil but are meant to guarantee a harmonious coexistence free from the use of force. They differ as to the social rank of the individual. They can be taught and aim to convince by common sense. It is only in the New Kingdom that ethics is directed at man's reflective nature and thus comes into intimate relationship with religion. The first conceptions of this kind can be found in the First Intermediate Period, but it is only later that they become popular belief. Right actions spring from a god-fearing mind, and God inspires man's mind in the right way.

FORMS OF EGYPTIAN RELIGION

Funerary religion. Many rites deal with the provisions for life after death. The dead man's body is preserved by mummification; numerous amulets guarantee its invulnerability. Offerings at the burial and on funerary festivals in the tombs provide him with food and drink. The mortuary equipment, with all its useful and luxurious articles, is at his disposal. All serve the subsistence of the dead, who live in the tombs or return to them. Imagination created various conceptions of the hereafter. In one of them it is situated in heaven. The soul ascends or flies to heaven, lives among the stars or the entourage of the sun, and finds food and all its needs in the fields of the blessed. According to other conceptions, the hereafter is situated in the underworld. The soul finds its way there through passages and doors guarded by doorkeepers, trying to ward it off. It also has to cross watercourses. In books of the hereafter, the underworld is divided into 12 regions, visited by the sun on its nightly course, bringing life for a short period. The idea of a judgment of the dead developed from the hope that the dead person could find a tribunal in the hereafter, where he could sue for his claims. In the early period the dead king or Re or Osiris is the supreme judge. From the First Intermediate Period on, the conception changes; the dead person has to prove his right to enter the hereafter. It is not until the 18th dynasty that the classical picture of the judgment of the dead is formed, with Osiris and 42 judges, before

Conceptions of the afterlife

By courtesy of the Metropolitan Museum of Art, New York, Museum Excavations, 1928–29, and Rogers Fund, 1930



The judgment of the dead. While Osiris, the ruler of the dead, watches from his throne, the jackal-headed god Anubis weighs the heart of the deceased against the figure of Ma'at. Funerary papyrus from Thebes, *c.* 1025 BC.

whom the heart is weighed against *ma'at* and the deceased recites all the uncommitted transgressions listed in chapter 125 of the Book of the Dead. In spite of all the darkness and dangers, the hereafter is still part of life. It guards against the Egyptians' greatest dread: complete extermination, "repeated death."

Divine kingship: pharaoh as god

The state religion. Pharaoh, the incarnation of the god Horus here on earth, heir to the kingship of Atum, son of the sun god Re, is the centre of the world. The world of the gods is determined by his existence insofar as certain gods (Re, Atum, Ptah, etc.), following his example, acquire the rank of "King of the Gods," are arranged in dynasties, and are placed in traditional histories as god-kings before the earthly kings. The existence of a legitimate pharaoh guarantees the order of the cosmos. This is expressed above all in the king's designation as "Lord of the Ritual." In theory he is the only priest who can perform the daily temple ritual and the festival rituals in accordance with *ma'at*, thus preserving the order of the cosmos, the fertility of the land, and the order and welfare of the empire.

The temple cult. The temple was "god's house," in which the god lived in the form of his cult statue. The daily ritual included the opening of the shrine, the decorating and dressing of the statue, and the offering of food and drink following the daily routine of the king.

Priesthood. The ritual was performed by official priests as representatives of the king. Knowledge of the ritual and maintenance of the rules of purity were the essential requisites for priestly service; a differentiation between priests and laymen seems to have occurred late. The priesthood itself consisted of numerous classes. Large temples had four main priests ("prophets"), numerous minor priests, singers, musicians, scribes, and temple employees. As each temple owned land, villages, workshops, and sometimes even quarries and foreign landholdings, it was virtually an independent economic

unit, with personnel corresponding to its size. During the temple service, hymns and songs describing his nature and activity accompanied the appearance of the god. Formulas naming the objects (incense, food offerings, jewelry, clothes) and identifying both the ritual as a whole and the different objects with mythical elements were part of the ritual. Thus the ritual acts were elevated to a mythical level and drawn into manifold connections with mythical ideas.

Festivals. The numerous festivals were partly of local, partly of general importance. They were held on birthdays of gods, on special calendar days (New Year, first days of the months), at seasonal events (harvest festival), king's ceremonies (coronation, so-called jubilees), and victory celebrations. Important festivals lasted several days. Most of the temple festivals were connected with a procession of the god's image and often with a visit to other temples. On these occasions the god was visible to the public and gave oracles.

Religious, social, and economic roles. Since Egyptian religion was centred on temple ritual, individual faith (*i.e.*, personal piety) was never of primary importance and was therefore perceptible only in later periods. On the other hand, religion played a decisive part in social and economic life. A large part of the population was integrated into the official hierarchy of priests and temple

Table 2: Minor Gods and Goddesses

	alternative names	areas of responsibility	principal sanctuary	relations to other deities	symbols and emblems
Anhur	Onouris	symbolized the creative power of the sun; frequently invoked against enemies and noxious animals	Sebennytus, This	probably a war-like personification of Re	warrior wearing a long embroidered robe and a headdress with four tall plumes; often carries a lance
Anuket	Anqet, Anquet	personification of the Nile as the nourisher of the fields	island of Sahal near Elephantine	second wife of Khnum; sister of Sati; member of the triad of Elephantine	woman wearing a crown of feathers and carrying a sceptre and the *ankh*
Apepi	Apep, Apophis, Rerek	personification of darkness and evil; attempted to obstruct the passage of the solar barque	as the embodiment of evil, Apepi had no worshippers	eternal enemy of the sun god Re	huge serpent
Hapi	Hap, Hep	god of the Nile, particularly of the inundation; bringer of fertility; nourisher of the gods and man	region of the First Cataract	husband of Nekhbet	bearded man coloured green or blue, with female breasts and a clump of water plants on his head
Khenty-Imentiu	Khenti-amentiu, Ophis (or Ophois)	warrior god; god of the dead until deposed by Osiris; at times piloted the sun's boat during its nocturnal voyage	Abydos, Asyut	associated with Osiris during that god's conquest of the world, later identified with Upuaut (Wepwawet)	male figure with the head of a wolf or a jackal
Kheper	Khepera, Khepri	symbolized the daily cycle of the sun across the sky; sometimes thought to be a symbol of the rising sun	Heliopolis	one of the forms of the sun god Re	a scarab pushing the sun across the sky or a man with a scarab on his head
Khnum	Khnemu	god of the cataract region; earlier associated with the abyss of the netherworld; in some places thought to be the potter god who formed the world and its inhabitants	Elephantine	husband of Sati and Anuket; member of the triad of Elephantine	ram-headed male figure
Khons	Khensu, Khonsu, Chons	moon god; during the New Kingdom, an exoroist and healer	Thebes; later all of Egypt	adopted son of Amon and Mut	swathed figure with its head shaven except for the braid of royal progeny and the skull cap holding the lunar disk and crescent
Mont	Mentu, Month	war god of Upper Egypt; during the Middle Kingdom, lord of the sun and sky with Re	Hermonthis	a form of the sun god Re	male figure with the head of a hawk surmounted by the solar disk and two plumes; often carried a curved scimitar
Nefertum	Nefertem, Nefertemu	god of the lotus, thus having a life-giving aspect	Memphis	son of Ptah and Sekhmet (or Bast), third member of the Memphite triad	lion-headed man wearing a headdress composed of a lotus flower, two plumes and two menats
Neith	Neit	goddess of war and the loom	Sais	self-begotten virgin; according to some legends, the mother of Re	carried arrows crossed on an animal skin; on her head, wore the Red Crown of Lower Egypt or a weaver's shuttle
Nephthys	Nebhat	associate of Isis in the Osiris cycle	Heliopolis	mother of Anubis, wife of Seth; sister of Isis, Osiris, Seth, and Horus	female figure wearing horns and the solar disk on her head
Nu	Nun	personification of the primordial watery abyss from which the world was created; and, therefore, the oldest and wisest of all gods	Hermopolis	guarded Shu and Tefnut at birth	male or female figure, sometimes with the head of a frog; occasionally shown plunged up to his waist in water
Sati	Satet, Satis	goddess of the inundation and of fertility	Island of Sahal near Elephantine	principal wife of Khnum; sister of Anuket; member of the triad of Elephantine	female figure wearing the crown of the South and a pair of horns, and holding a sceptre and the *ankh* (or a bow and arrows)
Seker	Sokar, Sokaris	deity of darkness and decay in the earth	Memphis	identified with Ptah and Osiris as Ptah-Seker-Osiris	hawk or hawk-headed mummy
Seshat	Sesheta	goddess of writing and history; measurer of time; foundress of temples	Hermopolis Magna	principal wife of Thoth	female figure wearing a leopard skin and holding a writing reed and a scribe's palette; on her head, a crescent (or two turned-down horns) and two plumes
Taurt	Apet, Opet, Taweret, Thoueris	goddess of maternity, presided over childbirth; during the New Kingdom, a protective household deity	Thebes	often appeared with Bes	pregnant hippopotamus standing on hind legs and holding two plaits of rolled papyrus
Tefnut	Tefnet	personification of the life-giving dew and moisture; received, with Shu, the sun as it broke free from the eastern horizon	Heliopolis	sister and wife of Shu	lioness or a woman with the head of a lion wearing the solar disk and the *uraeus* (occasionally, represented as the *uraeus* itself)

personnel. As knowledge of the holy scriptures and the ability to interpret and reformulate them was required, the higher priesthood achieved a high standard of education. Since the temples were economic units, which at times owned large estates, they constituted a major factor in the national economy. The lands administered by them were in many cases more favourably taxed than private property.

Magic, divination, and popular religion. Magical power—through knowledge of the name, possession of the image, application of the "right" word, and the use of amulets—was a weapon in man's fight against supernatural powers. The boundaries between magic and ritual were fluid. The results desired from ritual actions (fertility, order in nature, etc.) can be ascribed to both magical power and divine activity. Oracles certainly date from early periods, but their importance increased from the New Kingdom on. Political decisions (*e.g.*, the election of the king) and personal affairs (*e.g.*, the beginning of a journey, the choice of a profession) as well as judicial investigations were brought before the god in the form of alternative questions and were apparently decided by movements of the image.

Oracles

Since the public was generally excluded from the temple cult (apart from festival processions), they turned to divine beings and objects outside the temples (*e.g.*, the "Western Peak" at Thebes) or to certain images of gods and kings in the outer districts of the temples (*e.g.*, "Ptah in the Door-Frame" at Madīnat Habu). Statues and tombs of famous deceased were also adored. Some of these "popular gods" were recognized by Egyptian theology and integrated into theological systems (*e.g.*, Imhotep, the architect of King Djoser, is first patron of scribes, later a god and the son of Ptah). The religious revival following the heresy of Amenhotep IV (Akhenaton) originated apart from the official theology and influenced religion for centuries.

RELIGIOUS SYMBOLISM AND ICONOGRAPHY

In accordance with the faculty for abstraction and the limited sense of individuality typical of Egyptian culture as a whole, the representation of religious conceptions was characterized by a series of religious symbols and emblems rather than by the creation of individual god figures. Of all the images of gods, that of Isis with the child Horus on her lap, perhaps the prototype of the Madonna with the Child, is the most noteworthy. The image of the youthful sun god on a lotus blossom also exerted an influence outside Egypt. Among the symbols that were used as ornaments were the *crux ansata* (the symbol of life), the divine eye as symbol of prosperity, the scarab beetle standing for eternal regeneration, the knot of Isis as protective symbol, the crook as sign of kingship, the winged sun as symbol of the sun god, the *djed*-pillar as symbol of endurance, and the *was*-sceptre (a shepherd's staff) and *nefer*-sign as symbols of both fortune and welfare. Significant among architectural forms are pyramids, obelisks, and cavetto cornices, the last, however, having no symbolic value (see also EGYPT, HISTORY OF).

HISTORICAL DEVELOPMENT

A large number of the gods go back to prehistoric times. The images of a cow and star goddess (Hathor), the falcon (Horus), and human-shaped figures of the fertility god (Min) can be traced back to that period. Some rites, such as the "running of the Apis-bull," the "hoeing of the ground," and other fertility and hunting rites (*e.g.*, the hippopotamus hunt), presumably date from early times. To explain the existence of a variety of gods at different places, the following theory is often referred to: Egypt consisted of numerous nomes, or provinces, each having its own gods. When they were united at the beginning of the 3rd millennium BC, they retained their local gods, thus creating the pantheon. This theory cannot be proved. Connections with the religions in southwest Asia cannot be traced with certainty. It is doubtful whether Osiris can be regarded as equal to Tammuz or Adonis or whether Hathor is related to the "Great Mother." There are closer relations with northeast African religions. The numerous animal cults (especially bovine cults and panther gods) and details of the ritual dresses (animal tails, masks, grass aprons, etc.) probably are of African origin. The kingship in particular shows some African elements, such as the king as the head ritualist (*i.e.*, medicine man), the limitation and renewal of the reign (jubilees, regicide), and the position of the king's mother (a matriarchal element). Some of them can be found among the Ethiopians in Napata and Meroe, others among Prenilotic tribes (Shilluk).

The pantheon as synthesis of earlier local gods

The religion of independent Egypt. The history of Egyptian religion is less one of change in cults and gods than one of change in ideas and beliefs within the existing mythological and ritual sphere. In the Old Kingdom the state religion in particular was developed: the king as Horus and son of Re at the head of an ordered and well-organized world. Evidence concerning personal beliefs is almost wholly lacking. The collapse of the Old Kingdom brought with it the dissolution of the old social and intellectual bonds. At this time questions concerning the origin of evil, the end of the world, the responsibility of man, and life in the hereafter were passionately discussed, with the traditional religious forms and figures receding into the background. The concept of the gods became personal and undogmatic.

Although the Middle Kingdom returned to the old forms, the following period showed that the ideas of the First Intermediate Period had not been lost. At the beginning of the New Kingdom, the Egyptian sphere expanded enormously. The boundaries of the empire reached from the Fourth Cataract of the Nile to the upper Euphrates, and the acquaintance with Syrian and Palestinian cultures and religions enriched Egyptian thinking. Foreign gods (Resheph, Astarte, Baal) were worshipped in Egypt. This expansion took place under the supremacy of the god Amon-Re at Karnak. Tensions with his priesthood led to the heresy of Amenhotep IV. Opposing the claim to authority of Amon-Re of Karnak and his priesthood and intending to give an all-uniting god to the empire, Amenhotep IV (Akhenaton) made the sun disk, Aton, the supreme god and proclaimed himself his only interpreter. The attempt to establish a new religion followed the old sun belief and was accompanied by the attempt to create a new style of life and art. His attempt failed because of the god's claim to exclusiveness, his struggle against the existing economic order, his political failure, and, last but not least, his underrating of the belief in Osiris. His death was followed by a restoration period, which tried to extinguish all traces of his innovations. The breakdown of the movement was followed by an apparent restoration of the traditional cults. But religious thought was strongly influenced by the rise of a feeling of guilt and sin that brought the individual into a new ethical relationship to the gods. Individual confessions and prayers and moral justifications of the teachings were symptomatic of the new religious climate.

Amenhotep IV (Akhenaton): the Amarna heresy

In the Late Period various tendencies of religious development can be distinguished. The ethicizing and personalizing of religion continued. The importance of Amon-Re, who was closely linked to the policy of the New Kingdom, decreased, while that of Osiris and his so-called mysteries increased in response to the religious demands of the time. In addition, popular beliefs and their cults came increasingly to the fore: animal cults and oracles grew in number, local gods in importance. Cult prescriptions and purity laws apparently separated priests from laity to a greater extent than before. A discrepancy between cult forms and beliefs can be said to have begun.

Egyptian religion in the Greco-Roman world. When Egypt came under the rule of the Macedonian Ptolemies and later of the Roman emperors, the ancient cults continued to exist in their external forms. According to representations in the temples, the Ptolemaic kings and Roman emperors took the role of the pharaoh. Yet differences are still apparent. The layout of the temples (Dan-

darah, Edfu, Ombos, etc.) was standardized. All sorts of religious texts were written on temple walls, often in purposely mysterious spelling. A special cult for the Ptolemaic kings and queens was introduced in the temples, and some Hellenistic ideas seem to have been adopted and combined with native ideas by the Egyptian priests. The influence of the Egyptian on the Hellenistic mind was undoubtedly stronger than the reverse influence. In the Hellenic world Egyptian gods and cults were widely known through the histories of Hecataeus and Herodotus and the accounts of Greek merchants and mercenaries. Herodotus thought the Egyptian gods to be identical with the Greek ones (Zeus = Amon, Pan = Min, Athena = Neith, etc.).

For political reasons Ptolemy I Soter wanted to create a cult common to Egyptians and Macedonians. The deified Memphite Apis-bull, Osiris-Apis, became the prototype of the god Sarapis and acquired a Greek cultic image alleged to be from Sinope on the Black Sea. Its cultic centre was the new capital, Alexandria, whence the cult migrated to Europe more than to Egypt. The Hellenistic world, which favoured mystery religions with special zeal, adopted the cults of Osiris, Isis, and Horus (Harpocrates), transforming them in the manner of the mysteries of the late classical world. They were spread over the entire Roman Empire. The hymns to Osiris or Harpocrates preserved in Greek are translations of Egyptian texts. The conception of Isis as the goddess of navigation was formed in Alexandria. Not only Sarapis but also Osiris, Isis, and Horus adopted new forms in Hellenistic and Roman art; similarly Anubis was seen as a Roman warrior and Apis as an emperor. Particularly significant as the initiator of theosophy—a system of beliefs concerning God and the world based on mystical insights—is Hermes Trismegistos (in Egyptian, "Thoth the Very Great"), who is said to be the author of the Hermetic Scripts. As the so-called Thrice Great he was acknowledged by Christianity to some degree, because he was supposed to have expounded the Trinity. Egyptian theories often found their way, in many cases newly interpreted, into the philosophy of the late classical world (Plotinus, Neoplatonism) and Gnosticism—a Hellenistic religious movement that viewed spirit as good and matter as evil—(see also HELLENISTIC RELIGIONS; MYSTERY RELIGIONS; PLATONISM AND NEOPLATONISM).

Mystery religions

Egyptian conversion to Christianity and the coming of Islām. Egypt was Christianized very early, probably in the time of the Apostles. St. Mark, the author of the First Gospel, is regarded as the founder of Egyptian Christianity. It is certainly no accident that the Egyptian Christian Church (the Coptic Church) adhered to the Monophysite theory that Christ had one nature (a view held heretical by the Council of Chalcedon in AD 451). This is surely connected with the Egyptian understanding of deity, which made a distinction between a divine and human person unnecessary. Besides that, however, hardly anything of ancient Egyptian religion can be found in Egyptian Christianity. Yet even after the Islāmic conquest (AD 641), many individual features have continued to survive in popular belief and may still be found in the veneration of saints, and in popular customs and superstitions.

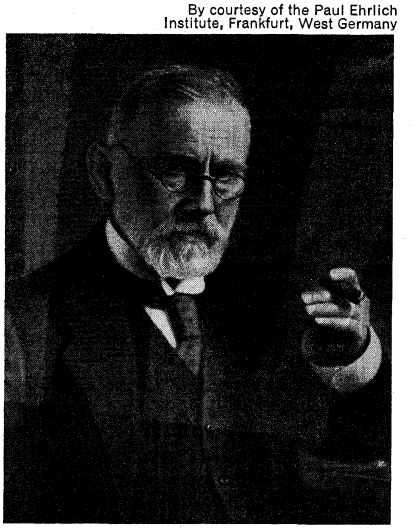
BIBLIOGRAPHY. H.B.C. BONNET, *Reallexikon der ägyptischen Religionsgeschichte* (1952), a dictionary of the names of gods, temples, and holy places, and religious terms of the pharaonic period and of the Ptolemaic and Roman times; J.H. BREASTED, *Development of Religion and Thought in Ancient Egypt* (1912) and *The Dawn of Conscience* (1934), two books that attempt to represent the historical development of the religious ideas and ritual customs; HENRI FRANKFORT, *Ancient Egyptian Religion: An Interpretation* (1961), a discussion of the special way of Egyptian understanding and thinking as a starting point of religious forms; HERMANN KEES, *Der Götterglaube im Alten Aegypten*, 2nd ed. (1956), a precise analysis of local gods, their forms and history, and the development of local cults; *Totenglauben und Jenseitsvorstellungen der alten Ägypter* (1956), a detailed study of the beliefs of the hereafter, the existence of the dead in the sky and in the netherworld, and the judgment of the dead; and *Das Priestertum im ägyptischen Staat* (1953), a discussion of the role and duties of the priesthood; K.H. SETHE, *Die altaegyptischen Pyramidentexte*, 4 vol. (1908–22), ritual texts written in the pyramids of the kings and queens of the 5th and 6th dynasties; a new translation and shorter commentary with supplements to these texts trans. by R.O. FAULKNER, *The Ancient Egyptian Pyramid Texts* (1969); ADRIAAN DE BUCK and A.H. GARDINER (eds.), *The Egyptian Coffin Texts*, 7 vol. (1935–61), texts written on wooden coffins (and a few papyri) in the time of the First Intermediate Period and the Middle Kingdom, especially important for the beliefs of this time; J.B. PRITCHARD (ed.), *Ancient Near Eastern Texts Relating to the Old Testament*, 2nd ed. (1955), translations of religious, literary, and historical texts from Egypt, Sumer, Mesopotamia, and Ugarit; THEODOR HOPFNER, *Fontes Historiae Religionis Aegyptiacae* (1922–25), excerpts from Greek and Latin authors relating to Egyptian religion from the time of Homer to the Middle Ages; R.T.R. CLARK, *Myth and Symbol in Ancient Egypt* (1960), an intelligible introduction to the mythological world with full interpretation of the great myths and explanations of the nature of the principal gods, their cult places, and major religious symbols; JOHN GARSTANG, *The Burial Customs of Ancient Egypt* (1907), an extensive description of the tombs of Beni Hassan and their equipment; ALEXANDRE MORET, *Du Caractère religieux de la royauté Pharaonique* (1902), a detailed description, though somewhat antiquated, of the dogma of the sacred kingship and its ceremonies at the birth, the coronation, at feasts, and at the death of the pharaohs; GEORGES POSENER, *De la Divinité du Pharaon* (1960), descriptions of the human character of the historical pharaohs and their image in literature; SERGE SAUNERON, *Les Prêtres de l'ancienne Égypte* (1957; Eng. trans., *Priests of Ancient Egypt*, 1960), the essential features of the Egyptian religion as shown by the description of the priesthood; FRANCOIS LEXA, *La Magie dans l'Égypte antique* (1925), a collection of magical texts and charms, amulets, sacred signs, and symbols of all periods; ERIK IVERSEN, *The Myth of Egypt and Its Hieroglyphs in European Tradition* (1961), on the survival of the Egyptian culture, especially the religion, in Mediterranean countries.

(E.O.)

Ehrlich, Paul

A German medical scientist of great intuitive capacity and rare versatility, Paul Ehrlich, who helped to develop the medical specialties of hematology and immunology, is best remembered for his pioneering work in chemotherapy, especially as the discoverer of the first cure for syphilis. An advocate of "chemical thinking" in medicine and biology, he was among the first to promote teamwork between research and industry. He shared the Nobel Prize for Physiology and Medicine in 1908 for his work on immunity.

By courtesy of the Paul Ehrlich Institute, Frankfurt, West Germany



Ehrlich.

Ehrlich was born on March 14, 1854, at Strehlen, in Prussian Silesia (now Strzelin, Poland), of a Jewish family prominent in business and industry. Although he lacked all formal training in experimental chemistry and applied bacteriology, he did independent work even as a

medical student at the University of Leipzig, writing about the distribution of foreign substances in the body and proving, with the aid of early discoveries in colour analysis, that the chemical mechanism that operates in therapeutic and toxic processes also determines cell nutrition. As a result of this work, Ehrlich was invited to do research at the famous Charité Hospital in Berlin, where he invented a new staining technique for the tuberculosis bacillus discovered by the German bacteriologist Robert Koch. This was an advance that proved to be of decisive importance for the microscopic diagnosis of tuberculosis. While developing new methods for the dyeing of live tissue, he discovered the uses of methylene blue in the treatment of nervous disorders. In other diagnostic advances, he traced a specific chemical reaction in the urine of typhoid patients, tested various medications for reducing or removing fever, and made valuable suggestions for the treatment of eye diseases. Of the 37 scientific contributions that he published between 1879 and 1883, Ehrlich considered the last as the most important: "The Requirement of the Organism for Oxygen." In it he established that oxygen consumption varies with different types of tissue and that these variations constitute a measure of the intensity of vital cell processes. In 1883 he married Hedwig Pinkus, with whom he had two daughters.

A bout with tuberculosis forced Ehrlich to interrupt his work and to seek a cure in Egypt. When he returned to Berlin in 1889, the disease had been permanently arrested. After working for some time in a tiny and primitive private laboratory, he transferred to Koch's "Institute for Infectious Diseases," where he concentrated on the problem of immunity. Very little was known at this time about the precise manner in which bacteria bring about disease, and even less was known about the body's defenses and about therapeutic effects. The working hypothesis that Ehrlich developed and by which he proceeded was the side-chain theory, according to which each cell has a vital centre of protein substance and a series of side chains, or receptors, which absorb and assimilate nutrients and certain toxic substances as well. Only if the haptophore group of a toxic molecule combines with the side chain of the cell can a bacterial toxin act upon a cell. The affected organism then produces great quantities of side chains, all of them "gauged" to the disease-producing toxic agent. These immune bodies prevent a renewed infection, so that the organism is actively immunized (see IMMUNITY).

Side-chain theory

This much-debated hypothesis in all its ramifications was destined to have a profound influence on Ehrlich's later work and on that of his successors. Thus Ehrlich was able to show experimentally that rabbits subjected to a slow and measured increase of toxic matter were able to survive 5,000 times the fatal dose. In the end, he established the precise quantitative patterns of immunity. These findings assumed great importance in 1890, when he met Emil von Behring, who had succeeded in creating an antitoxin against diphtheria. Behring had tried to prepare a serum that could be used in clinical practice, but it was only by adopting Ehrlich's technique of using the blood of live horses that the preparation of a serum of optimum antitoxic effectiveness became possible. Ehrlich developed a way of measuring the effectiveness of sera that was soon adopted all over the world for the standardization of diphtheria serum.

On the basis of these achievements Ehrlich was made director of a government-supported institute in Berlin, which was transferred to Frankfurt on Main in 1899 as the "Royal Institute for Experimental Therapy." No restrictions of any kind were placed upon the direction of his research. While this corresponded to Ehrlich's own talents and inclinations, it did not please Behring, who endeavoured to have his colleague specialize in immunology and serum therapy. The strained relationship between the two men was exacerbated by personality differences. Ehrlich, utterly indifferent to monetary rewards, had no ambition to become an industrialist like Behring; he was content to do his research in study and laboratory.

He had by now recognized the limitations of serum therapy. Many infectious disorders, in particular those caused by protozoa rather than bacteria, failed to respond to serum treatment. The recognition of this fact marks the birth of chemotherapy. Ehrlich started experimenting with the identification and synthesizing of substances, not necessarily to be found in nature, that could kill parasites or inhibit their growth without damaging the organism. He began with trypanosomes, a species of protozoa that he unsuccessfully attempted to control by means of coal-tar dyes. There followed compounds of arsenic and benzene; other compounds proved to be too toxic. Instead of declaring himself vanquished by these difficulties, Ehrlich now turned his attention to the *Spirochaeta pallida*, the causal organism of syphilis.

He had at this time several institutes at his disposal as well as sizeable research funds. He also had a staff of highly competent collaborators; in fact, his colleague, Sahachiro Hata, contributed much to Ehrlich's eventual success in combating syphilis. His preparation 606, later called Salvarsan and still later Neosalvarsan, was extraordinarily effective and harmless despite a large arsenic content. Already the first tests, undertaken in the spring of 1910, proved to be surprisingly successful in the treatment of a whole spectrum of diseases; in the case of yaws, a tropical disease akin to syphilis, a single injection was sufficient. It seemed, indeed, as if a "magic bullet," to use a favourite expression of Ehrlich's, had been found.

Cure for syphilis

Given the devastation wrought by syphilis, a worldwide demand soon arose for a new weapon against the disease. Ehrlich, however, would not yet release his discovery for general use, believing as he did that the usual few hundred clinical tests did not suffice in the case of an arsenic preparation, the injection of which required special precautions. In an unheard-of transaction, the manufacturer with whom Ehrlich had collaborated closely, Farbwerke-Hoechst, released a total of 65,000 units gratis to physicians all over the globe. Although harmful side effects remained nominal in number, some envious competitors did not hesitate to attack Ehrlich. Eventually he was forced to sue the most slanderous among them, and this detractor was given a jail sentence. The greatest distinction bestowed by the Prussian State was the title, "Wirklicher Geheimer Rat," or Privy Councillor, with the predicate of "Exzellens." Along with many other honors, Ehrlich was given honorary doctorates by the Universities of Oxford, Chicago, and Athens, and an honorary citizenship by Frankfurt am Main, where the institute he founded still bears his name.

Having suffered a first stroke in December 1914, Ehrlich succumbed to a second stroke on August 20 of the following year. In its obituary the London *Times* acknowledged Ehrlich's achievement in opening new doors into the unknown, saying, "the whole world is in his debt."

BIBLIOGRAPHY. F. HIMMELWEIT edited Ehrlich's writings in 1956, with a bibliography and four volumes of text in German, English, and French, but this publication does not cover all the posthumous work. There is as yet no comprehensive biography of Ehrlich. See the article by C.H. BROWNING, "Emil Behring and Paul Ehrlich: Their Contributions to Science," *Nature*, 175:570–575, 616–619 (1955); HEINRICH SATTER, *Paul Ehrlich, Begründer der Chemotherapie* (1962); H. LOEWE, *Paul Ehrlich* (1950); and W. GREILING, *Im Banne der Medizin-Paul Ehrlich* (1954).

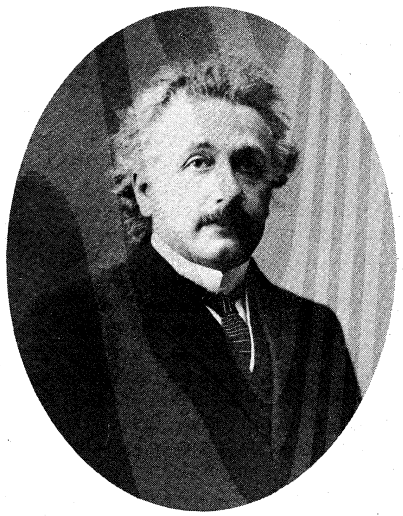
(H.Sa.)

Einstein, Albert

Recognized in his own time as one of the most creative intellects in human history, Albert Einstein, in the first 15 years of the 20th century, advanced a series of theories that for the first time asserted the equivalence of mass and energy and gave entirely new concepts of space, time, and gravitation. His theories of relativity were a profound advance over the old Newtonian physics and revolutionized scientific and philosophic inquiry.

Herein lay the unique drama of Einstein's life: a self-confessed lone traveller, his mind and heart soared with the cosmos, yet he could not armour himself against the

intrusion of the often horrendous events of the human community. Almost reluctantly he admitted that he had a "passionate sense of social justice and social responsibility." His celebrity gave him an influential voice that he used to champion such causes as pacifism, liberalism, and Zionism. The irony for this idealistic man was that his famous postulation of an energy–mass equation,



Einstein.
By courtesy of the Nobelstiftelsen, Stockholm

which states that a particle of matter can be converted into an enormous quantity of energy, had its spectacular proof in the creation of the atomic and hydrogen bombs, the most destructive weapons ever known.

Early life and career. Albert Einstein was born in Ulm, Germany, on March 14, 1879. The following year his family moved to Munich, where Hermann Einstein, his father, and Jakob Einstein, his uncle, set up a small electrical plant and engineering works. In Munich Einstein attended rigidly disciplined schools. Under the harsh regimentation of 19th-century German education, which he found intimidating, he showed little scholastic ability. At the behest of his mother, Einstein also studied music, and, though he played exclusively for relaxation, he became an accomplished violinist. It was then only the influence of Uncle Jakob that stimulated in Einstein a fascination for mathematics and the influence of Uncle Cäsar Koch that stimulated a consuming curiosity about science.

By the age of 12 Einstein had decided to devote himself to solving the riddle of the "huge world." Three years later, with poor grades in history, geography, and languages, he left school with no diploma and went to Milan to rejoin his family, who had recently moved because of his father's business setbacks. Albert Einstein resumed his education in Switzerland, culminating in four years of physics at the renowned Polytechnic Academy in Zürich.

After his graduation in the spring of 1900, he became a Swiss citizen, worked for two months as a mathematics teacher, and then was employed as examiner at the Swiss patent office in Bern. With his newfound security, Einstein married his university sweetheart, Mileva Marić, in 1903.

First contributions to science

Early in 1905 Einstein published in the prestigious German physics monthly *Annalen der Physik* a thesis, "A New Determination of Molecular Dimensions," that won him a Ph.D. from the University of Zürich. Four more important papers appeared in *Annalen* that year and forever changed man's view of the universe.

The first of these, "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen" ("On the Motion of Small Particles Suspended in a Stationary Liquid According to the Molecular Kinetic Theory of Induction") provided a theoretical explanation of Brownian motion. In "Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt" ("On a Heuristic Viewpoint Concerning the Production and Transformation of Light"), Einstein postulated that light is composed of individual quanta (later called photons) that, in addition to wavelike behaviour, demonstrate certain properties unique to particles. In a single stroke he thus revolutionized the theory of light and provided an explanation for, among other phenomena, the emission of electrons from some solids when struck by light, called the photoelectric effect.

Einstein's special theory of relativity, first printed in "Zur Elektrodynamik bewegter Körper" ("On the Electrodynamics of Moving Bodies") had its beginnings in an essay Einstein wrote at age 16. The precise influence of work by other physicists on Einstein's special theory is still controversial. The theory held that if, for all frames of reference, the speed of light is constant and if all natural laws are the same, then both time and motion are found to be relative to the observer.

In the mathematical progression of the theory, Einstein published his fourth paper, "Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?" ("Does the Inertia of a Body Depend Upon Its Energy Content?"). This mathematical footnote to the special theory of relativity established the equivalence of mass and energy, according to which the energy E of a quantity of matter, with mass m, is equal to the product of the mass and the square of the velocity of light, c. This relationship is commonly expressed in the form $E = mc^2$.

Public understanding of this new theory and acclaim for its creator were still many years off, but Einstein had won a place among Europe's most eminent physicists, who increasingly sought his counsel, as he did theirs. While Einstein continued to develop his theory, attempting now to encompass with it the phenomenon of gravitation, he left the patent office and returned to teaching —first in Switzerland, briefly at the German University in Prague, where he was awarded a full professorship, and then, in the winter of 1912, back at the Polytechnic in Zürich. He was later remembered from this time as a very happy man, content in his marriage and delighted with his two young sons, Hans Albert and Edward.

In April 1914 the family moved to Berlin, where Einstein had accepted a position with the Prussian Academy of Sciences, an arrangement that permitted him to continue his researches with only the occasional diversion of lecturing at the University of Berlin. His wife and two sons vacationed in Switzerland that summer and, with the eruption of World War I, were unable to return to Berlin. A few years later this enforced separation was to lead to divorce. Einstein abhorred the war and was an outspoken critic of German militarism among the generally acquiescent academic community in Berlin, but he was primarily engrossed in perfecting his general theory of relativity, which he published in *Annalen der Physik* as "Die Grundlagen der allgemeinen Relativitätstheorie" ("The Foundation of the General Theory of Relativity") in 1916. The heart of this postulate was that gravitation is not a force, as Newton had said, but a curved field in the space–time continuum, created by the presence of mass. This notion could be proved or disproved, he suggested, by measuring the deflection of starlight as it travelled close by the Sun, the starlight being visible only during a total eclipse. Einstein predicted twice the light deflection that would be accountable under Newton's laws.

General theory of relativity

His new equations also explained for the first time the puzzling irregularity—that is, the slight advance—in the planet Mercury's perihelion, and they demonstrated why stars in a strong gravitational field emitted light closer to the red end of the spectrum than those in a weaker field.

While Einstein awaited the end of the war and the opportunity for his theory to be tested under eclipse conditions, he became more and more committed to pacifism, even to the extent of distributing pacifist literature to sympathizers in Berlin. His attitudes were greatly influenced by the French pacifist and author Romain Rolland, whom he met on a war-time visit to Switzerland. Rol-

land's diary later provided the best glimpse of Einstein's physical appearance as he reached his middle thirties:

> Einstein is still a young man, not very tall, with a wide and long face, and a great mane of crispy, frizzled and very black hair, sprinkled with gray and rising high from a lofty brow. His nose is fleshy and prominent, his mouth small, his lips full, his cheeks plump, his chin rounded. He wears a small cropped mustache. (By permission of Madame Marie Romain Rolland.)

Einstein's view of humanity during the war period appears in a letter to his friend, the Austrian-born Dutch physicist Paul Ehrenfest:

> The ancient Jehovah is still abroad. Alas, he slays the innocent along with the guilty, whom he strikes so fearsomely blind that they can feel no sense of guilt. . . . We are dealing with an epidemic delusion which, having caused infinite suffering, will one day vanish and become a monstrous and incomprehensible source of wonderment to later generations. (From Otto Nathan and Heinz Norden [eds.], *Einstein on Peace;* Simon and Schuster, 1960.)

It would be said often of Einstein that he was naïve about human affairs; for example, with the proclamation of the German Republic and the armistice in 1918, he was convinced that militarism had been thoroughly abolished in Germany.

International acclaim. International fame came to Einstein in November 1919, when the Royal Society of London announced that its scientific expedition to Príncipe Island, in the Gulf of Guinea, had photographed the solar eclipse on May 29 of that year and completed calculations that verified the predictions made in Einstein's general theory of relativity. Few could understand relativity, but the basic postulates were so revolutionary and the scientific community was so obviously bedazzled that the physicist was acclaimed the greatest genius on Earth. Einstein himself was amazed at the reaction and apparently displeased, for he resented the consequent interruptions of his work. After his divorce he had, in the summer of 1919, married Elsa, the widowed daughter of his late father's cousin. He lived quietly with Elsa and her two daughters in Berlin, but, inevitably, his views as a foremost savant were sought on a variety of issues.

Proof of the general theory of relativity

Despite the now deteriorating political situation in Germany, Einstein attacked nationalism and promoted pacifist ideals. With the rising tide of anti-Semitism in Berlin, Einstein was castigated for his "Bolshevism in physics," and the fury against him in right-wing circles grew when he began publicly to support the Zionist movement. Judaism had played little part in his life, but he insisted that, as a snail can shed his shell and still be a snail, so a Jew can shed his faith and still be a Jew.

Although Einstein was regarded warily in Berlin, such was his demand in other European cities that he travelled widely to lecture on relativity, usually arriving at each place by third-class rail carriage, with a violin tucked under his arm. So successful were his lectures that one impresario guaranteed him a three-week booking at the London Palladium. He ignored the offer, but, at the request of the Zionist leader Chaim Weizmann, toured the United States in the spring of 1921 to raise money for the Palestine Foundation Fund. Frequently treated like a circus freak and feted from morning to night, Einstein nevertheless was gratified by the standards of scientific research and the "idealistic attitudes" prevailing in the United States.

During the next three years Einstein was constantly on the move, journeying not only to European capitals but also to the Orient, to the Middle East, and to South America. According to his diary notes, he found nobility among the Hindus of Ceylon, a pureness of soul among the Japanese, and a magnificent intellectual and moral calibre among the Jewish settlers in Palestine. His wife later wrote that on steaming into one new harbour, Einstein had said to her, "Let us take it all in before we wake up."

The Nobel Prize

In Shanghai a cable reached him announcing that he had been awarded the 1921 Nobel Prize for Physics "for your photoelectric law and your work in the field of theoretical physics." Relativity, still the centre of controversy, was not mentioned.

Though the 1920s were tumultuous times of wide acclaim, and some notoriety, Einstein did not waver from his new search—to find the mathematical relationship between electromagnetism and gravitation. This would be a first step, he felt, in discovering the common laws governing the behaviour of everything in the universe, from the electron to the planets. He sought to relate the universal properties of matter and energy in a single equation or formula, in what came to be called a unified field theory. This turned out to be a fruitless quest that occupied the rest of his life. Einstein's peers generally agreed quite early that his search was destined to fail because the rapidly developing quantum theory uncovered a probability factor in all measurements of the movement of electrons: the movement of a single electron simply could not be predicted because its position and speed at any instant could not be measured with equal precision, which means, in effect, that the future of any physical system at the subatomic level cannot be predicted. While fully recognizing the brilliance of quantum mechanics, Einstein rejected the idea that these theories were absolute and persevered with his theory of general relativity as the more satisfactory foundation to future discovery. He was widely quoted on his belief in an exactly engineered universe: "God is subtle but he is not malicious." On this point, he parted company with most theoretical physicists. The distinguished German quantum theorist Max Born, a close friend of Einstein, said at the time: "Many of us regard this as a tragedy, both for him, as he gropes his way in loneliness, and for us, who miss our leader and standard-bearer." This appraisal, and others pronouncing his work in later life as largely wasted effort, will have to await the judgment of later generations.

The year of Einstein's 50th birthday, 1929, marked the beginning of the ebb flow of his life's work in a number of aspects. Early in the year the Prussian Academy published the first version of his unified-field theory, but, despite the sensation it caused, its very preliminary nature soon became apparent. The reception of the theory left him undaunted, but Einstein was dismayed by the preludes to certain disaster in the field of human affairs: Arabs launched savage attacks on Jewish colonists in Palestine; the Nazis gained strength in Germany; the League of Nations proved so impotent that Einstein resigned abruptly from its Committee on Intellectual Cooperation as a protest to its timidity; and the stock market crash in New York City heralded worldwide economic crisis.

Crushing Einstein's natural gaiety more than any of these events was the mental breakdown of his younger son Edward, who had worshipped his father from a distance but now blamed him for deserting him and runining his life. Einstein's sorrow was eased only slightly by the amicable relationship he had with his older son, Hans Albert.

As visiting professor at Oxford University in 1931, Einstein spent as much time espousing pacifism as he did discussing science. He went so far as to authorize the establishment of the Einstein War Resisters' International Fund in order to bring massive public pressure to bear on the World Disarmament Conference, scheduled to meet in Geneva in February 1932. When these talks foundered, Einstein felt that his years of supporting world peace and human understanding had accomplished nothing. Bitterly disappointed, he visited Geneva to focus world attention on the "farce" of the disarmament conference. In a rare moment of fury, Einstein stated to a journalist,

> They [the politicians and statesmen] have cheated us. They have fooled us. Hundreds of millions of people in Europe and in America, billions of men and women yet to be born, have been and are being cheated, traded and tricked out of their lives and health and well-being.

Shortly after this, in a famous exchange of letters with the Austrian psychiatrist Sigmund Freud, Einstein suggested that people must have an innate lust for hatred

and destruction. Freud agreed, adding that war was biologically sound because of the love–hate instincts of man and that pacifism was an idiosyncrasy directly related to Einstein's high degree of cultural development. This exchange was only one of Einstein's many philosophic dialogues with renowned men of his age. With Rabindranath Tagore, Hindu poet and mystic, he discussed the nature of truth. While Tagore held that truth was realized through man, Einstein maintained that scientific truth must be conceived as a valid truth that is independent of humanity. "I cannot prove that I am right in this, but that is my religion," said Einstein. Firmly denying atheism, Einstein expressed a belief in "Spinoza's God who reveals himself in the harmony of what exists." The physicist's breadth of spirit and depth of enthusiasm were always most evident among truly intellectual men. He loved being with the physicists Paul Ehrenfest and Hendrick A. Lorentz at The Netherlands' Leiden University, and several times he visited the California Institute of Technology in Pasadena to attend seminars at the Mt. Wilson Observatory (now part of Hale Observatories), which had become world renowned for astrophysical research. At Mt. Wilson he heard the Belgian scientist Abbé Georges Lemaître detail his theory that the universe had been created by the explosion of a "primeval atom" and was still expanding. Gleefully, Einstein jumped to his feet, applauding. "This is the most beautiful and satisfactory explanation of creation to which I have ever listened," he said.

In 1933, soon after Adolf Hitler became chancellor of Germany, Einstein renounced his German citizenship and left the country. He later accepted a full-time position as a foundation member of the school of mathematics at the new Institute for Advanced Study in Princeton, New Jersey. In reprisal, Nazi storm troopers ransacked his beloved summer house at Caputh, near Berlin, and confiscated his sailboat. Einstein was so convinced that Nazi Germany was preparing for war that, to the horror of Romain Rolland and his other pacifist friends, he violated his pacifist ideals and urged free Europe to arm and recruit for defense.

Flight from Europe

Although his warnings about war were largely ignored, there were fears for Einstein's life. He was taken by private yacht from Belgium to England. By the time he arrived in Princeton in October 1933, he had noticeably aged. A friend wrote,

> It was as if something had deadened in him. He sat in a chair at our place, twisting his white hair in his fingers and talking dreamily about everything under the sun. He was not laughing any more.

Later years in the United States. In Princeton Einstein set a pattern that was to vary little for more than 20 years. He lived with his wife in a simple, two-story frame house and most mornings walked a mile or so to the Institute, where he worked on his unified-field theory and talked with colleagues. For relaxation he played his violin and sailed on a local lake. Only rarely did he travel, even to New York. In a letter to Queen Elisabeth of Belgium, he described his new refuge as a "wonderful little spot, . . . a quaint and ceremonious village of puny demigods on stilts." Eventually he acquired American citizenship, but he always continued to think of himself as a European. Pursuing his own line of theoretical research outside the mainstream of physics, he took on an air of fixed serenity. "Among my European friends, I am now called *Der grosse Schweiger* ("The Great Stone Face"), a title I well deserve," he said. Even his wife's death late in 1936 did not disturb his outward calm. "It seemed that the difference between life and death for Einstein consisted only in the difference between being able and not being able to do physics," wrote Leopold Infeld, the Polish physicist who arrived in Princeton at this time.

Development of the atomic bomb

Niels Bohr, the great Danish atomic physicist, brought news to Einstein in 1939 that the German refugee physicist Lise Meitner had split the uranium atom, with a slight loss of total mass that had been converted into energy. Meitner's experiments, performed in Copenhagen, had been inspired by similar, though less precise, experiments done months earlier in Berlin by two German chemists, Otto Hahn and Fritz Strassmann. Bohr speculated that, if a controlled chain-reaction splitting of uranium atoms could be accomplished, a mammoth explosion would result. Einstein was skeptical, but laboratory experiments in the United States showed the feasibility of the idea. With a European war regarded as imminent and fears that Nazi scientists might build such a "bomb" first, Einstein was persuaded by colleagues to write a letter to President Franklin D. Roosevelt urging "watchfulness and, if necessary, quick action" on the part of the United States in atomic-bomb research. This recommendation marked the beginning of the Manhattan Project.

Although he took no part in the work at Los Alamos, New Mexico, and did not learn that a nuclear-fission bomb had been made until Hiroshima was razed in 1945, Einstein's name was emphatically associated with the advent of the atomic age. He readily joined those scientists seeking ways to prevent any future use of the bomb, his particular and urgent plea being the establishment of a world government under a constitution drafted by the United States, Britain, and Russia. With the spur of the atomic fear that haunted the world, he said "we must not be merely willing, but actively eager to submit ourselves to the binding authority necessary for world security." Once more, Einstein's name surged through the newspapers. Letters and statements tumbled out of his Princeton study, and in the public eye Einstein the physicist dissolved into Einstein the world citizen, a kind "grand old man" devoting his last years to bringing harmony to the world.

The rejection of his ideals by statesmen and politicians did not break him, because his prime obsession still remained with physics. "I cannot tear myself away from my work," he wrote at the time. "It has me inexorably in its clutches." In proof of this came his new version of the unified field in 1950, a most meticulous mathematical essay that was immediately but politely criticized by most physicists as untenable.

Compared with his renown of a generation earlier, Einstein was virtually neglected and said himself that he felt almost like a stranger in the world. His health deteriorated to the extent that he could no longer play the violin or sail his boat. Many years earlier, chronic abdominal pains had forced him to give up smoking his pipe and to watch his diet carefully.

On April 18, 1955, Einstein died in his sleep at Princeton Hospital. On his desk lay his last incomplete statement, written to honour Israeli Independence Day. It read in part: "What I seek to accomplish is simply to serve with my feeble capacity truth and justice at the risk of pleasing no one." His contribution to man's understanding of the universe was matchless, and he is established for all time as a giant of science. Broadly speaking, his crusades in human affairs seem to have had no lasting impact. Einstein perhaps anticipated such an assessment of his life when he said "Politics are for the moment. An equation is for eternity."

MAJOR WORKS

SCIENTIFIC PAPERS: "Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt," in *Annalen der Physik* (1905); "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen," in *Annalen der Physik* (1905); " Zur Elektrodynamik bewegter Körper," in *Annalen der Physik* (1905), the initial paper on special relativity; "Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?" in *Annalen der Physik* (1905); "Zur Theorie der Brownschen Bewegung," in *Annalen der Physik* (1906), translated separately as *Investigations on the Theory of the Brownian Movement* (1926); "Zur Theorie der Lichterzeugung und Lichtabsorption," in *Annalen der Physik* (1906); "Planckschе Theorie der Strahlung und die Theorie der spezifischen Wärme," in *Annalen der Physik* (1907); "Entwurf einer Verallegemeinerten Relativitätstheorie und einer Theorie der Gravitation," in *Zeitschrift für Mathematik und Physik* (1913); "Grundlagen der allgemeinen Relativitätstheorie," in *Annalen der Physik* (1916), on the general theory of relativity; "Strahlungs-emission und -absorption nach der Quantentheorie," in *Verhandlungen der Deutschen physikalischen Gesellschaft* (1916); "Quantenthe-

orie der Strahlung," in *Physikalische Zeitschrift* (1917); "Quantentheorie des einatomigen idealen Gases," in *Sitzungsberichte der Preussischen Akademie der Wissenschaften* (1924 and 1925). Some of Einstein's important papers were collected in the joint work (with H.A. Lorentz and H. Minkowski), *H.A. Lorentz: Das Relativitätsprinzip, eine Sammlung von Abhandlungen* (1913; trans. as *H.A. Lorentz: The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity*, 1923). See also *The Meaning of Relativity*, which includes the generalized theory of gravitation (1953), the first edition of Einstein's unified field theory.

OTHER WORKS: *About Zionism: Speeches and Letters*, Eng. trans. by Sir Leon Simon (1931); *Builders of the Universe* (1932); with Sigmund Freud, *Warum Krieg? (Why War?*, Eng. trans. by Stuart Gilbert, 1933); with Leopold Infeld, *The Evolution of Physics* (1938); *The World As I See It* (Eng. trans. by Alan Harris, 1949); *Out of My Later Years* (1950).

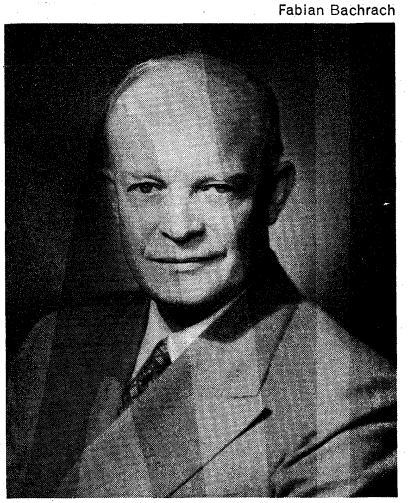
BIBLIOGRAPHY. PAUL A. SCHILPP (ed.), *Albert Einstein: Philosopher-Scientist*, 2nd ed. (1951), a discussion by eminent scholars of Einstein's impact on science and philosophy; PHILIPP FRANK, *Einstein: His Life and Times* (Eng. trans. 1947), a scientific biography on Einstein's early life and achievement; ANTONINA VALLENTIN, *Einstein: A Biography* (Eng. trans. 1954), a personal story of Einstein's European years; PETER MICHELMORE, *Einstein: Profile of the Man* (1962), a popular biography, richly anecdotal, of Einstein as man and scientist; LINCOLN BARNETT, *The Universe and Dr. Einstein*, 2nd rev. ed. (1957), a lucid exposition of Einstein's contribution to science; RONALD W. CLARK, *Einstein: The Life and Times* (1971), a distinguished, definitive work (well illustrated); BANESH HOFFMAN, with the collaboration of HELEN DUKAS, *Albert Einstein: Creator and Rebel* (1972), a significant biography, laced with a thorough but exciting interpretation of Einstein's scientific work.

(P.Mi.)

Eisenhower, Dwight D.

After a slow rise in the military hierarchy, Dwight David Eisenhower became commander of the Allied forces in World War II, and in 1952, a popular hero both through his achievements and his personality, was elected the 34th president of the United States. A colonel at the outbreak of World War II, his ability to transform strategic theory into effective action, his winning personality, and his singular diplomatic gifts won him rapid promotion. His command of the successful amphibious landing in North Africa in 1942 led to his appointment as supreme commander of the invasion of the continental mainland in 1944, and in 1952 he was swept to victory in the presidential election. Although his two terms in office were devoid of notable achievement either at home or abroad, he succeeded in keeping the United States out of war.

Fabian Bachrach



Eisenhower, 1952.

He was born in Denison, Texas, on October 14, 1890, the third of seven sons of David Jacob and Ida Elizabeth (Stover) Eisenhower. In the spring of 1891 the Eisenhowers returned to Abilene, Kansas, where their ancestors had settled as part of a Mennonite colony. Dwight's father worked in a creamery; the family was poor, and Dwight and his brothers were introduced to hard work and a strong religious tradition at an early age.

Early career. "Ike," as Dwight was called, was a fun-loving, personable youth who enjoyed sports and an occasional fistfight, while taking only a moderate interest in his studies. He graduated from Abilene High School in 1909, supported a brother's college education for a year, and then entered the United States Military Academy at West Point. He excelled in football but injured a knee in his second year at the Academy and was forced to stop playing. In the remarkable class of 1915—which was to produce 59 generals—he ranked 61st academically and 125th in discipline out of the total of 164 graduates.

Education

Commissioned a second lieutenant, he was sent to San Antonio, Texas, where he met Mamie Geneva Doud, daughter of a successful Denver meat packer. They were married in 1916 and had two sons: Doud Dwight, born in 1917, who died of scarlet fever in 1921, and John Sheldon Doud, born in 1922.

During World War I Eisenhower commanded a tank training centre, was promoted to captain, and received the Distinguished Service Medal. The war ended just before he was to be sent overseas. From 1922 to 1924 he was assigned to the Panama Canal Zone and there came under the inspiring influence of his commander, Brig. Gen. Fox Conner. With Conner's assistance, Eisenhower was selected to attend the army's command and general staff school at Ft. Leavenworth. Then a major, he graduated first in a class of 275 in 1926 and two years later graduated from the Army War College. He then served in France, writing a guidebook of World War I battlefields, and in Washington, D.C., before becoming an aide to army chief of staff Gen. Douglas MacArthur in 1933. Two years later he accompanied MacArthur to the Philippines to assist in the reorganization of the commonwealth's army and while there was awarded the Distinguished Service Star of the Philippines and the rank of lieutenant colonel. He returned to the United States shortly after Germany's invasion of Poland initiated the European phase of World War II and in March 1941 became a full colonel. Three months later he was made chief of staff of the 3rd Army and soon won the attention of army chief of staff Gen. George C. Marshall for his role in planning war games involving almost 500,000 troops.

Supreme commander in World War II. When the United States entered World War II in December 1941, Marshall appointed Eisenhower to the army's war plans division in Washington, where he prepared strategy for an Allied invasion of Europe. Having become a brigadier general in September 1941, he was promoted to major general in March 1942 and named head of the operations division of the War Department. In June, Marshall selected him over 366 senior officers to be commander of United States troops in Europe. Eisenhower's rapid advancement, after a long army career spent in relative obscurity, was due not only to his knowledge of military strategy and a talent for organization but also to his ability to persuade, to mediate, and to be agreeable. Men from a variety of backgrounds and nationalities, impressed by his friendliness, humility, and persistent optimism, liked and trusted him.

Promoted to lieutenant general in July 1942, Eisenhower was named to head Operation Torch, the Allied invasion of French North Africa. This first major Allied offensive of the war was launched on November 8, 1942, and successfully completed in May 1943. Eisenhower's decision to work during the campaign with the French admiral Jean-François Darlan, who had collaborated with the Germans, aroused a storm of protest from the Allies; but his action was defended by Pres. Franklin D. Roosevelt. A full general since February, Einsenhower then directed the amphibious assault of Sicily and the Italian mainland, which resulted in the fall of Rome on June 4, 1944.

During the fighting in Italy, Eisenhower participated in plans to cross the English Channel for an invasion of

Invasion of Europe

France. On December 24, 1943, he was appointed supreme commander of the Allied Expeditionary Forces, and the next month he was in London making preparations for the massive thrust into Europe. On June 6, 1944, he gambled on a break in bad weather and ordered the Channel crossed. About 1,000,000 men in 4,000 ships landed in Normandy and began to fight their way into the heart of France. On August 25 Paris was liberated. The Allied forces continued eastward, following Eisenhower's "broad-front" plan, which ultimately distributed armies along Germany's entire western border. After overcoming a fierce German counterattack in the Ardennes in December, the Allies crossed the Rhine on March 7, 1945. Germany surrendered on May 7, and the war in Europe was over. In the meantime, in December 1944, Eisenhower had been promoted to the five-star rank of general of the army.

He was given a hero's welcome upon returning to the United States for a visit in June 1945. His intended retirement was delayed by Pres. Harry S. Truman when he was named in November to replace General Marshall as chief of staff. For over two years Eisenhower directed demobilization of the wartime army and worked to unify the armed services under a centralized command. He ended speculation that he might run for the presidency in 1948 by refusing to permit his name to appear on an early primary ballot.

In May 1948 he left active duty, the most popular and respected soldier in the United States, to become president of Columbia University. His book *Crusade in Europe*, published that fall, became a best seller and made him a wealthy man.

Eisenhower's brief career as an academic administrator was not especially successful. His technical education and military experience prepared him poorly for the post, and his distrust of intellectuals made relations with faculty members difficult. In the fall of 1950 President Truman asked him to become supreme commander of the North Atlantic Treaty Organization (NATO), and in early 1951 he flew to Paris to assume his new position. For the next 15 months he devoted himself to the task of creating a united military organization in western Europe to be a defence against the possibility of Communist aggression.

Supreme commander of NATO

First term as president. As early as 1943 Eisenhower was mentioned as a possible presidential candidate. His personal qualities and military reputation, rather than his often vague pronouncements on public policy, prompted both parties to woo him. As the campaign of 1952 neared, Eisenhower let it be known that he was a Republican, and the Eastern wing of the party, headed by Gov. Thomas E. Dewey of New York, made an intensive effort to persuade him to seek the Republican presidential nomination. His name was entered in several state primaries against the more conservative Sen. Robert A. Taft of Ohio. Although the results were mixed, Eisenhower decided to run. In June 1952 he retired from the army, after 37 years of service, returned to the United States, and began to campaign actively. At the party convention in July, after a bitter fight with Taft supporters, Eisenhower won the nomination on the first ballot. His running mate was Sen. Richard M. Nixon of California. Democrats nominated Gov. Adlai E. Stevenson of Illinois for president and Sen. John Sparkman of Alabama for vice president.

Despite his age, Eisenhower campaigned tirelessly, impressing millions with his warmth and sincerity. He urged economy and honesty in government and promised to visit Korea to explore the possibilities for ending the Korean War, which had broken out in 1950 between Communist North Korea and pro-Western South Korea and soon involved United Nations (mainly United States) and Communist Chinese forces. Many Republicans, including Nixon, spoke of pro-Communist disloyalty within the Truman administration and called for stringent anti-subversive measures. The Eisenhower–Nixon ticket won handily, carrying 39 states, winning the electoral vote, 442 to 89, and collecting more than 33,000,000 popular votes. The Republican Party won control of Congress by a slim margin but lost both houses two years later.

Conception of the presidency

Unlike his two Democratic predecessors, Eisenhower did not believe in strong executive leadership of the federal government. Because of his military background and limited knowledge of government and politics, he liberally delegated authority to his advisers and cabinet members and did not participate actively in the detailed work of the executive departments. His assistant, former New Hampshire governor Sherman Adams, was a powerful figure in the administration. The Chief Executive, Adams wrote,

> expected me to manage a staff that would boil down, simplify and expedite the urgent business that had to be brought to his attention and to keep as much work of secondary importance as possible off his desk.

Policies emanating from the White House, even after Adams' resignation in 1958, usually involved the President only in the final stages.

Eisenhower's basically conservative views on domestic affairs were shared by his influential secretary of the treasury, George M. Humphrey. The Administration's domestic program, which came to be labelled "modern Republicanism," called for reduced taxes, balanced budgets, a decrease in government control over the economy, and the return of certain federal responsibilities to the states. Controls over rents, wages, and prices were allowed to expire, and in 1954 there was a slight tax revision. At Eisenhower's insistence, Congress transferred title to valuable tideland oil reserves to the states. But there was no sharp break with policies inherited from previous Democratic administrations. The needs of an expanding population (from 155,000,000 to 179,000,000 during the Eisenhower era) and the nation's overseas commitments caused budget deficits during five out of eight years. The minimum wage was increased to $1 per hour, the Social Security System was broadened, and in the spring of 1953 the Department of Health, Education, and Welfare was created.

The right wing of his party clashed with the President more often than did the Democrats during his first term. In part to preserve party unity, he refused to condemn publicly Sen. Joseph R. McCarthy's irresponsible charges of Communist influence within the government. Although privately Eisenhower expressed his distaste for the Senator, at times he seemed to encourage the attacks of McCarthyites. Hundreds of federal employees were fired under his expanded loyalty-security program. With his approval, Congress passed a law designed to outlaw the American Communist Party. Following the sensational hearings on McCarthy's charges against army officers and civilian officials, televised nationally for five weeks in the spring of 1954, McCarthy's popularity waned, as did much of the anti-Communist hysteria of the postwar years.

Foreign policy

Foreign affairs drew much of Eisenhower's attention; he and his secretary of state John Foster Dulles worked hard at achieving world peace and constructing collective defense agreements designed to check the spread of Communism. As promised, Eisenhower visited Korea shortly after his inauguration. Partly, perhaps, because of Joseph Stalin's death in March 1953, the President was able to negotiate a truce for the Korean War in July 1953. In December of that year, he proposed that the nations of the world pool atomic information and materials under the auspices of an international agency. This "Atoms for Peace" suggestion bore fruit in 1957, when 62 countries formed the International Atomic Energy Agency.

In July 1955 the President met with leaders of Great Britain, France, and the Soviet Union at a summit conference in Geneva. His "open-skies" proposal, by which the United States and the Soviet Union would permit continuous air inspection of each other's military installations, was welcomed by world opinion but was rejected by the U.S.S.R.

In September 1954, Eisenhower and Dulles succeeded in creating the Southeast Asia Treaty Organization (SEATO) to prevent further Communist expansion in that

part of the world. It was composed of the United States, France, Great Britain, Australia, New Zealand, the Philippines, Thailand, and Pakistan. NATO was strengthened in 1955 by the inclusion of West Germany.

Critics contended that there were frequent disparities between the administration's words and deeds in the field of foreign relations. While threatening to "unleash" Nationalist Chinese leader Chiang Kai-shek, the United States signed a defense treaty with Nationalist China in December 1954 that inhibited Chiang's ability to attack the Communist Chinese. Dulles spoke of "liberating" captive peoples in Communist nations, but the administration limited itself to protests when uprisings occurred in East Germany (1953) and Hungary (1956). While the Secretary of State promised "massive retaliation" against Communist aggression, the President made the decision to limit America's role in the Indochina crisis between France and Ho Chi Minh's guerillas to only financial and military aid. Many observers praised Eisenhower for his restraint and moderation under pressure, however.

Second term. A heart attack in September 1955, and an operation for ileitis in June 1956, raised considerable doubt about Eisenhower's ability to serve a second term. But he recovered quickly, and the Republican convention unanimously endorsed the Eisenhower–Nixon ticket on the first ballot. Democrats again selected Adlai E. Stevenson and named Sen. Estes Kefauver of Tennessee as his running mate. Eisenhower's great personal popularity turned the election into a landslide victory, the most one-sided race since 1936. He won 35,500,000 votes against Stevenson's 25,000,000; the electoral count was 467 to 74. The Democrats, however, once more captured both houses of Congress, a feat they were to duplicate in 1958. Eisenhower was the first president to serve with three Congresses controlled by the opposition party.

The election campaign of 1956, however, had been complicated by a crisis in the Middle East over Egypt's seizure of the Suez Canal. The subsequent attack by Great Britain, France, and Israel on Egypt and Egypt's support by the Soviet Union prompted the President to go before Congress in January 1957 to urge adoption of what came to be called the Eisenhower Doctrine, a pledge to send United States armed forces to any Middle Eastern nation requesting assistance against Communist aggression. In July 1958 American troops were sent to Lebanon to protect its government from attack by Communist-inspired rebels.

Role in racial desegregation

When the United States Supreme Court, on May 17, 1954, declared racial segregation in public schools unconstitutional, controversy and violence broke out, especially in the south. In September 1957 Eisenhower dispatched 1,000 federal troops to Little Rock, Arkansas, to halt an attempt by Gov. Orval E. Faubus to obstruct a federal court order integrating a local high school. This action was the most serious challenge of his presidency. On several occasions Eisenhower had expressed his distaste for racial segregation, though he doubtless believed that the process of integration would take time. The Civil Rights Act of 1957 was the first such law passed since 1875.

On October 4, 1957, the Soviet Union launched Sputnik I, the first man-made satellite to orbit the Earth. Americans were shocked by the achievement, and many blamed Eisenhower for the administration's insistence on low military budgets and its failure to develop a space program. Steps were taken to boost the nation's missile and space research and to provide funds to increase the study of science. The National Aeronautics and Space Administration was created in July 1958.

The administration again came under fire in the fall of 1957 for a recession in the economy that lasted through the following summer. The President refused to lower taxes or increase federal spending to ease the slump for fear of fuelling inflation.

Following the death of Dulles in the spring of 1959, Eisenhower assumed a more personal role in the direction of American foreign policy. He travelled to Europe, Africa, and Asia in December 1959 and toured Latin America in February and March 1960 seeking friendship for the United States.

The U-2 incident

To improve relations with the Soviet Union, the President invited Premier Nikita S. Khrushchev to visit the United States. Khrushchev toured parts of the country in September 1959 and held private talks with Eisenhower. Another summit meeting was planned, and a new era of personal diplomacy seemed at hand. But when a United States U-2 reconnaissance plane was shot down over the U.S.S.R. in May 1960 Khrushchev scuttled the talks and angrily withdrew his invitation to the President to visit the Soviet Union. Eisenhower admitted that the flights had gone on for four years and shouldered much of the blame for the ill-timed affair. In January 1961, during the last weeks of the Eisenhower administration, the United States broke diplomatic relations with Cuba, for two years under the control of Fidel Castro.

Though his administrations had a great many critics, Eisenhower remained extraordinarily popular with the American people. When he left office, the oldest president in the nation's history, Congress restored his rank as general of the army. He retired to his farm in Gettysburg, Pennsylvania, and devoted much of his time to his memoirs. In 1963 he published *Mandate For Change*, which was followed in 1965 by *Waging Peace*. A lighter work, *At Ease: Stories I Tell to Friends*, appeared in 1967. On March 28, 1969, after a series of illnesses, he died of heart failure in a hospital in Washington, D.C. He was buried in Abilene, Kansas.

BIBLIOGRAPHY. For a convincing interpretation of Eisenhower's entire career, see WILLIAM CHILDS, *Eisenhower: Captive Hero* (1958). STEPHEN E. AMBROSE, *The Supreme Commander: The War Years of General Dwight D. Eisenhower* (1970), is scholarly and sound. ARTHUR LARSON, *Eisenhower: The President Nobody Knew* (1968), is a favourable view of the Eisenhower presidency by a friend. EMMET JOHN HUGHES, *The Ordeal of Power* (1963), has a brilliant critique of the first term. DEAN ALBERTSON (ed.), *Eisenhower As President* (1963), contains first-rate assessments.

(T.C.R.)

Eisenstein, Sergey

Three of the undisputed classics of the art of the cinema —*Potemkin* (1925), *Alexander Nevsky* (1938), and *Ivan the Terrible* (*Part One*, 1944; *Part Two*, 1946, released 1958)—are the work of the Russian director and theorist Sergey Eisenstein, whose writings, largely on the achievement of rhythms of mood and meaning through editing, or montage, and films have been instrumental in creating a film aesthetic. Since even his lesser films reflect his searching genius, his total body of work has had an immense impact on the development of the art. If sensitivity to the expressive possibilities of his medium is the mark of a great artist, then Eisenstein must be counted, as many of his followers proclaim him, the greatest master of the art of the film.

Birth and early life

Of Jewish descent through his paternal grandparents, Sergey Mikhaylovich Eisenstein was born on January 23, 1898, at Riga, where his father, Mikhail, a civil engineer, worked in shipbuilding. In 1910 the family left Riga for St. Petersburg (now Leningrad).

After studying in 1916–18 at the Institute of Civil Engineering, Eisenstein decided on a career in the plastic arts and entered the School of Fine Arts.

With the outbreak of the Russian Revolution of 1917, he enlisted in the Red Army and helped to organize and construct defenses and to produce entertainment for the troops. Having now found his vocation, he entered, in 1920, the Proletkult Theatre (Theatre of the People) in Moscow as an assistant decorator. He rapidly became the principal decorator and then the co-director. As such, he designed the costumes and the scenery for several notable productions. At the same time, he developed a strong interest in the Kabuki theatre of Japan, which was to influence his ideas on film. For his production of *The Wise Man*, an adaptation of Aleksandr Ostrovsky's play, he made a short film, "Glumov's Diary," which was shown as part of the performance in 1923. Soon afterward the cinema engaged his full attention, and he pro-

Eisenstein, on location for *October* in 1927.
Sovfoto

duced his first film, *Strike*, in 1924, after having published his first article on theories of editing in the review *Lef*, edited by the great poet Vladimir Mayakovsky. He said there that in place of the static reflection of an event, expressed by a logical unfolding of the action, he proposed a new form: the "montage of attractions"—in which arbitrarily chosen images, independent of the action, would be presented not in chronological sequence but in whatever way would create the maximum psychological impact. Thus, the film maker should aim to establish in the consciousness of the spectators the elements that would lead them to the idea he wants to communicate; he should attempt to place them in the spiritual state or the psychological situation that would give birth to that idea.

First films and cinematic theory

These principles guided Eisenstein's entire career. In the realistic films that he undertook, however, such a technique is effective only when it utilizes the concrete elements implicit in the action; it loses validity when its symbols are imposed upon reality instead of being implied by it. Thus, in *Strike* (1924), which recounts the repression of a strike by the soldiers of the tsar, Eisenstein juxtaposed shots of workers being mowed down by machine guns with shots of cattle being butchered in a slaughterhouse. The effect was striking, but the objective reality was falsified.

Possessed by his theory, Eisenstein was bound to succumb often to this failing. *Potemkin* (1925), also called *The Battleship Potemkin*, happily escaped it. Ordered by the Central Executive Committee of the U.S.S.R. to commemorate the Revolution of 1905, the film, made in the port and the city of Odessa in 1925, had a momentous impact and still remains among the masterpieces of the world cinema. (In 1958 it was voted the best film ever made, by an international poll of critics.) Its greatness lies not merely in the depth of humanity with which the subject is treated, nor in its social significance, nor in the formal perfection of its rhythm and editing; but rather, it is each of these magnified and multiplied by the others.

Having by this accomplishment earned recognition as the epic poet of the Soviet cinema, Eisenstein next made a film entitled *October* (1928) or *Ten Days that Shook the World*, which in the space of two hours dealt with the shifts of power in the government after the 1917 Revolution, the entrance on the scene of Lenin, and the struggle between the Bolsheviks and their political and military foes. If the film was sometimes inspired, it was also disparate, chaotic, and often confused.

Also uneven, but better balanced, was *Old and New* (originally titled *The General Line*), filmed in 1929 to illustrate the collectivization of the rural countryside. Eisenstein made of it a lyric poem, as calm and as expansive as *Potemkin* had been violent and compact.

In 1929, putting to profit a visit to Paris, he filmed *Romance sentimentale* (*Sentimental Melody*), an essay in counterpoint of images and music. Engaged by Paramount studios in 1930, he left for Hollywood, where he worked on adaptations of the novels *L'Or* ("Sutter's Gold"), by Blaise Cendrars, and *An American Tragedy*, by Theodore Dreiser. Refusing to modify his scripts to meet studio demands, however, he broke the contract and went to Mexico in 1932 to direct *Que viva Mexico!*, with capital collected by the novelist Upton Sinclair.

The film never was completed. After disagreements between Eisenstein and Sinclair, some of the negatives were sold and released as the films *Thunder over Mexico*, *Eisenstein in Mexico*, and *Death Day* (1933–34). In 1939 a fourth film, entitled *Time in The Sun*, was made from the footage. None of these films bears more than a distant resemblance to the original conception.

After his return to Moscow in 1933, Eisenstein undertook *Bezhin Meadow*. Several weeks before its completion, however, he was ordered to suspend its production. The scenes already shot were put together by Eisenstein, but the film, which was never released, was attacked as "formalistic" because of its poetic interpretation of reality. Eisenstein thus suffered from the same governmental policies toward art that had embroiled the composer Sergey Prokofiev, the writer Isaac Babel, and many other artists in difficulties with Soviet officialdom.

The Russian epics

Having expressed contrition for the errors of his past works, Eisenstein was able to make a film recounting the medieval epic of *Alexander Nevsky*, in accordance with Stalin's policy of glorifying Russian heroes. Made in 1938, this film transfigured the actual historical events, majestically leading to a final resolution that represented the triumph of collectivism. As in medieval epics, the characters were the strongly stylized heroes or demigods of legend. Produced in close collaboration with Prokofiev, who wrote the score, the film represented a blend of images and music into a single rhythmic unity, an indissoluble whole.

During World War II Eisenstein achieved a work of the same style as *Alexander Nevsky* and even more ambitious—*Ivan the Terrible*—about the 16th-century tsar Ivan IV, whom Stalin admired. Begun in 1943 in the Ural Mountains, the first part was finished in 1944, the second at the beginning of 1946. A third part was envisaged, but Eisenstein, suffering from an angina of the chest, had to take to his bed for several months. He was about to return to work when he was laid low again. He died on February 11, 1948, only a few days after his 50th birthday.

Most critics would agree that though Eisenstein's three greatest films stand far above the others, all of his work is significant; their flaws are those common to artists probing the limits of their craft. It may be that in the entire history of motion pictures, no other film maker has surpassed him in his understanding of his art.

MAJOR WORKS

FILMS: *Strike* (1924); *The Battleship Potemkin* (1925); *October* (*Ten Days That Shook the World;* 1928); *Old and New* (*The General Line;* 1929); *Alexander Nevsky* (1938); *Ivan the Terrible, Part One* (1944); *Ivan the Terrible, Part Two* (1958).

COLLECTIONS OF WRITINGS: *The Film Sense* (1942); *Film Form* (1949); *Notes of a Film Director* (1958); *Film Essays* (1968).

BIBLIOGRAPHY. MARIE SETON, *Eisenstein* (1952; Eng. trans., 1952); JEAN MITRY, *S.M. Eisenstein* (1955), in French; LEON MOUSSINAC, *Serge Eisenstein* (1964; Eng. trans., 1970).

(Je.M.)

Ekman, Vagn Walfrid

Vagn Walfrid Ekman, a Swedish physical oceanographer who pioneered in developing the science of oceanography, is particularly noted for his fundamental hydrodynamic theories concerning ocean currents.

Born on May 3, 1874, in Stockholm, Ekman was the youngest son of Fredrik Laurentz Ekman, a Swedish physical oceanographer. After finishing secondary school in Stockholm, Ekman studied at the University of Uppsala, where he majored in physics. But lectures on hydrodynamics in 1897 by Vilhelm Bjerknes, one of the founders of meteorology and oceanography, definitely decided the direction of Ekman's work.

While still a student at Uppsala, Ekman made important contributions to oceanography. When it was observed, during the Norwegian North Polar Expedition, that drift ice did not follow the wind direction but deviated by 20° to 40°, Bjerknes chose Ekman to make a theoretical study of the problem. In his report, published in 1902, Ekman took into account the balance of the frictions between the wind and sea surface, within layers of water, and the deflecting force due to the earth's rotation (Coriolis force).

After taking his degree at Uppsala in 1902, he joined the staff of the International Laboratory for Oceanographic Research in Oslo, where he remained until 1909. During those years he proved to be a skilled inventor and experimentalist. The Ekman current meter, an instrument with a simple and reliable mechanism, has been used, with subsequent improvements, to the present, while the Ekman reversing water bottle is used in freshwater lakes and sometimes in the ocean to obtain water samples at different depths with a simultaneous measurement of water temperatures. He displayed his theoretical and experimental talents in his study of "dead water," which causes slow-moving boats to become stuck because of a thin layer of nearly fresh water spreading over the sea from melting ice. This phenomenon, often occurring in fjords, seriously impeded the Norwegian explorer Fridtjof Nansen in Arctic waters. Ekman demonstrated by experiments in a wave tank that the resistance to the motion of the vessels is increased by the waves that are formed at the interface between layers of water of different densities.

By courtesy of the Lunds Universitet, Sweden

Ekman, 1928.

Study of "dead water" from melting ice

He also derived an empirical formula for the mean compressibility (compression ratio divided by pressure) of seawater as a function of pressure and temperature. This formula is still in use today to determine density of deep seawater which is compressed by hydrostatic pressure.

From 1910 to 1939, Ekman was professor of mechanics and mathematical physics at the University of Lund in Sweden, where he pursued his main interest, the dynamics of ocean currents. He published theories on wind-driven ocean currents including the effects of coasts and bottom topography, and on the dynamics of the Gulf Stream. He also tried, with partial success, to solve the complex problem of ocean turbulence.

In 1925 Ekman participated in a cruise of a German research ship to the Canary Islands. Again, on finding that data on currents obtained for several days at several marine stations between the Bay of Biscay and the Canary Islands, were not sufficient to obtain an average figure, he and a colleague, during the years 1922–29, improved the technique for measuring currents for a prolonged period by collecting data from an anchored ship. After several preparatory cruises for that purpose off the Norwegian coast aboard a Norwegian research vessel, they made a cruise to the trade-wind region off northwestern Africa in the summer of 1930 to determine the average current at various ocean depths at stations occupied for two weeks or longer. Preliminary reports were published soon after the cruise, but Ekman wrote the final report in 1953 at the age of 79. The long delay in publication, partly because of the loss of important data during the German occupation of Norway, also indicates the unparalleled care he took with his work.

Influence and assessment

His name is used in current terminology: "Ekman layer" denotes either the layers of the ocean or the layers of the atmosphere at the air–sea, air–ground, and sea–ground interfaces; "Ekman spiral" for velocity distributions of the wind or ocean currents in these layers; and also "Ekman transport" for water transport caused by wind.

Although his name and achievements were well known among oceanographers, he rarely attended international meetings—but his genuine kindliness prevented him from becoming a recluse. Most of his teachers and friends, such as Nansen and Bjerknes, were Norwegian; he spent many vacations in Bergen. He sang a beautiful bass, spent much time at the piano, and occasionally composed music. In the fall of 1953, he began a study of turbid currents, which he continued until a few days before his death, on March 9, 1954 at Gostad, near Stockaryd, Sweden.

BIBLIOGRAPHY. A. DEFANT, *Physical Oceanography*, vol. 1, pp. 41–43, 382–450 and vol. 2, pp. 521–527 (1961), a description of Ekman's work on the density of ocean water, on wind-driven circulation theories, and on "dead water"; W. HANSEN, "Vagn Walfrid Ekman: In Memoriam," *Dt. Hydrogr. Z.*, 7:65–66 (1954), a summary of Ekman's many scientific works; B. KULLENBERG, "Vagn Walfrid Ekman, 1874–1954," *J. Cons. Perm. Int. Explor. Mer*, 20:140–143 (1954), a brief biography emphasizing Ekman's personality and career; G. NEUMANN, *Ocean Currents*, pp. 15–17, 180–197 (1968), a description of Ekman's current meter and wind-driven current theories.

(T.Ic.)

Elasticity

In the broadest sense, a body has elasticity when it has the power of recovery. Thus, a china cup is not considered elastic when it falls to the floor and breaks. A straight strip of tempered steel will spring back from a bent position, even when it has been held so for a long time, whereas a strip of lead will not; the steel is considered to be elastic, the lead is not. The word elasticity was coined in the 17th century when some appreciation was being gained of the fact that all substances are more or less elastic, whether inanimate or part of living organisms, though elastic behaviour terminates rather early with some substances due either to their plasticity or to their brittleness. Even a gas may be said to spring back to its original volume when it has been compressed, and in 1662 Robert Boyle, the English physicist, noted the "elastical power" of air. In 1678, Robert Hooke, another English physicist, maintained that in any "springy body" displacement is proportional to load. Not until much later were stress and strain given precise meanings. This lack of definition, however, did not deter the mathematicians of the 18th century from examining problems such as flexure of beams; the essential steps toward a sound theory of elasticity were taken between 1820 and 1830 by French mathematicians, notably by Louis-Augustin Cauchy.

The phenomenon of elasticity. In 1807 Thomas Young, the English physicist and physician, proposed an elastic modulus, or a constant, relating unidirectional stress and strain. If a bar of uniform cross-sectional area

A is stretched by a force T, so that the stress (σ) denotes the force per unit area, then stress is expressed by $\sigma = T/A$. Under this stress the bar, of original length L, will stretch by a length indicated by δL to give a total length of $L + \delta L$. Unidirectional strain e is the fractional elongation; that is, the increase in length divided by the original length, so that $e = \delta L/L$. From this, the modulus of elasticity, called Young's modulus, can now be defined. A modulus is a numerical measure of some effect, in this case of elasticity, stated in terms of a unit condition: force per unit area of the medium divided by the fractional elongation, or the strain, is a measure of the effect of elasticity. Young's modulus, therefore, is the constant of proportionality between stress and strain: $E = \sigma/e$. It expresses the inherent stiffness of the material under tensile stress (stress resulting in a longitudinal stretching), regardless of the size of the bar.

Young's modulus

Although Young had some knowledge of numerical modulus values, it was somewhat later that testing machines came into regular use for examining constructional materials. Sensitive extensometers for measuring extensions of the order of $\frac{1}{100,000}$ of an inch (10^{-5} inch) were also needed for accurate determination of elastic behaviour. Notable advances in the techniques of testing materials, particularly in the use of optical extensometers, were made in German research schools from about 1870.

At the present time, dynamic methods are often used for measuring Young's modulus. A bar of uniform section may be vibrated along its length (in its fundamental longitudinal mode) using a piezoelectric vibrator. The periodic time is the time required for a stress pulse to travel from one end of the bar and back, or twice the bar length L; however, the velocity v of the propagation of this stress pulse turns out to be such that $v^2 = E/\rho$, ρ being the density of the bar. Hence E, the modulus, may be calculated from the equation that relates E to the product of the density ρ and the square of the length of the bar, L^2, and the square of observed resonant frequency, f^2, that is, $E = 4\rho L^2 f^2$.

The idea that torsional stiffness (resistance to twisting) depends on a shear modulus was accepted before the theory of torsion was put on a firm general basis by the French engineer A.J.C. Barré de Saint-Venant in 1853. A cylindrical shell with length L and radius r may be thought of as fixed at one end. If the other end of the shell is twisted through an angle theta (θ) the engineering shear strain phi (ϕ) is defined as the angle θ multiplied by the ratio of radius r to length L; that is, $\phi = \theta r/L$. Shear modulus G (also called the rigidity modulus) is the constant that relates shear strain phi, (ϕ), to shear stress tau, (τ), in an elastic body: shear stress equals the shear modulus times the shear strain, so that $\tau = G\phi$. By considering a solid bar of outside radius R to be made up of a series of such shells, it is possible to calculate the total torsional moment M as equal to the product of four terms: (1) pi (π), (2) the outside radius to the fourth power, (3) the sheer modulus, and (4) the shear strain, all divided by twice the length of the shell. In equation form, $M = \pi R^4 G\theta/2L$. Hence, from a graph of measured M and θ values, G can be found.

Shear modulus

Elasticity in viscous and crystalline bodies. Although matter in a condensed state is inherently elastic, by reason of bonding forces between molecules, distortional elasticity can become evident only when relative sliding between molecules is inhibited and the resulting resistance to flow is relatively high. In materials made up of crystals, the atoms within each crystal are tightly packed in a regular array. Other materials are amorphous; that is, they consist of randomly assembled atoms, for example, glass and viscous liquids. On an atomic scale, flow in an amorphous substance depends on the existence within the material of holes sufficiently large to allow the transfer of an atom from one position to another. Because there is thermal agitation of atoms within all materials at normal temperatures, the frequency with which an atom will gain enough energy to accomplish a transfer may be predicted theoretically by the application of certain equations of the Maxwell-Boltzmann type derived from kinetic molecular theory. Such methods of statistical mechanics (see THERMODYNAMICS, PRINCIPLES OF) have been used to explain the linear relation between stress and strain rate in a viscous liquid (which is then termed a Newtonian liquid) and also the effect of temperature and free volume on the coefficient of viscosity (equal to shearing stress divided by the rate of shear). It is a familiar observation that amorphous materials, such as wax or pitch, will soften progressively when heated but will behave elastically when cold or when subjected to stress applied rapidly.

When a liquid contains a moderate concentration of solid particles in suspension, its apparent viscosity (a property that inhibits the motion of one part of the liquid relative to another) is increased. Many concentrated solutions of organic substances, however, tend to form a gel (jellylike, having elastic properties), when allowed to rest, through the growth of a network of solute molecules. Gelatinous material, for example, develops a definite yield strength, or resistance to stress, on setting that enables it to oscillate when it is shaken.

Polymers, substances of high molecular weight made up of a large number of identical (or practically identical) smaller molecules, can exist as fluids or gels, or they may be glassy or crystalline. A polymer such as rubber contains long molecules twisted around each other, and the extension by pulling or twisting of such a substance involves straightening or crumpling of molecular chains. Distortion involves viscous shearing as parts of the molecular chains of rubber move relative to each other. On release of stress the chains resume their former convoluted shape and the strain due to elasticity is gradually recovered. Fillers such as carbon modify the open structure of rubber, increasing its elastic modulus. Internal structure obviously has an important effect on the elastic modulus of fibre-reinforced plastics; it is at a maximum when stress is applied in the direction of the fibres, and when the fibres are kept straight during molding.

A piece of metal generally consists of an aggregate of large numbers of crystals formed during solidification, and these crystals persist through subsequent heating or forging. Although creep (a slow change in shape) can be due partly to movements of atoms around vacant lattice sites, the plastic deformation of metals takes place chiefly through movement of dislocations. These defects in the stacking arrangement of atoms within crystals have the effect of reducing the yield strength to about one-hundredth part of the strength of a perfect crystal. Studies of dislocations, carried out extensively in many countries since about 1945, have succeeded in explaining many aspects of the mechanical behaviour of metals.

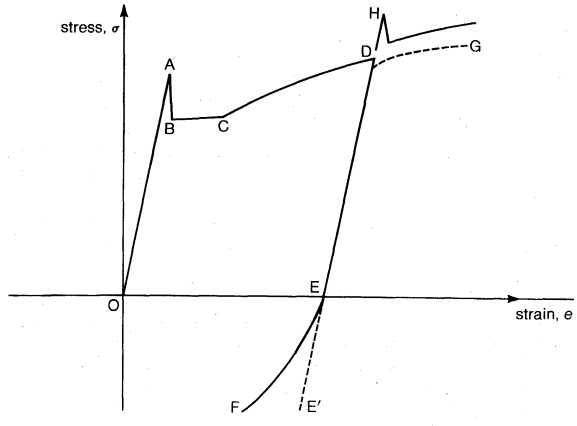


Figure 1: Behaviour of low-carbon steel during stretching in a testing machine (see text).

If, for example, a specimen of low-carbon steel is stretched in a testing machine, stress (σ) and strain (e) may be plotted as in Figure 1. The first application of any stress produces strain indicated by the straight line OA, its slope being equal to Young's modulus. An upper yield point A is reached and plastic extension, as strain is

increased, occurs along BC at a slightly lower stress; this reduction is associated with the spread of what are called Lüders bands, lines or markings on the formerly smooth surface, through the specimen. After C, strain-hardening occurs along CD. When stress is released at point D, recovery along DE is substantially elastic. Application of compressive stress at point E, however, is not found to give an elastic compression EE′, but instead a curved inelastic line EF is followed. This is called the Bauschinger effect, and it is associated with an easy reverse flow of dislocations that have become piled up against obstacles such as locked groups of dislocations and grain boundaries. An immediate reapplication of tensile stress gives flow DG, as if loading had not been interrupted. If the specimen is rested for a few days, a new yield point H is encountered.

Bauschinger effect

Yield points in steel can be associated with the presence of interstitial atoms of carbon and nitrogen. These diffuse through the iron (ferrite) lattice, and by gathering around dislocations, prevent their movement until an upper yield stress is applied. The large numbers of free dislocations present in a plastically strained material are thus gradually immobilized. The effect is called Cottrell-Bilby locking, and it is accelerated by heating to about 200° C. This treatment restores elasticity and eliminates the Bauschinger effect.

When a piece of steel is subjected to continuous rapid cycles of stress that do not exceed the static yield stress, a thin hysteresis loop develops (a graph of the resistance to a cycle of applied force due to a former state of the system, most often plotted for the changes in the magnetic permeability of a substance; see MAGNETISM). If the hysteresis is attributed to a limited number of atoms breaking away, it can be understood why the loop is widened by overstraining that provides more free dislocations, and why it is reduced in width by heat treatment.

It is well known that a bar of steel emits a ringing sound when it is struck, but a bar of cast iron does not because of the higher internal damping (resistance to oscillations) of cast iron. As an idealization for use in studying inelastic response of materials, a constitutive equation has been proposed that includes rates of stressing and straining with respect to time. Constitutive equations give the relationship between deformation and force (see DEFORMATION AND FLOW; also see below *Mathematical considerations*, equation [1]).

Constitutional equations

Three-dimensional strain and stress. The general properties of strain are concerned with the shape of an object, whereas those of stress derive from equilibrium of forces; therefore, these properties are independent of whether or not the body is elastic. If the position of a particle in the unstrained body is described by coordinate values x_1, x_2, and x_3, and if the amount of distortion is small, the position of this particle can be described with sufficient accuracy by the same coordinate values, even though small displacements have occurred. Then, if these displacements are u_1 in the x_1 direction, u_2 in the x_2 direction, and u_3 in the x_3 direction, normal strain will be equal to fractional elongation. Taking an element of length dx_1 lying along the x_1 axis of the coordinate system in the body, the displacement of the outer end will exceed that of the inner end by the partial derivative of the displacement in the x_1 direction with respect to x_1 multiplied by dx_1: $(\partial u_1/\partial x_1)dx_1$.

This is the rate of change of u_1 with respect to x_1, multiplied by the incremental distance. Hence the component of strain becomes $e_{11} = \partial u_1/\partial x_1$, in which e_{11} represents normal strain along the x_1 coordinate. The other normal strains follow by simply replacing the coordinate x_1 with either coordinate x_2 or x_3, and replacing the displacement u_1 with either displacement u_2 or u_3.

A rectangular element, Figure 2, may be considered next, having sides initially parallel to the x_1 and x_2 axes. Shear strain in the theory of elasticity, according to the mathematical definition, is equal to half the decrease in angle of the corner that is outermost with respect to the positive coordinate axes. (The angle A becomes B under the stress.) Small changes in inclination of the outer sides can be written in terms of displacement derivatives (Figure 2), leading to an expression for shear strain in which e_{12} represents strain in two dimensions:

$$e_{12} = \frac{1}{2}\left(\frac{\partial u_2}{\partial x_1} + \frac{\partial u_1}{\partial x_2}\right).$$

Strain-displacement equations

Two further components are obtained by replacing symbols x_1,x_2 with x_2,x_3 or x_3,x_1 and by replacing symbols u_1,u_2 with u_2,u_3 or u_3,u_1. These equations of strain-displacement give six components of strain in terms of three displacement components, and thus continuity of displacements in a solid body imposes compatibility conditions on any permissible strain distribution.

A stress component in the theory of elasticity is specified

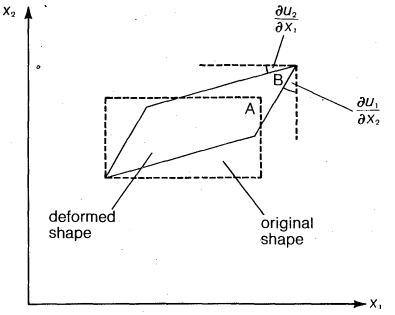


Figure 2: Calculation of shear strain in two dimensions, using angular changes (see text).

by means of two subscripts. Normal stress, σ_{11}, means the stress having a line of action parallel to the x_1 axis and acting on an element of area perpendicular to the x_1 direction. It is a positive value when tensile stress is being measured. Shear stress, σ_{12}, means a stress acting along a line parallel to the x_1 direction on an element of area perpendicular to the x_2 direction. In the x_1,x_2 plane this stress is positive if it acts in the positive x_1 direction on the side of an element situated at the larger positive x_2 value, or if it acts in the negative x_1 direction on the opposite side of the element. Equilibrium of moments (*e.g.*, forces that can rotate a body) of shear stresses acting on the edges of rectangular elements leads to equivalences between the following shear stresses: $\sigma_{12} = \sigma_{21}$, $\sigma_{32} = \sigma_{23}$ and $\sigma_{13} = \sigma_{31}$ or in general $\sigma ij = \sigma ji$ for $ij = 1,2,3$. Hence, there are three independent shear stresses that, with the three normal stresses, make a total of six components acting at any point. These components can be transformed so as to give the components referred to any alternative set of axes. The particular set of three axes on which shear stresses are zero are known as the principal axes.

Stresses acting on any element must be in equilibrium. If stresses, σ_{11}, acting on opposite sides of the element are equal, they will exert no resultant force. It is assumed that this normal stress may increase across the element so as to have an excess value $(\partial\sigma_{11}/\partial x_1)\delta x_1$ on the outside face (Figure 3), which has an area represented by the product

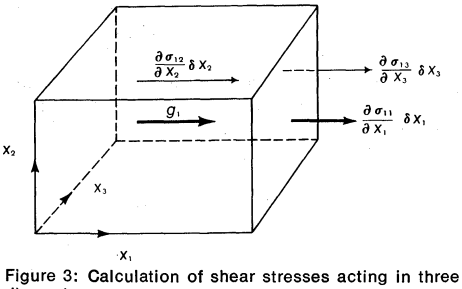


Figure 3: Calculation of shear stresses acting in three dimensions.

of the two deltas, representing the elements of length, δx_2 δx_3. Similar increases are permitted in shear stresses acting in the x_1 direction, and these excess values of stress are shown in Figure 3. Also, a body-force vector such as might be produced by gravity or inertia, is supposed to have component g_1 acting in the x_1 direction, this being a force per unit volume. By equating to zero the sum of

all net forces acting in the x_1 direction, and removing the common factor equal to the volume of the element, an equation is obtained that may be written:

Stress equilibrium equations

$$0 = g_1 + \frac{\partial \sigma_{11}}{\partial x_1} + \frac{\partial \sigma_{12}}{\partial x_2} + \frac{\partial \sigma_{13}}{\partial x_3}.$$

Two more stress equilibrium equations representing equilibrium in the X_2 and X_3 directions can be obtained by interchange of coordinate values.

These equations define the restrictions to permissible distribution of stress, and they form the starting point in the analysis of stress in any deformable medium.

Linear elasticity and elastic constants. Arguments in favour of the existence of two independent elastic moduli were clarified by Sir George Gabriel Stokes, British physicist and mathematician, in 1845. He maintained that a change in volume and a change in shape are physically distinct processes. A bulk modulus K may be defined by the statement that average stress, σ, produces a fractional change in volume. If e is the average linear strain, the law of volume elasticity is represented by an equation in which average stress—that is, the average of the three stress components—equals three times the product of the bulk modulus and average linear strain: $\sigma = 3Ke$.

Isotropic material

In an isotropic material—that is, a material having equal properties in all directions—a single constant can be used for relating reduced stresses and strains because this constant is a measure of rigidity, or of resistance to change of shape. General components of stress and the linear relation can be expressed in an equation that contains the shear modulus. The proportionality between stresses and small strains is calculated without reference to the property of springing back on removal of stress, though this property would need to be assumed if one wished to predict the final strain produced by a succession of stresses having dissimilar signs.

It should also be noted that, although a structure may consist of parts made from linear elastic material, there may be reasons due to its shape why the structure gives a nonlinear variation of deflection with applied load. Examples are struts loaded near the point of elastic instability and thin plates having large transverse deflections imposed (for calculations, see below *Mathematical considerations*, Equation 2).

It is found that Young's modulus for an alloy is not much affected by mechanical and thermal treatments, and it is given approximately by the (volumetric) average of values (of the modulus) for each of the constituent metals. For example, all steels with a content of alloying elements less than about 10 percent can be assumed for most purposes to have the same elastic constants as mild steel. Progressive cold working of metals, however, may lead to varying orientations of the crystallographic axes of individual grains, with development of elastic anisotropy; that is, of a variation in modulus depending on the direction of the specimen axis relative to the direction of cold-working.

Molecular packing

The bulk modulus increases with pressure because of a decrease in the space between molecules. Water at 20° C, for example, has a bulk modulus value $K = 320{,}000$ pounds per square inch (psi). According to experimental results this should be multiplied by a factor containing pressure. (The factor is $[1 + 20 \times 10^{-6}p]$, in which p is pressure in psi). The increase in K with pressure is in part caused by a decrease in size of the molecules themselves. Because this decrease is relatively small, it might be expected that solids having closely packed molecules would show only a small variation of K with pressure; experiment proves this to be true. For similar reasons, the increase in shear modulus G is small for most metals. (The correction factor here is approximately $[1 + 0.15 \times 10^{-6}p]$). The elastic moduli of metals usually decrease with temperature due to the atoms moving farther apart, and there is a rapid fall at higher temperatures. This change in modulus may be nonlinear and sometimes irregular, but for many metals a decrease of two to six percent per 100° C may be expected over moderate ranges of temperature on either side of room temperature (for a treatment of invariants, energy, and yield criteria, see below, *Mathematical considerations*).

Mathematical theory of elasticity. It has been seen that three sets of equations must be satisfied at all interior points in a linear-elastic body; namely, the strain-displacement equations, the stress-equilibrium equations, and the stress-strain relations. A typical problem posed in the theory of elasticity is to find stresses and displacements at all interior points, given the external shape of the body and the stresses or displacements imposed over the boundary of the body or part of the boundary. Occasionally the basic equations can be used as they stand, by inserting a simple system of stresses or displacements into the first order differential equations.

Various idealized stress systems can be formulated, such as that appropriate for symmetry on planes, or about an axis, or for flexure of plates. The procedure is to look for solutions of the differential equations that can be made to fit the specified boundary conditions. Rectangular coordinates cannot be conveniently used for describing stresses and displacements on curved boundaries. In such problems, the differential equations need to be recast in curvilinear coordinates, which may call for many specialized techniques of applied mathematics.

It is often possible to verify theoretical solutions by means of experimental methods, for example, by using photoelastic models or electrical resistance strain gauges. The extensive range of techniques in modern experimental stress analysis can also be used in situations where no theoretical methods can be applied; for example, in the measurement of internal stresses set up in components during manufacturing processes.

Mathematical considerations. Equations (1) and (2), referred to earlier, as well as invariant functions, strain energy, and yield criteria, are treated below.

Equation (1): The constitutive equation of an anelastic solid is expressed

$$\sigma + t_1 \frac{d\sigma}{dt} = E\left(e + t_2 \frac{de}{dt}\right)$$

in which t_1 and t_2 are time constants. For very slow straining, the rate terms can be dropped, leaving the static elasticity relation $\sigma = Ee$. If straining is extremely rapid so that the rate terms predominate, however, the apparent modulus is given by $E' = Et_2/t_1$. Response to periodic stressing at frequency $\omega/2\pi$ can be found by inserting $\sigma = \sigma_0 \sin \omega t$ and $e = e_0 \sin(\omega t - \phi)$. The angle of lag ϕ is then obtained as

$$\tan \phi = \omega(t_2 - t_1)/(1 + \omega^2 t_1 t_2).$$

This predicts a maximum angle of lag at imposed frequency $\omega = 1/\sqrt{t_1 t_2}$, and it is found that many materials do exhibit such a maximum value.

Equation (2): It is usually convenient to use total stresses and strains rather than reduced values. New elastic constants, ν, called Poisson's ratio, and E may be introduced to obtain equations relating normal stress and strain of the type

$$e_{11} = \frac{1}{E}[\sigma_{11} - \nu(\sigma_{22} + \sigma_{33})],$$

provided the new constants are related to the previous ones by

$$E = \frac{9KG}{3K + G} \text{ and } \nu = \frac{3K - 2G}{2(3K + G)}.$$

These expressions lead to the useful expressions for Young's modulus $E = 2(1 + \nu)G = 3(1 - 2\nu)K$. To illustrate the significance of Poisson's ratio ν, the axis of a tensile specimen is made to coincide with the z direction, and the longitudinal stress is set equal to $\sigma_{zz} = T$, with transverse stresses $\sigma_{xx} = \sigma_{yy} = 0$. The strains are then $e_{zz} = T/E$ and $e_{xx} = e_{yy} = -\nu T/E$. This means that the fractional transverse contraction is ν times the fractional longitudinal extension.

An invariant function of components of stress or strain

Invariants, energy, and yield criteria

at a point is a function having the same numerical value for any choice of coordinate axes. There are three independent invariants of stress. The first is average stress σ, defined as one third of the sum of the three normal components of stress. The second is conveniently written in terms of reduced stresses in which the reduced values of stress are distinguished by a prime, $\sigma'_{11} = \sigma_{11} - \sigma$, $\sigma'_{12} = \sigma_{12}$, and similarly for the other components;

$$2\bar{\sigma}^2 = \sigma'_{11}{}^2 + \sigma'_{22}{}^2 + \sigma'_{33}{}^2 + 2\sigma_{12}{}^2 + 2\sigma_{23}{}^2 + 2\sigma_{31}{}^2.$$

It is seen that when only a single shear stress σ_{12} is present, the other components being zero, then $\bar{\sigma} = \sigma_{12}$, so that this invariant is sometimes called the equivalent shear stress. The third invariant is essentially a third-order combination of stress components that may be shown to have a one-to-one relationship with Lode's variable, $\mu = 3\sigma'_2/(\sigma_1' - \sigma_3')$. These reduced principal stresses are designated in such a way that $\sigma'_1 > \sigma'_2 > \sigma'_3$. This variable, as defined here, can have values ranging from minus one for uniaxial tensile stress to plus one for compressive stress.

Strain energy U per unit volume is essentially the summation of products of internal forces and displacements, and for any linear material this can be written

$$U = \frac{1}{2}(\sigma_{11}e_{11} + \sigma_{22}e_{22} + \sigma_{33}e_{33} + 2\sigma_{12}e_{12} + 2\sigma_{23}e_{23} + 2\sigma_{31}e_{31}).$$

These stresses and strains may be separated into isotropic and reduced parts so that the basic stress-strain relations for an isotropic solid can be used. As might be expected, the result can be expressed in terms of stress invariants, $U = \frac{1}{2}(\sigma^2/K + \bar{\sigma}^2/G)$. Alternatively, total stresses may be used together with elastic constants E and ν to obtain the strain energy per unit volume,

$$U = \frac{1}{2E}[\sigma_{11}{}^2 + \sigma_{22}{}^2 + \sigma_{33}{}^2 - 2\nu(\sigma_{11}\sigma_{22} + \sigma_{22}\sigma_{33} + \sigma_{33}\sigma_{11}) + 2(1 + \nu)(\sigma_{12}{}^2 + \sigma_{23}{}^2 + \sigma_{31}{}^2)].$$

A similar expression can be obtained, if required, in terms of strain components alone.

The stress at which plastic flow commences in crystalline materials is not affected by superposition of moderate values of average stress. This fact has been demonstrated by carrying out tensile and torsion tests with specimens enclosed within a pressure vessel and subjected to high fluid pressures. Hence, the criterion of plastic yielding in an isotropic solid can be written in terms of the remaining stress invariants, in the form $\bar{\sigma}(\mu) = k$, in which k is some fixed numerical value. This function of μ is called a yield locus and can be represented in various geometrical ways. The simplest kind of yield locus is when $\bar{\sigma} = k$, which is known as the Mises criterion. An alternative, the Tresca criterion, states that yielding occurs when the maximum shear stress acting within the material reaches a certain value, and as this shear stress is equal to half the difference between the greatest and least of the principal stresses, this criterion becomes $(\sigma_2 - \sigma_3) = 2k$.

BIBLIOGRAPHY. Apart from isolated chapters in many physics textbooks, most works dealing with elasticity are for advanced studies. The standard texts include: S.P. TIMOSHENKO, *History of the Strength of Materials* (1953); A.H. COTTRELL, *The Mechanical Properties of Matter* (1964); W.D. BIGGS, *The Mechanical Behaviour of Engineering Materials* (1965); J.M. LESSELLS, *Strength and Resistance of Metals* (1954); A.M. FREUDENTHAL, *Introduction to the Mechanics of Solids* (1966); I.S. SOKOLNIKOFF and R.D. SPECHT, *Theory of Elasticity* (1946); and N.I. MUSKHELISHVILI, *Some Basic Problems of the Theory of Elasticity* (1953; orig. pub. in Russian, 1935).

(D.S.D.)

Elbe River

The Elbe, one of Europe's major waterways, runs from Czechoslovakia to the North Sea. It is formed at Spindleruv Mlyn by the confluence of numerous headwater streams originating in the Riesengebirge (Giant Mountains) on the border of Czechoslovakia and Poland. Starting southward, the river makes a wide arc to the west as it flows across Bohemia (where it is known as the Labe). It cuts northwest through the Erzebirge (Ore Mountains) in a gorge four miles long and enters Germany. From Dresden it flows northwest to Magdeburg and then north to Werben, where it again turns northwestward. Between Wittenberg and Lauenburg the boundary between the Federal Republic of Germany and the German Democratic Republic lies in midstream. Above Hamburg the Elbe splits into several branches; these rejoin farther downstream and the river now broadens from a width of 0.3 mile at Hamburg to a width of 9 miles at Cuxhaven, where it flows into the North Sea. Although Hamburg lies 68 miles upstream from the mouth of the river, the width of the estuary has made that city the largest seaport on the continent of Europe.

The total length of the Elbe is 724 miles (1,165 kilometres), of which 352 miles are in the German Democratic Republic. Its total drainage area is 55,620 square miles. Major tributaries are the Vltava (Moldau) and the Ohře (Eger), the Mulde and the Saale, all of which join it from the left; and the Iser, Black Elster, Havel, and Alster, from the right.

The flow of water in the Elbe varies considerably with the amount of precipitation and thawing in its drainage basin. At Dresden, the discharge rate averaged 14,609 cubic feet per second in the period 1937–47, but the rate varied from a low of 9,387 cubic feet in one year to 25,054 in another. At Neu-Darchau, about 140 miles above the mouth, the discharge rate was 24,700 cubic feet per second in the period 1926–65, with extremes of 5,117 and 127,741. These great variations sometimes hinder navigation. While there are dams on the Saale River in the Thüringer Wald (Thuringia Forest), these are not sufficient to control the water level of the Elbe.

The lower course of the Elbe is tidal as far as Altengamme, above Hamburg, where the river flow periodically reverses its direction. The average tide at Hamburg is 8.2 feet, and during storms the water may rise much higher, occasionally even flooding parts of the city.

Traffic on the Elbe

By means of the Elbe and its connecting waterways, vessels from Hamburg can navigate to Berlin, central Germany, and Czechoslovakia. The Mittelland Kanal, a short distance below Magdeburg, runs westward about 202 miles to the Dortmund–Ems Kanal, carrying barges of up to 1,000 tons to the industrial cities of Osnabrück, Hannover, Salzgitter, Hildesheim, Peine, and Wolfsburg, and connecting with the Weser and Rhine rivers. The Havel Kanal carries traffic from Magdeburg eastward to the network of waterways around Berlin and farther on to Poland. The Kiel Canal runs from the mouth of the Elbe to the Baltic Sea, and the Elbe–Lübeck Kanal, starting at Lauenburg, also runs to the Baltic, following an older (14th-century) canal. Another canal connects the lower Elbe with Bremerhaven on the Weser River. In the early 1970s the Nord–Süd Canal was under construction, intended to connect the Elbe a short distance upstream from Hamburg with the Mittelland Kanal at Fallersleben, thus linking the harbour of Hamburg directly with the industrial district of Salzgitter. The Elbe itself is navigable for 700-ton barges as far as Usti nad Labem in Bohemia, and for smaller barges as far as Prague through the tributary of the Vltava (Moldau). In East Germany it serves the important river ports of Magdeburg, Schönbeck, Aken, Dessau, Riesa, and Dresden, carrying bituminous coal, lignite, coke, metal, potash, grain, and piece goods. A new six-line railway tunnel under the Elbe at Hamburg, under construction in the early 1970s, was likely to have a considerable effect on trans-European traffic flows.

Historical importance

The basin of the Elbe has been settled since prehistoric times. Until the Middle Ages the river was the western boundary of the area inhabited by the northern Slavs. In the 12th century the Germans began to colonize the lands east of the Elbe and along the Baltic Sea. In World War II a point on the Elbe, near Torgau, was the meeting place of the United States and Soviet armies. Later, the river formed part of the demarcation between the German Democratic Republic and the Federal Republic of Germany.

The city of Hamburg dates from about 810. Together with Lübeck, Hamburg established the Hanseatic League in 1241. Today it is West Germany's largest city, excluding West Berlin, with a population (in 1970) of about 1,794,000. Another ancient city on the Elbe is Magdeburg, which in the early 9th century was a trading post on the border between the Germans and the Slavs. In the 13th century it was a flourishing commercial city and an important member of the Hanseatic League. Today it is the largest inland harbour of the German Democratic Republic. The other chief city of the Elbe is Dresden, founded about 1200. During the 18th century Dresden developed into a great centre of the fine arts, known as "Florence on the Elbe." Much of its beautiful architecture was destroyed during World War II. (H.F.)

Elburz Mountains

The Elburz (alternate spellings are Elburs, Alburz, or Albourz) Mountains in northern Iran separate the lowlands along the Caspian coast from the inland plateau. In its widest sense, Elburz designates the whole chain of ranges, some 560 miles (900 kilometres) long, which extends in a crescent-shaped arc from the Soviet frontier southwest of the Caspian Sea to the Khorāsān *ostān* (province) southeast of the sea, where it merges into Ālādāgh, the more southerly of the two principal ranges there. More commonly, however, the mountains northwest of the Safīd Rūd gorge, which lies to the southwest of the sea, are called the Talish (Talysh, Talesh, or Tavālesh) Range, or the Bogrov Dāgh; the Elburz Range, in its strictest sense, forms part of the central stretch of the chain, which also includes Iran's two highest peaks, Mt. Damāvand and Alām Kūh. The Elburz system traverses four of Iran's *ostāns* (provinces)—Āzarbāījān-e Khavari (East Azerbaijan), Gīlān, Central (Tehrān), and Māzandarān.

Geology. The Elburz chain is not as truly alpine (*i.e.*, resembling the European Alps) in its structure as is often suggested. On the one hand, continental conditions regarding sedimentation are reflected by thick Devonian sandstones (formed 395,000,000 to 345,000,000 years ago), called Old Red Sandstone in Britain, and by shales containing coal seams; on the other hand, marine conditions are reflected by Permo-Carboniferous (*i.e.*, younger Paleozoic) strata that are composed mainly of limestones, as well as by very thick beds of green volcanic rocks known as tuffs and lavas. Orogenetic (mountain-building) phases of importance date from the younger Tertiary (Mio-Pliocene) Epoch (between 26,000,000 and 2,500,000 years ago). Over large areas they produced only a loose folding; but in the Central Elburz a number of folds were formed into blocks thrust mainly southward but in places northward, with cores made of Paleozoic rocks. There is a distinct zonation of the whole area according to which rock formations predominate in forming mountain crests. Structurally and topographically, the Elburz system is less clearly defined on the southern than on the Caspian side of the chain, since various offbranching elements interconnect it on the southern side with the adjoining Iranian Plateau.

The Western, Central, and Shāhkūh ranges

Relief. The Western Elburz Range runs south-southeastward for 125 miles. Varying in width from 15 to 20 miles, it consists of a single asymmetric ridge, the long slope facing the Caspian. Few of its peaks approach or exceed 10,000 feet in height. There is a low pass west of Āstārā, near the Soviet frontier, 5,000 feet above sea level. The Safīd Rūd, formed by the junction of the Qezel Owzan (Qisil Uzun) and Shāhrūd rivers, is the only river to cross the whole width of the chain: its gorge, giving access to the low pass of Qazvīn, offers the best passage, although by no means an easy one, between Gīlān on the shores of the Caspian and the inland plateau.

The Central Elburz is 250 miles long. East of the longitude of Teheran (*q.v.*), which lies to the south of the range, it reaches a width of 75 miles. Longitudinal valleys and ridges follow the structural pattern; some of them are important centres of settlement, with towns such as Deylamān, Razan, Kojūr, and Namar located on the Caspian side, and Shāhrūd, Lār, Damāvand, and Fīrūzkūh on the southern side. There are likewise many gorges, by which the rivers find their way down one or another of the slopes. Only two passes allow a relatively easy crossing in a single ascent—these are the Kandevān Pass, between the Karaj and the Chālūs rivers, and the Gadūk Pass, between the Hableh and the Tālā rivers. The main divide runs generally south of the highest crest, which—with the exception of the towering and isolated cone of the extinct volcano Damāvand (18,386 feet [5,604 metres])—culminates in the glaciated massif (mountainous mass) of Takht-e Sulaimān, with Alām Kūh (15,827 feet, [4,824 metres]) at its summit.

The Eastern, or Shāhkūh, Elburz runs about 185 miles in a northeasterly direction. Since two ranges branch off on its southern side and no compensatory elements appear on the northern side, its width dwindles to less than 30 miles. With the exception of the Shāhkūh Range proper (which reaches an altitude of 12,359 feet, [3,767 metres]), the chain decreases in height toward the east. Longitudinal valleys are found less and less frequently east of the Shāhkūh. There are several passes at low altitudes.

Climate, soils, and vegetation. The Caspian and the inland, or southern, slopes of the Elburz differ markedly from each other in climatic and vegetational aspects.

Characteristics of the northern and southern slopes

The Caspian slope has a distinctly humid climate, thanks to northerly air movements, enriched with moisture from the sea, which collide with the steep faces of the mountains to cause precipitation. This precipitation amounts to more than 40 inches annually in the lowlands of Gīlān and is even more plentiful at higher altitudes. Although it decreases toward the east, it still suffices to nourish a humid forest for the whole length of the chain on the Caspian side, where the soils are mostly of the brown-forest type.

The natural vegetation of this slope grows in distinct zones: the luxuriant Hyrcanian forest on the lowest levels; a beech forest in the middle zone; and a magnificent oak forest from the altitude of 5,500 feet up to the levels where gaps in the divide allow the moist air to overflow into the inland basins. Above the cloudy zone, which expands and recedes depending on conditions, a steady current of dry air from the south or southwest causes the vegetation of the inland slope to overlap the divide on all the higher ranges; this circumstance changes in winter and in summer, when frequent large-scale cyclonic wind and pressure systems from the Mediterranean make their way eastward over the Iranian Plateau. In some sheltered valleys there are extensive stands of wild cypress; sheltered valleys adjacent to the Safīd Rūd constitute the only olive-growing areas of note in Iran.

The southern slope of the Elburz, by direct contrast, shares the arid character of the Iranian Plateau. Annual precipitation varies between 11 and 20 inches and is very irregular. The soils are of the type associated with steppe (treeless, grassy, or shrubby type of vegetation) or, less widely, are of the chestnut variety. The slope has become even more steppe-like ever since the almost complete destruction of its original dry forest of junipers.

Animal life. The Hyrcanian tigers for which the Caspian forests were famous are now very rare; but other wild cats, such as the leopard and the lynx, are still numerous in the Elburz. The bear, the wild boar, red and roe deer, the moufflon (wild mountain sheep), and the ibex are also present. Eagles and pheasants are notable among birds.

Demography. The basic elements in the population of the two *ostāns* bordering the Caspian—Gīlān and Māzandarān—are considered to possess distinguishing characteristics; they speak Iranian dialects of their own. Their tendencies to separatism, however, have caused successive governments to implant tribal groups of Turkic, Kurdish, and even Arab origin in these provinces. A gradual migration from the highland into the humid area, where manpower has always been short—not least because of the high mortality due to malaria—was also

noticeable. The total population of the Elburz chain, estimated at about 1,000,000 in 1966, is very unevenly distributed. While large areas are almost empty—some being occupied only by nomads and others having been depleted by Turkmen raids in the 19th century—there are still several well-settled districts, including Deylamān, Alāmut, Tālaqān, and Lārījān (at the foot of Mt. Damāvand), all of which have various historical associations.

The Caspian and the inland slopes have different characteristic types of settlement; the landscape of the Caspian slopes is characterized by forest clearings with shingle-roofed loghouse villages and by lush fields and pastures. The landscape of the inland slopes is of the oasis type. *Qeshlāqs*, or winter villages, are found both on the inland foothills and in the low valley bottoms (terraced for rice growing) of the Caspian side. *Yeylāqs*, or summer villages, mostly half deserted in winter, are spread over a broad zone higher up. Extensive grain cultivation occurs on both slopes, and cattle raising occurs on the Caspian side. Alpine pastures, seasonally dotted with thousands of flocks, with the black tents of the semi-nomads, and with the huts or white tents of the village shepherds, cover an extensive zone yet higher. The land-distribution pattern prevailing in the Elburz includes a high proportion of peasant ownership. The holdings often are much fragmented.

Summer and winter villages

Many of the traditional ways of livelihood of the mountaineers, including charcoal burning (now prohibited because of devastation of the forests), the transportation of goods (especially of rice and of charcoal for Tehrān) by pack animals, and the working of hundreds of small coal mines, have been displaced by the 20th-century modernization of Iran. Modernization likewise, together with the lure of the nearby capital city of Tehrān, has not only set a new course for seasonal emigration but also has tended to turn that seasonal movement into a definite exodus from the countryside.

Transportation. Apart from the main line of the trans-Iranian railroad, which links Tehrān with Bandar-e Shāh via the Gadūk Pass, there are six asphalted roads across parts of the Elburz. From west to east, these run between Ardebīl and Āstārā, between Qazvin and Rasht, between Tehrān and Chālūs, between Tehrān and Āmol (via Damāvand); between Tehrān and Bābol (via Fīrūzkūh), and between Shāhrūd and Gorgān (via Kotal-e Zardāneh Pass).

Economic resources. The wild (natural or original) forests of the Elburz Mountains cover more than 8,000,000 acres, of which some 3,000,000 can be exploited commercially for timber and other wood. There are also a few modern coal mines, which yield about 250,000 tons of coal a year, as well as some deposits of iron and other ores. But most important is the water of the rivers, which is used for irrigation, for generating hydroelectric energy, and for supplying the fast-growing Tehrān. Spectacular dams have been built. These include the Safīd Rūd Dam, used for the irrigation of the Safīd Rūd Delta; the Karaj Dam and the Jajrūd Dam, used mainly for supplying water to Tehrān and partly for irrigation; and a series of dams on other rivers of the Māzandarān *ostān* also used for irrigation.

In the early 1970s, further dams were either under construction or projected, and it was anticipated that the establishment of such facilities in the mountains as skiing, health, and tourist resorts could provide another economic resource in the future.

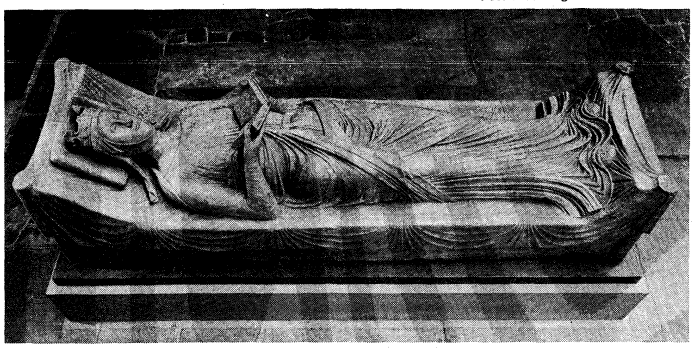
BIBLIOGRAPHY. W.B. FISHER (ed.), *The Land of Iran*, vol. 1 of *The Cambridge History of Iran* (1968), with bibliography; H. BOBEK, *Iran: Probleme eines unterentwickelten Landes alter Kultur*, 3rd ed. (1967); A. GANSSER and H. HUBER, "Geological Observations in the Central Elburz, Iran," *Schweiz. Miner. Petrogr. Mitt.*, 42:583–630 (1962); J. STOCKLIN, "Structural History and Tectonics of Iran: A Review," *Bull. Am. Ass. Petrol. Geol.*, 52:1229–1258 (1968), with bibliography.

(H.B.)

Eleanor of Aquitaine

Queen consort of both Louis VII of France and Henry II of England, and mother of Richard I the Lion-Heart and John of England, Eleanor of Aquitaine was probably the most powerful woman of her era. Her political daring made her the equal of most contemporary sovereigns, and, as patroness of the arts, she provided at the court of Poitiers a congenial environment for the burgeoning literary culture of 12th-century Europe.

Foto Marburg



Eleanor of Aquitaine, plaster cast of her tomb effigy, 13th century, in the abbey church at Fontevrault-l'Abbaye, France.

Eleanor, born about 1122, was the daughter and heiress of William X, duke of Aquitaine and count of Poitiers, who possessed one of the largest domains in France—larger, in fact, than those held by the French king. Upon William's death in 1137 she inherited the duchy of Aquitaine and, in July 1137, married the heir to the French throne, who succeeded his father Louis VI the following month. Eleanor became queen of France, a title she held for the next 15 years. Beautiful, capricious, and adored by Louis, Eleanor exerted considerable influence over him, often goading him into undertaking perilous ventures.

Heiress of Aquitaine

From 1147 to 1149 Eleanor accompanied Louis on the Second Crusade to protect the fragile Latin kingdom of Jerusalem, founded after the First Crusade only 50 years before, from Turkish assault. Eleanor's conduct during this expedition, especially at the court of her uncle Raymond of Poitiers at Antioch, aroused Louis's jealousy and marked the beginning of their estrangement. After their return to France and a short-lived reconciliation, their marriage was annulled in March 1152. According to feudal customs, Eleanor then regained possession of Aquitaine, and two months later she married the grandson of Henry I of England, Henry Plantagenet, count of Anjou and duke of Normandy. In 1154 he became, as Henry II, king of England, with the result that England, Normandy, and the west of France were united under his rule. Eleanor had only two daughters by Louis VII; to her new husband she bore five sons and three daughters. The sons were William, who died at the age of three; Henry; Richard, the Lion-Heart; Geoffrey, duke of Brittany; and John, surnamed Lackland until, having outlived all his brothers, he inherited, in 1199, the crown of England. The daughters were Matilda, who married Henry the Lion, duke of Saxony and Bavaria; Eleanor, who married Alfonso VIII, king of Castile; and Joan, who married successively William II, king of Sicily, and Raymond VI, count of Toulouse. Eleanor would well have deserved to be named "grandmother of Europe."

During her childbearing years, she participated actively in the administration of the realm and even more actively in the management of her own domains. She was instrumental in turning the court of Poitiers, then frequented by the most famous troubadours of the time, into a centre of poetry and a model of courtly life and manners. She was the great patron of the two dominant poetic movements of the time: the courtly love tradition, conveyed in the romantic songs of the troubadours, and the historical matière de Bretagne, or "legends of Britanny," which originated in Celtic traditions and in the *Historia regum Britanniae*, written by the chronicler Geoffrey of Monmouth some time between 1135 and 1139.

Cultural activities

The revolt of her sons against her husband in 1173 put her cultural activities to a brutal end. Since Eleanor, 11 years her husband's senior, had long resented his infidelities, the revolt may have been instigated by her; in any case, she gave her sons considerable military support. The revolt failed and Eleanor was captured while seeking refuge in the kingdom of her first husband, Louis VII. Her imprisonment for nearly 15 years (1174–89) in England ended only with the death of Henry II. On her release, Eleanor played a greater political role than ever before. She actively prepared for Richard's coronation as king, was administrator of the realm during his crusade to the Holy Land, and, after his capture by the Duke of Austria on Richard's return from the east, collected his ransom and went in person to escort him to England. During Richard's absence, she succeeded in keeping his kingdom intact and in thwarting the intrigues of his brother John Lackland and Philip II Augustus, king of France, against him.

In 1199 Richard died without leaving an heir to the throne, and John was crowned king. Eleanor, nearly 80 years old, fearing the disintegration of the Plantagenet domain, crossed the Pyrenees in 1200 in order to fetch her granddaughter Blanche from the court of Castile and marry her to the son of the French king. By this marriage she hoped to insure peace between the Plantagenets of England and the Capetian kings of France. In the same year she helped to defend Anjou and Aquitaine against her grandson Arthur of Brittany, thus securing John's French possessions. In 1202 John was again in her debt for holding Mirebeau against Arthur, until John, coming to her relief, was able to take him prisoner. John's only victories on the Continent, therefore, were due to Eleanor.

Assessment

She died on April 1, 1204, in the monastery at Fontevrault, Anjou, where she had retired after the campaign at Mirebeau. Her contribution to England extended beyond her own lifetime; after the loss of Normandy (1204), it was her own ancestral lands and not the old Norman territories that remained loyal to England. She has been misjudged by many French historians who have noted only her youthful frivolity, ignoring the tenacity, political wisdom, and energy that characterized the years of her maturity. "She was beautiful and just, imposing and modest, humble and elegant"; and, as the nuns of Fontevrault wrote in their necrology: a queen "who surpassed almost all the queens of the world."

BIBLIOGRAPHY. AMY KELLY, *Eleanor of Aquitaine and the Four Kings* (1950), is a major work with complete notes and a good bibliography and sources. H.G. RICHARDSON, "The Letters and Charters of Eleanor of Aquitaine," *English Historical Review*, 74:193–213 (1959), adds some unpublished sources to those gathered by Amy Kelly. REGINE PERNOUD, *Aliénor d'Aquitaine* (1965), gives more attention to the personality and politics of Eleanor herself, independently from the history of her two husbands.

(R.Pe.)

Eleaticism

One of the principal schools of ancient Pre-Socratic philosophy was Eleaticism, so-called from its seat in the Greek colony of Elea (or Velia) in southern Italy. This school, which flourished in the 5th century BC, was distinguished by its radical monism; *i.e.*, its doctrine of the One, according to which all that exists (or is really true) is a static plenum of Being as such, and nothing exists that stands either in contrast or in contradiction to Being. Thus, all differentiation, motion, and change must be illusory. This monism is also reflected in its view that existence, thought, and expression coalesce into one.

The Eleatic school vis-à-vis rival movements. The Eleatic school was founded by Parmenides, a 5th-century Rationalist, and included two scholars who may have been his pupils, Zeno of Elea, renowned for his paradoxes of motion (see below), and Melissus, an admiral of Samos.

Each member of the Eleatic school espoused a distinctive variety of Eleaticism: Parmenides pursued a direct and logical course of thought and viewed Being as finite and timeless; Zeno concurred in Parmenides' doctrines but employed the indirect methods of *reductio ad absurdum* and infinite regress (see below); and Melissus modified the doctrines, viewing Being as infinitely extensive and eternally temporal. A fourth thinker, the Sicilian Sophist Gorgias of Leontini, though not an adherent to Eleaticism, employed the methods of the Eleatic Zeno to defend its opposite—a nihilism that affirmed Non-Being instead of Being.

Contrasts with other movements

Eleaticism represents a reaction against several tendencies of thought. Methodologically, it spurned the empirical (observational) approach taken by earlier cosmologists, such as the 6th-century Milesians Thales and Anaximenes, who discerned ultimate reality in water and in air (or breath), respectively, for these substances are materializations of Being—analogous to the materialization that occurs in Pythagoreanism in passing from an abstract line or plane or three-dimensional form to a solid perceptible body—rather than Being itself; or, at best (as some scholars hold), the substances are mythological representations of Being. The Eleatics, on the contrary, ignoring perceptual appearances, pursued a rationalistic—*i.e.*, a strictly abstract and logical—approach and thus found reality in the all-encompassing, static unity and fullness of Being and in this alone. Thus Parmenides was the father of pure ontology.

Ontologically—in its view of the nature of Being—the Eleatic school, holding to what Parmenides called the "way of truth" ("what is"), stood opposed to two other "ways of research" that were then current: first, to the "way of opinion" (or Seeming; later developed at length by Plato in the *Sophist*), which held that a being comprises or is defined not only by what it is but also by what it is not—*i.e.*, by its contrast with other things; and, second, to a way recognizable as that of Heracleitus, a caustic and often cryptic philosopher then living in Ephesus, who maintained—still more radically—that the essence of a being lies in part in its involvement in, or even its identity with, its opposite. Finally, as an aspect of Parmenides' opposition to the way of opinion, he was in reaction also against Anaximander, another Milesian scientist and philosopher. Though Anaximander's basic principle, the apeiron ("boundless"), was duly abstract and not a part of the world itself (as were water and air), his philosophy depended, nonetheless, upon the world's contrast with the infinite apeiron, from which all things come and to which they return "in accordance with the ordinance of Time." This contrast—which, in a Pythagorean version, envisioned the world as breathing in voidness from the infinite outer breath in order to keep things apart or discreet—thus spawned "a many" that contradicts the Eleatic One.

Archaeological and literary sources. The sources for the study of Eleaticism are both archaeological and literary. Archaeologists have ascertained that, at the time of Parmenides, Elea was a large town with many temples, a harbour, and a girdle of walls several miles long. They have also unearthed a site presumed to be that of the medical school that Parmenides founded and an inscription bearing Parmenides' name.

The literary sources consist of fragments preserved by later classical authors—19 from Parmenides, 4 from Zeno, and 10 from Melissus—of which all but 3 from Parmenides and 2 from Melissus are 10 lines or less in length. The interpretation is also affected by the weight of scholarly opinion that now holds (as against Karl Reinhardt, an early-20th-century philologist and historian of philosophy) that Parmenides reacted to Heracleitus and not vice versa. Moreover, the biases of the citing authors must be weighed. Heracleitans and Parmenideans of the second generation, for example, saw their masters, simplistically, only as the prophets of movement and immobility; and the ancient Skeptic Sextus Empiricus distorted Parmenides' thinking into problems of epistemology (theory of knowledge), because this is what his Skeptical eye saw in his writings.

The rigorous Ontologism of Parmenides and Melissus. For a long time Xenophanes of Colophon, a religious thinker and rhapsode of the 6th–5th century BC, was con-

sidered the founder of the Eleatic school and Parmenides' mentor. This ancient claim, however, has been successfully criticized by Reinhardt. It is even possible that, on the contrary, Xenophanes was an older pupil of Parmenides. In any case, his monistic view of a cosmic God, whom he may have equated pantheistically with Being itself, was Eleatic in its contention that God is one and ungenerated, that his seeing, thinking, and hearing are equally all-pervading (*i.e.*, he is not a composite), and that he "always remains in the same place, not moving at all."

Parmenides' poem and the two ways of research

Parmenides' poem *Peri physeōs* ("On Nature") is divided into three parts: (1) a proem (preface), in which his chariot ride through the heavens to the very seat of the goddess Alētheia (Truth) is described and their initial conversation is related, in which she announces that he is "to learn all things, both the unshaken heart of well-rounded truth and also what seems to mortals, in which is no true conviction"; (2) the "Way of Truth," the main part, in which the real and unique Being is depicted; and (3) the "Way of Opinion" (or Seeming), in which the empirical world—*i.e.*, the single things as they appear every day to every man—is presented.

Logical and linguistic approach. Thus, at the very heart of Parmenides' philosophy lies the distinction made by the goddess (in Fragment 2) between the two "ways of research." As noted earlier, the first is the antinomy (or paradox) of those who think and say that everything is Being and who shun all assertions of Non-Being; and the second is that of those who think and say that something is in a way and is not in another way—that a book is a book, for example, and not a table. There is, however, also a third way that is far more erroneous and fallacious than the second: that of Heracleitus, who acknowledged, just as Parmenides did, the ontological antinomy of is and is not but reversed it, holding that the real way of understanding things is to grasp their essential contradiction, their intrinsic opposition to everything else. In this view, one must say that to be a table is also not to be just a table and that to be a chair is not to be just a chair but to be also a table, because not only opposite things but also things that are merely different are bound to each other. Thus, life is death to Heracleitus, death is life, and justice would be meaningless if it had no injustice to defeat.

In essence, then, the possible ways are three: (1) that of renouncing all contradictions whatsoever (truth); (2) that of contradicting oneself relatively (seeming); and (3) that of contradicting oneself completely and absolutely (Heracleitus). And Eleaticism chose the first, the absolutely noncontradictory way that says that only what is, Being, is really true.

Coalescence of Being, thought, and expression

Not-Being, in fact, can neither be recognized nor expressed, for, as Parmenides then added, "for the same thing can be thought and can exist." And—if one may guess at the words (now lost) that probably followed—what-is-not you can neither know nor say; thus, to think is indeed the same as to say that what you think is. To this coalescence of existing reality and the intellectual grasping of it, Parmenides also added the linguistic communication of such knowledge. Each way of research, in fact, is at the same time a way of speculation and a way of diction; *i.e.*, both a way of searching for truth with one's mental eyes and of expressing it in words. The primal source of the Eleatic philosophy thus lies in the archaic sense of language, according to which one cannot pronounce "yes" and "no" without deciding about the reality or unreality of the objects of the statements. Thus, "yes," or "is," becomes the name of the truth; and "no," or "it is not," becomes that of its opposite.

This Eleatic principle may be illustrated by a passage from Aeschylus, a leading Greek dramatist, who, in his *Septem contra Thebas* (Eng. trans., *Seven Against Thebes*), judged it very appropriate that Helen would have destroyed Troy, because her name—naïvely derived from *helein* ("destroy") and *naus* ("ship")—marked her as a destroyer of ships. Here *nomen est omen:* the language is not merely a symbol; it corresponds to reality in its very structure. Thus, the Eleatic could not imagine a truth that is only expressible but not thinkable nor one that is only thinkable but not expressible.

Monistic theory of Being. From the premise of the essential coalescence of language and reality follows Parmenides' theory of Being, which comprises the heart of his philosophy. The only true reality is Eōn—pure, eternal, immutable, and indestructible Being, without any other qualification. Its characterizations can be only negative, expressions of exclusions, with no pretense of attributing some special quality to the reality of which one speaks.

Conflict with Melissus on the eternity of Being

In Fragment 8, verse 5, Parmenides said that the absolute Being "neither was nor will be, because it is in its wholeness now, and only now." Thus, its presence lasts untouched by any variation in time; for no one can find a genesis for it, either from another being (for it is itself already the totality of Being) or from a Not-Being (for this does not exist at all).

Obviously, this Parmenidean conception of the eternal presence of the Being conflicts with Melissus' idea of the perpetual continuation of the Being in the past, in the present, and in the future. Thus, if Eleaticism had been founded by Melissus, no one could have really understood its actual doctrine. One could suspect in it only an aspiration to have things capable of being really enduring. But even then the theory would hardly be understandable, because what one wants is not stable things in general; one wants good things to be firm and stable and bad things to be ephemeral. The perpetual continuity of existence as espoused by Melissus was despised by Parmenides just because "will be" and "has been" are not the same as "is." Only "is" is the word of the reality—just because it is the right name for the right thinking of the right Being.

Among the consequences of this Eleatic conception is the rejection of every change (birth, movement, growth, death) as pertaining only to the second-rate reality, which is known and expressed through the second "way of research." Thus, the true and noncontradictory reality is extraneous to all of those happenings, great or small, that make the constant stuff of all history.

Secondly, the real Being has no difference, no lack, no variety whatsoever in itself. Melissus is here the true pupil of Parmenides, who said that the Eōn is so closely connected in itself that "all Being is neighbour of all Being": for Melissus developed this theory by the negation of every form of *kenon* ("void"): the Being is an absolute plenum just because every lack in its plentifulness would amount to a presence of some Not-Being.

The paradoxes of Zeno. The position of the other great pupil of Parmenides, Zeno of Elea, was clearly stated in the first part of Plato's dialogue *Parmenides.* There Zeno himself accepted the definition of Socrates, according to which he did not really propose a philosophy different from that of Parmenides but only tried to support it by the demonstration that the difficulties resulting from the pluralistic presupposition of the *polla* (the multiple beings of man's daily experience) were far more severe than those that seemed to be produced by the Parmenidean reduction of all reality to the single and universal Being.

Paradoxes of plurality

The arguments by which Zeno upheld his master's theory of the unique real Being were aimed at discrediting the opposite beliefs in plurality and motion. There are several arguments against plurality. First, if things are really many, everything must be infinitely small and infinitely great—infinitely small because its least parts must be indivisible and therefore without extension and infinitely great because any part having extension, in order to be separated from any other part, needs the intervention of a third part; but this happens to such a third part, too, and so on ad infinitum.

Very similar is the second argument against plurality: if things are more than one, they must be numerically both finite and infinite—numerically finite because they are as many things as they are, neither more nor less and numerically infinite because, for any two things to be separate, the intervention of a third thing is necessary, ad infinitum. In other words, in order to be two, things

must be three, and, in order to be three, they must be five, and so on. The third argument says: if all-that-is is in space, then space itself must be in space, and so on ad infinitum. And the fourth argument says: if a bushel of corn emptied upon the floor makes a noise, each grain must likewise make a noise, but in fact this does not happen.

Paradoxes of motion

Zeno also developed four arguments against the reality of motion. These arguments may also be understood (probably more correctly) as proofs *per absurdum* of the inconsistency of any presupposed multiplicity of things, in so far as these things may be proved to be both in motion and not in motion. The first argument states that a body in motion can reach a given point only after having traversed half of the distance. But, before traversing half, it must traverse half of this half, and so on ad infinitum. Consequently, the goal can never be reached.

The second argument is that of "Achilles and the tortoise." If in a race the tortoise has a start on Achilles, Achilles can never reach the tortoise; for, while Achilles traverses the distance from his starting point to that of the tortoise, the tortoise will have gone a certain distance, and, while Achilles traverses this distance, the tortoise goes still further, ad infinitum. Consequently, Achilles may run indefinitely without overtaking the tortoise. This argument is fundamentally identical to the previous one, with the only difference being that here two bodies instead of one are moving.

The third argument is the strongest of them all. It says the following: so long as anything is in a space equal to itself, it is at rest. Now, an arrow is in a space equal to itself at every moment of its flight; therefore, even the flying arrow is at rest all of the time. And the final argument says: two bodies moving at equal speed traverse equal spaces in an equal time. But, when two bodies move at equal speed in opposite directions, one passes the other in half of the time that a moving body needs to pass a body that is at rest.

The difficulty with all these arguments is that of really understanding them in their historical frame, which neither Aristotle—who was mainly concerned to confute Zeno—nor many modern scholars—who are concerned with developing new theories for the calculation of infinitesimal quantities—really tried to do. Moreover, the role of the author of the paradoxes in the history of Greek philosophy is itself paradoxical, for many of the same arguments by which Zeno proved the self-contradictory nature of the unity considered as the smallest element of a pluralistic reality (the Many) were later similarly used by Gorgias and Plato to demolish the Parmenidean One-Totality itself.

The decline of Eleaticism. This problem is also connected with that of the correct interpretation of the second part of Plato's *Parmenides*. Here the great dialectical discussion to which Zeno submits the young Socrates is meant as a serious exemplification of the logical training that Socrates still needs if he wants to make progress in philosophy. But the result is simply comic—a "fatiguing joke," according to Plato's *Parmenides*—because Zeno always starts from the mere principles of the pure Being or the One and arrives at absurd conclusions: everything is shown to be true as well as false and deducible and not deducible from everything else.

Gorgias' nihilism and Plato's *Sophist*

Such dialectical futility had been anticipated by the nihilism of Gorgias, presented in a work ironically entitled *Peri tou mē ontos ē peri physeōs* ("On That Which Is Not or Nature"), in which he said (1) that nothing exists; (2) that if anything exists, it is incomprehensible; and (3) that if it is comprehensible, it is incommunicable —and in so doing he applied Parmenides' coalescence of Being and thought and expression to Not-Being instead of to Being and thus signalled the decline of Eleaticism.

The serious discussion and criticism of the Eleatic philosophy, however, and the positive interpretation of every Not-Being as a *heteron* (*i.e.*, as a being characterized only by its difference from "another" being) is neither in Gorgias nor in the *Parmenides* but in the *Sophist* of Plato. There Plato argued that the antinomy between *On* and *Mē-On* (Being and Not-Being) does not really exist, the only real antinomy being that of *Tauton* and *Heteron; i.e.*, only that of a single object of consciousness in its present determination and all other things from which it is distinguished.

The real story of ancient Eleaticism thus ends with Plato and with Democritus, who said that Being exists no more than Not-Being, the thing no more than the no-thing. But many thinkers, and great thinkers at that—from Aristotle to Kant and from Hegel to Marx—have continued to work or to fight with the antinomy of Being and Not-Being.

BIBLIOGRAPHY

Sources: The main edition of the fragments of the Eleatic, as of all the Pre-Socratic, philosophers is still HERMANN DIELS, *Die Fragmente der Vorsokratiker* (1903; 11th ed. rev. by WALTHER KRANZ, 3 vol., 1964), which is a critical edition of the Greek fragments with German translations; for English translations see KATHLEEN FREEMAN, *Ancilla to the Pre-Socratic Philosophers* (1948, reprinted 1962); PILO ALBERTELLI, *Gli Eleati: testimonianze e frammenti* (1939), is a good Italian translation with commentary.

General works: Much of the best material is in general works on ancient Greek or Pre-Socratic philosophy. The most comprehensive treatment is in EDUARD ZELLER, *Die Philosophie der Griechen*, 6–7th ed., 3 vol., ed. by W. NESTLE (1920–23); in the revised Italian trans. of Zeller, ed. with extensive additions by ROLDOLFO MONDOLFO, the portion on the Eleatic school by GIOVANNI REALE (1967) occupies the whole 3rd volume of the 1st section and contains a good selected bibliography; an English translation of Zeller is entitled *A History of Greek Philosophy, from the Earliest Period to the Time of Socrates*, 2 vol. (1881). See also JOHN BURNET, *Early Greek Philosophy*, 4th ed. (1930, reprinted 1963); GEOFFREY KIRK and JOHN E. RAVEN, *The Presocratic Philosophers*, pp. 263–306 (1957); W.K.C. GUTHRIE, *A History of Greek Philosophy*, vol. 2 (1965); and GUIDO CALOGERO, *Storia della logica antica*, vol. 1, pp. 109–208 (1967).

Special works: KARL REINHARDT, *Parmenides und die Geschichte der griechischen Philosophie* (1916); GUIDO CALOGERO, *Studi sull'Eleatismo* (1932; German trans., *Studien über den Eleatismus*, 1970); JOHN E. RAVEN, *Pythagoreans and Eleatics* (1948), for specialists; J.H.M.M. LOENEN, *Parmenides, Melissus, Gorgias: A Reinterpretation of Eleatic Philosophy* (1959).

Articles: KURT VON FRITZ, review (in German) of Calogero's *Studi* in *Gnomon*, 14:91–109 (1938); GREGORY VLASTOS, review (in English) of Jean Zafiropoulo's *L'École Éléate* in *Gnomon*, 25:166–169 (1953); G.E.L. OWEN, "Eleatic Questions," *Classical Quarterly*, 10:84–102 (1960).

(G.C./L.H.St.)

Electoral Processes

Elections are means of making political choices by voting. They are used in the selection of leaders and in the determination of issues. This conception of elections implies that the voters are presented with alternatives, that they can choose among a number of proposals designed to settle an issue of public concern. The presence of alternatives is a necessary condition, for although electoral forms may be employed to demonstrate popular support for incumbent leaders and their policies, the absence of alternatives disqualifies such devices as genuine elections.

Background

ORIGINS IN ANTIQUITY

The origins of election as a means of choosing among alternatives is shrouded in mystery. Literary, rather than genuinely historical evidence from the Greece of Homer and the Germany of Tacitus suggests that acclamation of the victorious leader in tribal wars, by the vanquished and winning sides alike, may have served the important function of re-establishing political peace. Although a practically compulsory form of "choice" only remotely related to what is an election, the losing side's recognition through acclamation of the new state of affairs, served to reduce tension and integrate the community. Taking sides in the selection of leaders or the settlement of controversy seems to have been the next step in the evolution of elections. A 6th-century Athenian law prescribed that in case of civil turmoil all citizens were to join one side or another or lose their status as citizens, for abstention from the conflict by too many would facilitate tyranny.

Elections in Greece

Although elections are nowadays invariably associated with one form or another of representative government, the practice of voting probably preceded the institution of representation, and quite certainly preceded any theoretical understanding of representation. The Greek system of representation was not highly developed and was thought of as limiting rather than extending the participation of citizens in government. Insofar as elections served the purpose of filling public offices, they were used when the duties involved required special knowledge and skill, such as military leadership—a reversal of the modern practice of staffing technical offices through appointment of qualified experts. In general, however, public offices were filled by lot from among tribal demes (units of local government), in a system of short-term rotation, as in the Athenian Council of Five Hundred. Any other procedure was feared to jeopardize the principle of equality and open the door to ambitious men who might seek to exploit their electoral victory for usurping political power.

Voting on measures in the Greek and Roman assemblies was usually open and by show of hands although acclamation by voice was sometimes used. In some voting situations, as in votes of ostracism and lawsuits, decisions were made by use of secret ballots in the form of white and black pebbles, marked and unmarked shells, or carved wooden tablets.

In ancient northern and eastern Europe, acclamation in support of leaders or decisions by the clashing of spears or shields signified unanimity. Unanimous consent was sought because of the prevailing feeling that what could not be done unanimously should not be done at all. As all participants were deemed equal, they all had to take part in making decisions, and the requirement of unanimity guaranteed that they did. All were winners and none were losers. But once the condition of equality gave way to inequality in status and power, as in the medieval Polish diet or the Aragonese cortes, the unanimity rule proved unworkable because it represented, in effect, a form of veto that made collective choice impossible.

THE MEDIEVAL PERIOD

Decline of elections

In the period between the decline of the Roman Empire and the emergence of absolute monarchy in the 16th century, elections probably survived at the local level in towns and regional assemblies. But the medieval social structure was not congenial to electoral practices. In the church election of superiors by their peers was practiced to facilitate and legitimize ascendancy through the hierarchy, but the electorates involved were very small, and to what extent election differed from appointment, it is difficult to say.

Throughout the medieval period the evolution of electoral processes was intimately connected with the tortuous efforts to create workable representative institutions and government by consent. Although relevant solutions, stemming largely from the Roman legal tradition, were proposed by ecclesiastical and constitutional theorists, neither the corporate feudal structures of the early Middle Ages nor the de facto bargaining mechanisms of feudalism in decline were propitious environments for collective decision making. Constitutional development, beginning in 13th-century England, was more in response to rapidly changing political requirements than in pursuit of formal doctrine. The king had a growing need to surround himself with advisers whose voice "counted" (that is, who could be counted), and to solicit the support of the increasingly numerous and densely populated towns on whose consent the success of raising taxes for the royal treasury depended. At the same time, petition for royal favours or the redress of grievances, initially a privilege, came to be regarded as a right. Although it is impossible to trace the history of voting as such, it is possible to see that the need for consent, on the one hand, and the practice of petition, on the other hand, could be readily reconciled in voting.

Throughout the medieval period, in contrast to both antiquity and modern times, the units of counting were not individuals but representatives of landed interests or geographical localities such as towns and boroughs. Although medieval writers, following the Roman tradition, often recognized election as the potential source of popular sovereignty, there was no significant concern with electoral processes. In reality, inheritance of positions and offices became the effective procedure of regulating succession. In the royally convened assemblies voting either called for unanimous consent or involved manipulation by the king's men to overcome recalcitrant opposition. Except in England and Sweden, the age of absolutism in the 16th and 17th centuries represented a period of discontinuity in the evolution both of representative institutions and of those feeble vote mechanisms that may have been used in the medieval period.

EXTENSION OF THE SUFFRAGE

The extension of the suffrage has largely been a consequence of the rapidity of change in a society's social structure and the customary relations among its people. Where social change was relatively slow, as in ancient Rome and in postmedieval England through the 19th century, ruling groups could successfully maintain effective power through gradual extension of the suffrage to would-be opposition groups, thereby co-opting them into the governmental process. On the other hand, where, after a protracted period of social stagnation, changes in social structure went unrecognized, as in many European areas in the postfeudal era through the period of absolutism to the French Revolution, the later battle for the suffrage was not only often violent, but the speed in extending it often found the new electorate unprepared to meet its obligations, thereby contributing to governmental instability.

The need for effective organization

Although the extension of the suffrage to previously disenfranchised population groups is a concomitant of change in a society's social structure, the effectiveness with which the suffrage can be pursued, and the probability of using it successfully once gained, depend on effective organization of the groups demanding the vote. Constitutions often provide for a wide franchise, but the utilization of the right to vote may falter for lack of organization among suddenly enfranchised population groups.

The extension of the suffrage to the working class in 19th-century Europe was anticipated by the formation of workingmen's associations and parties. The extension of the suffrage to women depended on successful prior organization, and intensive, indeed, often shrill, pressure for the vote. By way of contrast, the failure of an effective suffrage for blacks in the American South, following emancipation and formal enfranchisement by the Fifteenth Amendment of the U.S. Constitution, was probably due as much to the organizational weakness and lack of leadership among the former slaves, at it was due to intimidation and biassed administration of the election laws. It was not until almost a century later that the black population could provide sufficient organization and leadership to make the right to vote and other civil rights an effective demand on society.

Early fears about the extension of the suffrage to groups with presumed low stakes in the social order—the propertyless, the enslaved, the uneducated, the itinerant, the disinterested, or the young—appear, in retrospect, as little more than either ill-founded prejudices or strategems to maintain the status quo. Devices invented to offset the effects of extending the suffrage, such as giving property owners or the highly educated more than one vote or highly restrictive residence requirements, failed in their purpose. Social and economic developments hardly to be anticipated at the end of the 18th century so radically transformed modern societies in the next century that few of the arguments advanced in favour of property, income, or educational qualifications could be defended in either logic or fact. Indeed, as it became evident that more and more of the population had a stake in society, the suffrage came to be recognized as an instrument of governance that is both a private right and a public duty.

Although female suffrage was instituted in three or four countries before World War I, it was the new role of women during that war that provided the major impetus for their enfranchisement. Resistance to the demands of the suffragettes crumbled as women, in the course of wartime manpower shortages, entered into a previously male-dominated social and economic order. During the 1918–20 period alone, women achieved the vote in Belgium, the Netherlands, Germany, Poland, Canada, the United States and many other countries. In Great Britain women over 30 years of age were granted the vote in 1918, but not until 1928 was the age requirement equalized at 21 for both men and women. Where women were excluded from voting in the interwar years, as in France, Italy, and Japan, it was their role in World War II that finally led to their inclusion in the electorate. The late 1940s also witnessed the extension of the vote to women in most South American countries. In the Arab nations female suffrage is still subject to numerous qualifications. In the Communist regimes, male and female suffrage has been universal from their inception.

Voting ages

The next stage in the extension of the suffrage has been a movement to reduce the voting age, normally 21, to 18. As large numbers of the young are called upon to do military service and as a large proportion of them receive more and more higher education, the significance of their role in the social order cannot be denied. Eighteen year-olds vote in several countries, including Britain and the United States.

The extension of the vote to women and younger age groups, as well as relaxed residence requirements, has greatly broadened the eligible electorate, although discriminatory practices, such as literacy tests and poll taxes, continue to restrict the franchise.

ELECTIONS AND REPRESENTATIVE GOVERNMENT

The widespread use of elections in the modern world is largely due to the gradual emergence of representative government from the 17th century on. For proper appraisal, however, it is important to distinguish between the form and substance of elections. Electoral forms may be present but the substance may be missing. The substance is, of course, that the voter has a free and genuine choice between at least two alternatives. In a purely formal sense, the great majority of the approximately 130 contemporary nations have what are called "elections," but probably only a third of these have more or less competitive elections; perhaps a fifth have one-party elections; and in some others the electoral situation is highly ambiguous.

The discovery of the individual as the unit to be counted was, from the 17th century on, the critical factor in the emergence of modern electoral processes. The counting of individuals, in turn, was a by-product of the change from the holistic conception of representation in the Middle Ages to an individualistic conception. The British Parliament, for instance, was no longer seen as representing estates, corporations, and vested interests, but as standing for actual human beings. The movement abolishing "rotten boroughs"—boroughs of small population controlled by one person or family—that culminated in the Reform Act of 1832 was a direct consequence of the individualistic conception of representation. Once governments were not only believed to derive their powers from the consent of the governed but were expected to seek consent, the only remaining problem was to decide who was to be included among the governed whose consent was to be sought. The democratic answer was, of course, universal adult suffrage.

Although it is common to equate representative government, elections, and democracy, and although competitive elections under universal suffrage are in many respects a defining characteristic of political democracy, universal suffrage is not a necessary condition of competitive electoral politics. An electorate may be limited by formal legal requirements—as was the case before universal adult suffrage—or it may be limited by citizens' failure to take advantage of the vote, as is often the case in American municipal and other elections. Although such legal or self-imposed exclusion affects the democratic quality of elections and may ultimately affect the legitimacy of government, it does not impede decision-making by election, provided the voter is presented with alternatives among which to choose.

Access to the political arena during the 18th century depended largely on membership in some aristocracy, and participation in elections was regulated mainly by local customs and arrangements. With the American and French revolutions every citizen was declared formally equal to every other citizen, but the vote remained an instrument of political power possessed by very few.

Even with the arrival of universal suffrage, the ideal of one man, one vote was not achieved. Systems of plural voting were maintained in some countries giving certain social groups an electoral advantage. In Great Britain, for example, university graduates and owners of businesses in constituencies other than those in which they lived continued to have an extra vote until 1948. Before World War I, both Austria and Prussia had three classes of weighted votes that effectively kept electoral power in the hands of the upper social strata.

Elections in authoritarian systems

Whereas in the Western nations of the 19th and 20th centuries the increasing use of competitive mass elections in selecting governments had the purpose and the effect of institutionalizing the diversity of modern societies, in the Eastern, one-party, Communist regimes, mass elections came to have quite different purposes and consequences. Elections in these nations were not designed to offer choice of leadership or to influence policy formation. They differ from competitive elections in that only one team of candidates is presented, and the voter has only the choice of voting for the official candidate. They are in the nature of the 19th century Napoleonic plebiscites, intended to demonstrate the unity rather than the diversity of the people, and, above all, they do not make or unmake governments.

Although the Soviet Union's constitution of 1936 provided for elections on the Western model, the election mechanism is manipulated to produce almost universal turnout and unanimous support of the regime. Elections are dramatic occasions for Communist agitation and propaganda to strengthen the citizen's identification with the Soviet way of life.

Because voting is considered a pledge of loyalty, it is easier to vote against the official candidate than to abstain. Although nonparticipation is not a legal offense, it may be socially punished. Although an unmarked ballot is considered a "yes" vote, dissent may be registered by crossing out the name of the candidate on the ballot, as some 2,000,000 Soviet citizens do in each election. These voters are easily identified because they are likely to choose voting booths that make secret voting possible. Perhaps for this reason less than 5 percent of the voters risk public scrutiny by secret voting. It may well be that dissenting votes are cast not so much because of dislike of Communism, but because of grievances involving the conduct of minor officials who stand for election to one of the numerous village, town, republic, or oblast soviets.

Not all elections under authoritarian governments follow the Soviet model. In Poland, for instance, more names appear on the ballot than there are offices to fill, and some degree of electoral choice is possible. Electoral participation is usually over 90 percent, but effective dissent is minimal. It is difficult to appraise election results in the Eastern, one-party, Communist nations because published data are unreliable and cannot be easily compared.

In the non-Communist authoritarian regimes of the world a variety of means is employed to eliminate free electoral choice, ranging from suppression of the opposition to the physical intimidation of voters.

Modern electoral systems

FUNCTIONS OF ELECTIONS

Fundamental to the use of elections is the contribution that they make to democratic government. Where the

members of the body politic cannot themselves govern and must entrust government to representatives, elections serve not only to select leaders acceptable to the voters but also to hold the leaders accountable for their performance in office. Accountability, however, is greatly jeopardized in electoral situations in which elected leaders, for want of ambition, do not care whether or not they are re-elected or in situations in which, for historical or other reasons, one party is so predominant as to preclude effective choice among alternate candidates or policies. Nevertheless, the possibility of controlling leaders by requiring them to submit to regular and periodic elections contributes to solving the problem of succession in leadership and, thereby, to the continuation of democracy. Moreover, where the electoral process is competitive and forces candidates or parties to expose their record of accomplishment and future intentions to popular scrutiny in election campaigns, elections serve as forums for the discussion of public issues, facilitate the expression of public opinion, and, more generally, permit an exchange of influence between governors and governed.

Democratic functions

Elections also serve to reinforce the stability and legitimacy of the political community in which they take place. Like national holidays commemorating common experiences, elections serve to link the members of a body politic to each other and thereby confirm the viability of the political community. By mobilizing masses of voters in a common act of governance, elections lend authority and legitimacy to the acts of those who wield power in the name of the people.

Elections can also confirm the worth and dignity of the individual citizen as a human being. Whatever other needs he may have, participation in an election serves to gratify the voter's sense of self-esteem and self-respect. It gives him an opportunity to have his say, and he can, through expressing partisanship and even through nonvoting, satisfy his sense of belonging to or alienation from the political community. It is for just this reason that the age-long battle for the right to vote and the demand for equality in electoral participation can be seen as the manifestation of a profound human craving for personal fulfillment.

In all forms of government, from the most democratic to the most totalitarian alike, elections have a ritualistic aspect. Elections and the campaigns preceding them are dramatic events which, depending on cultural or historical circumstances, may exude the gay atmosphere of a circus or the sombre atmosphere of a funeral. Rallies, banners, posters, buttons, headlines, and television call attention to the importance of participation in the event. Candidates and parties, from right to left, in addition to propagating their policy objectives through rhetoric and slogans, invoke the symbols of nationalism or patriotism, reform or revolution, past glory or future promise. Whatever the peculiar national, regional, or local variations, elections are events that, by arousing emotions and channelling them toward collective symbols, break the monotony of daily life and focus attention on the common fate.

DECISION RULES

Individual votes are totalled into collective decisions by rules of counting that voters and leaders have accepted as legitimate prior to the election. These decision rules may call for plurality voting, which requires that, among three or more alternatives, the winner need have only the highest number of votes; simple majority voting, which requires that the winner receive more than 50 percent of the vote; extraordinary majority voting, which requires some higher proportion for the winner, such as a two-thirds vote; or unanimity.

Plurality voting

The simplest means of deciding an election is the plurality rule. To win, a candidate need only poll more votes than any other single opponent; he need not, as required by the majority formula, poll more votes than the combined opposition. The more candidates contesting a constituency seat, the greater the probability that the winning candidate will receive less than 50 percent of the vote. The plurality formula is used in the national elections of such countries as Great Britain, Canada, and New Zealand. One half of the West German Bundestag is elected according to the plurality system, while the other half is elected according to a proportional representation formula (see below).

Under the majority rule, the party or candidate winning over 50 percent of the vote in a constituency is awarded the contested seat. The winning party or candidate must poll more votes than the combined opposition. This system thus ensures that the elected representative has the support of the majority of the voters. The majority formula is employed in the election of the Australian House of Representatives, and of the French National Assembly. It is usually applied only within single-member electoral constituencies.

Both the plurality and the majority-decision rules are employed in the election of U.S. presidents. The composition of the electoral college, which actually elects the president, is determined by a plurality vote taken within each state. Voters choose between the names of the presidential candidates, but they are in effect choosing the electors who will elect the president by means of a majority vote in the electoral college. All of a state's electoral votes (which are equal in number to its seats in Congress) are given to the presidential candidate who gains a plurality of the vote in the state election. It is thus possible for a president to be elected on the basis of a minority of the popular vote.

A critical difficulty with the majority formula is that in a multiparty political system, the formula may produce an electoral deadlock if no candidate secures 50 percent of the total vote. In order to break such deadlocks a second round of elections (second ballot) is required, if no candidate obtains a majority on the first round. In Australia, voters rank the candidates on an alternate preference ballot. If a majority is not achieved on the first round of elections, the weakest candidate is eliminated, and the votes he accrued are distributed to the other candidates according to the second preference on the ballot. This redistributive process is repeated until one candidate collects a majority of the votes. If no candidate secures a majority in the first round of the French National Assembly elections, another round of elections is required. In this second round, the candidate securing a plurality of the popular vote is declared winner.

Neither the majority nor the plurality formulas distribute legislative seats in proportion to the share of the popular vote won by the competing parties. Both formulas tend to award the strongest party disproportionately and to handicap the weaker parties. The latter is particularly true if small parties are rooted in ethnic, religious, or social minorities; small parties escape the inequities of the electoral system only if they have a regionally concentrated base. The plurality formula distorts the distribution of seats more than the majority system.

Proportional methods

Proportional representation requires that the distribution of offices be proportional to the distribution of the popular vote among competing political parties or candidates. It seeks to overcome the distribution imbalances that result from majority and plurality formulas, and to create a representative body that reflects the distribution of opinion within the electorate. Because of the use of multimember constituencies in proportional representation, parties with neither a majority nor plurality of the popular vote can still win legislative representation.

Proportional representation is an ideal that is sought after, but only approximated: the size of electoral districts is the critical factor; the larger the electoral district in terms of seats, the more proportional the representation will be. The number of seats assigned to a constituency is in fact a greater determinant of the proportionality of the outcome than is the specific type of proportional formula used. The different formulas of proportional representation are basically similar in their effect on the conversion of votes to political representation.

Developed in the 19th century in Denmark and in Britain, the single transferable vote formula—or Hare system (after one of its English developers, T. Hare)—em-

ploys a ballot that allows the voter to rank the competing candidates in order of preference. When the ballots are counted, any candidate receiving the necessary quota of first preference votes is awarded a seat. In the electoral calculations, votes received by a winning candidate in excess of the quota are transferred to other candidates according to the second preference marked on the ballot. Any candidate who then has the necessary quota is also awarded a seat. This process is repeated, with subsequent surpluses also being transferred, until all the remaining seats have been awarded. Five members constituencies are considered optimal for the working of the Hare system.

The single transferable vote formula, because it involves the aggregation of ranked preferences, necessitates complex electoral computations. This factor, plus the fact that the Hare system limits the influence of political parties, probably accounts for its infrequent use; at present it is used only in Eire and in the selection of the Australian and South African senates. The characteristic of the Hare formula, which distinguishes it from other proportional representation formulas, is its emphasis on candidates, not parties. The party affiliation of the candidates has no bearing on the computations.

The basic difference between the transferable vote formula and the list systems is that, in the latter, voters choose among party-compiled lists of candidates rather than among individual candidates. Although voters may have some limited choice among individual candidates, electoral computations are on the basis of party affiliation; seats are awarded in respect to party rather than candidate totals. The seats that a party wins are allocated to its candidates in the order in which they appear on the party list. Several types of electoral formulas are used, although they fall into two main categories; largest average formulas and greatest remainder formulas. All employ some type of electoral quota.

Largest average formula

In the largest average formula, the number of votes won by each party is divided by the number of seats held by the party, plus one. The first seat is awarded to the party with the highest number of votes, since, no seats yet having been allocated, the initial denominator is 1. When a party wins a seat, its formula denominator is increased by one and hence the party's chances of winning the next seat are reduced. The available seats are awarded one at a time to the party with the greatest average. Party totals, not candidate totals, are used in the calculations. No transfer of ballots takes place. Frequently named after its Belgian inventor, Victor d'Hondt, the largest average formula is used in Austria, Belgium, Finland, and Switzerland, and in the proportional representation half of the West German Bundestag elections.

The d'Hondt formula has a slight tendency to over reward large parties and to reduce the chance of small parties gaining legislative representation. The so-called Lague variation of the d'Hondt formula—used in Denmark, Norway, and Sweden—reduces the reward to large parties but increases further the handicap to small parties. By adjusting the denominator of the d'Hondt formula, the Lague formula increases the cost of both the initial seat of a party and that of additional seats. The Lague formula thus aids middle-size, political parties and reduces the number of legislatively represented small parties.

The greatest remainder method establishes a vote quota for each seat in an electoral district by dividing the total vote in the district by the number of competing parties. The total popular vote won by each party is then divided by the quota, and a seat is awarded as many times as the party total contains the full quota. If all the seats are awarded in this manner, the election is complete. Such an outcome is unlikely, however. Seats that were not won by full quotas are subsequently awarded to the parties with the largest remainder of votes after the quota has been subtracted from each party's total vote for each seat it was awarded. Seats are awarded sequentially to the parties with the largest remainder until all the district's allocated seats have been awarded.

Of all the proportional representation formulas, the greatest remainder formula, given large enough constituencies, yields results closest to the proportional ideal. Small parties fare better when the greatest remainder formula is used than when the largest average formula is employed. The greatest remainder formula is used in Israel, Italy, and Luxembourg and in some elections in Denmark. Italy, however, uses a special variety of the formula, called the Imperiali formula, whereby the electoral quota is established by dividing the total popular vote by the number of parties plus two. This modification increases the legislative representation of small parties, but leads to a greater distortion of the proportional ideal than the unaltered formula.

The choice of majority and plurality or proportional systems is, of course, not simply a matter of pure theory. Different methods of counting, just as different conceptions of representation, usually reflect cultural, social-structural, and political circumstances in a particular jurisdiction. Majority or plural methods of voting are most likely to be acceptable in relatively stable political cultures. In such cultures, fluctuations in electoral support, given to one party or another from one election to the next, reduce polarization and make for political centrism. Thus the "winner take all" implications of the majority or plurality formulas are not experienced as unduly deprivational or restrictive. Proportional representation, on the other hand, is more likely to be found in societies with traditional ethnic, linguistic, and religious cleavages or in societies experiencing pervasive class and ideological conflicts.

TYPES OF ELECTIONS

Elections of officeholders. Rather than deciding public policy, election to public office bestows the right to decide on a small group of elected officials. The elections grant officials the authority to make public policy decisions, and thereby legitimize the decisions that are made. The policy voice of the electorate in these elections may be weak. An electoral mandate results not so much in the implementation of the abstract preferences of the majority of the citizens, but rather in the implementation of the policy of those leaders who come to power in elections.

Role of parties

Most officeholder elections are interpretable only in terms of the role of political parties in the political process. Parties are tied so closely to the democratic functions of elections that their competition for the electorate's support is considered the main criterion by which democratic elections are distinguished from nondemocratic elections. Because of the difficulty in fixing individual responsibility in the complexity and vastness of modern governments, political parties, rather than representatives, have become the vehicle of democratic accountability.

Political parties shape the nature of the electoral decision. The selection and nomination of candidates, a vital first stage of the electoral process, lies in the hands of the political parties; the election serves only as the final process in the recruitment to political office. The party system can thus be regarded in a true sense as an extension of the electoral process. Political parties provide the pool of talent from which candidates are drawn, and they simplify and direct the electoral choice and mobilize the electorate at the registration and election stage. Political parties have transformed elections from a choice of men to a choice of governments.

But the fusion of political parties with the electoral process has not gone unchallenged. In North America a number of modifications have been made in the electoral structure to limit the influence of political parties. Nonpartisan elections (elections in which party affiliations are not formally indicated on ballots) are the norm for some municipalities in the United States and Canada and represent an attempt to remove political parties from the electoral scene. Nonpartisanship started as a reform movement in the early part of the 20th century and was designed to put local government on a more businesslike basis and to isolate local politics from state and national

politics. The exclusion of party labels from the ballot, however, does not guarantee that political parties will be inactive or without influence in nonpartisan elections.

Recall. The practice of recalling officeholders was, like most populist innovations, an attempt to minimize the influence of political parties on representatives. The recall was an electoral modification designed to ensure that an elected representative would act as a spokesman for his constituency rather than for his political party or the dictates of his own conscience or interests. It enabled constituents to recall their elected representative before the end of his term of office if, in their opinion, he was no longer acting in accord with the best interests of the constituency. The actual instrument of recall was often a letter of resignation signed by the elected representative *before* assuming office. During the term of office this letter could be evoked by a quorum of constituents if the representative's legislative performance failed to meet their expectations. In practice the recall is not often or widely used today, even where it is provided for constitutionally.

Referendum. A referendum is an election, held either at the initiative of the governors or the governed, in which the preferences of the community regarding a particular issue are assessed. The use of the referendum reflects a reluctance to entrust full decision-making power to elected representatives. Because of the small voting turnout for referendum elections, and the absence of affixed party labels to the referendum alternatives, voting in referendum elections is more open to leadership and interest group influence than is voting in officeholder elections.

Referenda usually concern the raising and spending of public money, although they are occasionally used to decide moral issues in which the elected bodies are deemed to possess no special competence. They may be legislatively binding, or merely consultative, but even consultative referenda are likely to be considered legislative mandates.

Plebiscites. Plebiscites are elections held to decide two paramount types of political issue: the form of government and the national boundaries adopted by the state. In the first case, the plebiscite is a means of establishing the legitimacy to rule, not a means of providing a choice among alternate governmental teams or legislative proposals. The incumbent government, seeking a popular mandate as a basis for legitimacy, employs a plebiscite to establish its right to speak for the nation. The incumbent rulers seek a single identity with the nation, and thus a vote for the nation is interpreted as a vote for the incumbent regime. Plebiscites of this nature are thought to establish a direct linkage between the rulers and the ruled; intermediaries such as political parties are bypassed in plebiscites, and in this respect they are antithetical to pluralism and competitive politics. Following the Revolution of 1789, the plebiscite had a great appeal in France, rooted as it was in the ideas of nationalism and popular sovereignty. In more recent times totalitarian regimes have employed plebiscites to legitimize their power.

Plebiscites have also been used as a device for establishing the boundaries of political units according to criteria of nationality. After World War I the plebiscite was employed as a means of fitting political boundaries to boundaries of nationality. Eleven plebiscites, based on universal suffrage, were proposed by the League of Nations, the most successful of which was held in 1935 in the Saar (which chose to remain a German territory rather than go to France). The potential of plebiscites for establishing or changing political boundaries is, however, severely limited, because they entail the prior agreement of two governments over what is usually a very contentious issue.

CONSTITUENCIES

The drawing up of constituencies—those subdivisions of the total electorate that send representatives to the local or central assembly—is inextricably linked with questions concerning the nature of representation and methods of voting. Of these questions, that of representation is primary. The problem of electoral representation turns on the question of what is to be represented. In this connection the functions of constituencies and the problem of determining them (districting) become relevant. The central problem is that the constituency, as a geographical area, may contain within its boundaries diverse, and sometimes incompatible, social, economic, religious, or ethnic interests, all of which seek to be represented.

Districting

The solution to this problem has been largely historically determined. Where the interests of electors have not been totally incompatible, and where ethnic, religious, social, and economic differences have been relatively free of passionate conflict, as in the Anglo-American countries, geographical areas (electoral districts) have usually been considered the constituency, and the method of counting has been some system of majority or plurality voting. The elected person represents the geographical unit as a whole, irrespective of internal divisions. In general, the district is a single-member constituency, although multimember constituencies are also possible.

On the other hand, where the electorate is composed of several minorities, none of which can hope to obtain a majority, or perhaps even a plurality sufficiently large to obtain representation, the geographical district can only be regarded as an administrative unit for counting votes. The effective constituency is the group of electors that can be identified as having voted for a given candidate. In this case, election is usually by some method of proportional counting whereby any candidate, party, or group receiving a requisite number of votes is entitled to a proportionate number of representatives.

The drawing up, or delimitation, of electoral districts is inextricably linked with differing conceptions of representation; and conceptions of representation are inextricably linked with alternate methods of vote counting. The virtues and vices of different electoral arrangements are still debated by constitution makers, but the arguments advanced in favour of any one set of arrangements remain inconclusive. Although proportional systems treating electoral districts as merely administrative conveniences and providing for multiple representation are believed to approximate more closely the one man, one vote principle of democratic theory than do single-member geographic constituencies providing for majority or plurality voting, they are probably conducive to governmental instability, and they place an inordinately heavy burden of conflict resolution on representative bodies. This often makes for political stalemate or immobilism, as has been the case at times in France and Italy. But it is doubtful that in nations with many interests seeking representation any other method of electoral districting and vote counting would be satisfactory.

Representation of geographic constituencies with majority or plurality voting, on the other hand, disguises many differences in the electorate's composition and preferences. In single-member, majority vote electoral districts, the minority losing at the polls may feel unrepresented. But it is often a precarious assumption that the majority itself is truly represented, since it is likely to be a loose coalition of diverse interests, and no single representative can usually do justice to the diversity of interests involved. In general, therefore, the representative, in so far as he is responsive, is more likely to play the role of broker among his district's diverse interests, rather than the role of spokesman for the district as a whole.

Whatever the role of representative, and however electoral units are conceived—as geographic areas with stable, compatible interests, or as purely administrative areas for the purpose of vote counting—districting must be distinguished from apportionment. In general, if the district is considered as only an administrative unit for counting votes, its boundaries can be drawn rather arbitrarily without injustice. The district as such has no stakes of its own to be represented. On the other hand, if the district is considered a true constituency, in the sense that it has unique interests that can be geographically and, as a result, socially defined, districting should not be

arbitrary. The area involved should at least be contiguous and compact so that its presumed interests can be fairly recognized.

The difficulty in treating geographic areas as genuine constituencies is, of course, that internal transformations in their economy and social structure may make their historical boundaries obsolete. As a result, historical areas serving as electoral districts (such as provinces, states, or counties), which once served as genuine communities of interest to be represented, may no longer do so. As these districts become increasingly diverse, making it difficult to identify particular constituency interests, representatives find it increasingly impossible to be responsive to constituency wishes or needs.

Apportionment

Problems of apportionment, in contrast to problems of districting, stem from efforts to reconcile the territorial and population bases of representation. If geographic areas, for instance, are assumed to have an equal right to be represented because the area is considered to be a viable constituency, malapportionment in terms of population size is inevitable. New York with a population of nearly 20,000,000 has, in the U.S. Senate, the same number of representatives as Nevada with a population of less than 1,000,000. This kind of "constitutional malapportionment" must not be confused with the kind of "electoral malapportionment" that defies the one man, one vote principle of equal representation. The latter is usually the result of population shifts. Under modern conditions of increasing urbanization, failure to reapportion the number of seats in representative bodies to take account of population changes has largely benefitted rural electoral districts. More recently, the shift of city people to the suburbs has led to possible underrepresentation of suburban vis-à-vis urban populations.

From a political rather than a legal perspective, malapportionment is usually considered "undemocratic" because it results in overrepresentation and underrepresentation of certain sectors of the population and, as a result, may culminate in public policies not acceptable to the true majority.

Apportionment is not a simple matter. In particular, the definition of the population among which a specified number of legislative seats is to be apportioned is not easy to arrive at. Total population, number of citizens of voting age, number of registered voters, or number of actual voters can serve as bases of apportionment. If there were a constant relation between these groupings, apportionment would not be a problem, and any one of them could serve as a basis. In fact, however, the relation is not constant because of variations in registration and voting turnout.

Gerrymandering

Malapportionment, whether of the constitutional or electoral sort, must not be confused with gerrymandering—a form of arbitrary districting used to benefit the party that at a given time controls the apportionment process. Gerrymandering is named after governor of Massachusetts Elbridge Gerry (1744–1814) who recognized the possibility of influencing electoral outcomes by manipulating the boundaries of electoral districts. The practice involves concentrating large percentages of the opposite party's votes into a few districts, and drawing the boundaries of the other districts in such a way that the gerrymandering party wins in them all, even though the majority, or, in three-cornered elections, the plurality, is relatively small.

Gerrymandering comes to politicians easily. Party managers prefer "safe" districts to those that are competitive. By not contesting seriously in the other party's safe districts, by maintaining one's own safe districts, and by carving competitive districts up in such a way that one's own party is favoured, party managers use gerrymandering to create artificial district boundaries that, though advantageous to the parties, may harm other social, economic, and political interests that need representation.

ELECTORAL PRACTICES

There is a direct relation between the size of an electorate and the formalization and standardization of its voting practices. In very small voting groups, in which political encounters are face-to-face and the members are bound together by ties of friendship or common experience, voting is mostly informal and may not even require counting, because the "sense of the meeting" emerges from the group's deliberations. The consensus of the Quaker meeting is of this order. An issue is discussed until a solution emerges to which all participants can agree or, at least, from which any one participant will not dissent.

By way of contrast, in modern mass electorates, in which millions of individual votes are aggregated into the collective choice, formalization and standardization of voting practices and vote counting are the rule. This is necessary in order to guarantee that the outcome can be considered valid, reliable, and legitimate. Validity means that the collective choice in fact expresses the sense of the electorate. Reliability means that each vote is accurately recorded and effectively counted into the total. Legitimacy means that the criteria of validity and reliability have been met; so that the result of the voting is acceptable and provides authoritative guidelines in subsequent political conduct.

The development of routinized and standardized electoral practices in mass electorates is a surprisingly recent phenomenon, not much older than 100 years. It is as much a corollary of the growth of rapid communication through telephone and telegraph as of the growth of the electorate and rational insistence on making electoral processes fair and equitable. Nevertheless, even today electoral practices around the world differ a great deal, depending not just on formal institutional arrangements, but even more on a country's political culture.

Secret voting. Once suffrage rights had been extended to masses of voters who, in theory, were assumed to be equal but who, in fact, were unequal (in order of birth, in intellectual endowment and educational accomplishment, in social status and the possession of property, and so on), open voting was no longer tolerable precisely because it could and often did involve undue influence, ranging from hidden persuasion and bribery to intimidation, coercion, and punishment. Equality, at least in voting, was not something given but something that had to be engineered; the secrecy of the vote was a first and necessary administrative step toward the one man, one vote principle. Equality in voting was possible only if each vote was formally independent of every other vote, and this suggested the need for strict secrecy.

The slow progress made in introducing secret voting, from the French Revolution well into the 20th century, attests to the fact that social engineering, no matter how desirable or lofty in purpose, depends on favourable conditions for success. One need not assume that the obstacles placed in the way of secret voting were the result of some conspiracy on the part of those who recognized, quite accurately, that the mobilization of large numbers of voters fundamentally changed the distribution of political power and who, through opposing the secret vote, hoped to maintain a stranglehold on the newly enfranchised electorates. Rather, the success of secret voting depended on the reduction of illiteracy and, at the cultural level, on the spread of the individualistic norms of privacy and anonymity to certain classes of the population, notably peasants and workers. Traditionally these groups took their cues from those accepted as superior in status, or from their peers. Secret voting required learning to free oneself as a citizen from one's customary associations and from pressures for conformity. The difficulty of introducing and practicing the secret vote in today's politically and economically less developed nations mirrors the tortuous advance of the secret ballot in the Western nations during the 19th and early 20th centuries.

Benefits of secret voting

Secret voting reduces drastically the possibility of undue influence on the voter. Without it, influence can range from outright purchase of votes to social chastisement or economic sanctions. This is not to say that bribery in voting is automatically eliminated by secret voting. Laws prohibiting and punishing the purchase of votes are on

the statute books of many countries and undoubtedly contribute to discouraging the practice.

Although informal social pressures on the voter are probably unavoidable and, in some respects, useful in reducing political rootlessness and contributing to political stability, secrecy in voting permits voters to break away from their social moorings and gives them a considerable degree of independence if they wish to take advantage of this electoral freedom. As a result it becomes ever more difficult for interest groups, whether labour unions, farmers' organizations, commercial or industrial associations, ethnic leadership groups, or even criminal syndicates, to "deliver the vote." The extent to which "deviant voting" occurs depends partly on the degree of rigidity in the social structure. In countries where caste or class barriers are high or where traditional social, economic, religious, or regional cleavages remain strong, deviant voting is less likely than in countries where social mobility is possible and where political conflicts cut across traditional social cleavages. In Western nations the increasing difficulties of labour, farm, religious, or ethnic leaders in influencing the voter attests to the success that secret voting seems to have had on freeing the individual from electoral bondage to his traditional or economic affiliations.

Balloting. The ballot makes secret voting possible. Its initial use seems to have been a means to reduce irregularities and deception in elections. This objective, however, could only be achieved if the ballot was not supplied by the voter himself, as was the case in much early voting by secret ballot, or by political parties, as is still the case in some countries. Ballot procedures differ widely, ranging from marking the names of preferred candidates to crossing out those not preferred or writing in the names of persons who are not formal candidates. Ballots also differ according to the type of voting system employed. Where plurality or majority voting is practiced, most elections employ classified ballots whereby the voter casts his vote for only one candidate or list of candidates. Where proportional methods are used, election is by ballots that enable the voter to rank the candidates according to his preferences.

Though evidence is hard to come by, it is commonly believed that the nature of the ballot influences the voter's choice. In jurisdictions where electors are called upon to vote not only for higher offices but also for a multitude of local positions and where, in addition, the election may include propositions in the nature of referenda, the length of the ballot seems to be a factor affecting vote outcomes. Overwhelmed by the length of the ballot, voters may be discouraged from expressing their preferences for candidates of whom they have not heard, or from deciding on propositions that they do not understand. Breaking up of the single ballot into separate short ballots helps overcome this problem but does not eliminate it. Election data show a rapid decline from votes cast for higher offices to those cast for lower offices and referendum-type propositions.

Ballot position also seems to have an effect on the votes cast for particular candidates, especially in the absence of cues as to party affiliation or other identifications. The first position on the ballot may be favoured, and in the case of a long ballot both first and last names may benefit, with candidates in the middle of the ballot suffering a slight handicap. Ballot position is likely to have its greatest impact in nonpartisan elections, primaries, and elections for minor offices.

Finally, the manner in which candidates are listed—by party column or office bloc—is likely to affect election outcomes. On party-column ballots it is possible to vote a "straight ticket" for all of a party's candidates by entering a single mark, although voting for individual candidates is usually possible. On the other hand, on the office-bloc ballot, voting is for individual candidates grouped by office rather than party. This discourages, though it does not eliminate, voting exclusively for members of one party. This can have important consequences for the structure of government, especially in systems with separated powers and federal territorial organization. If different offices are controlled by different parties, the governmental process may be marked by greater conflict than would otherwise be the case, and governmental decision-making often will be more difficult.

The introduction of voting machines has not substantially changed the balloting process, although it has made it faster, more accurate and economical, and virtually tamper-proof. The voting machine is not without some minor problems of its own, in that it may marginally depress the level of voting. In candidate elections the dampening effect is accounted for by improper use of the machine, a problem that is being overcome through improved machines and voter education.

Compulsory voting. In some nations, notably Australia and Belgium, electoral participation is legally required of all citizens, and nonvoters without legitimate excuses face money fines. The concept of compulsory voting reflects a strain in democratic theory in which voting is considered not merely a right but a duty. Its purpose is to ensure the electoral equality of all social groups. Whether made compulsory in law or through social pressure, it is doubtful that high voter turnout as such is a good indication of an electorate's capability for intelligent social choice. On the other hand, high rates of abstention or differential rates of abstention by different social classes are not necessarily signs of satisfaction with governmental processes and policies and may in fact indicate the contrary.

Electoral abuses. Corruption of electoral practices is, of course, not limited to bribery or intimidation of the individual voter. The possibilities are endless, ranging from the dissemination of scurrilous rumours about candidates, and deliberately false campaign propaganda, to tampering with the election machinery by stuffing the ballot box with fraudulent returns, dishonest counting or reporting of the vote, and total disregard of electoral outcomes by incumbent officeholders. The existence of these practices depends more on a population's adherence to political civility and the democratic ethos than on the prohibitions and sanctions written into the law.

The integrity of the electoral process is maintained by a variety of devices and practices. Permanent and up-to-date registries of voters are maintained to guarantee easy identification of those eligible to participate in elections, and procedures are designed to make the registration process as simple as possible. In most jurisdictions, elections are now held on a single day rather than staggered. Polling hours in all localities are the same, and opening and closing hours are fixed and announced, so that voters have an equal opportunity to participate. Polling stations are manned by presumably disinterested government officials or polling clerks under governmental supervision; and political party agents or party workers are given an opportunity to observe the polling process, enabling them to challenge irregularities and prevent abuses. Efforts are made to maintain order in polling stations, directly through police protection or indirectly through such practices as closing bars and liquor stores. The act of voting itself takes place in voting booths that protect privacy. Votes are counted and often recounted by tellers, watched by party workers to assure an honest count. The transmission of voting results from local polling stations to central election headquarters is safeguarded and checked.

PARTICIPATION IN ELECTIONS

The rate at which individuals participate in the electoral process depends on many factors, including the type of electoral system, the social groupings to which voters belong, the voters' personality and beliefs, their place of residence, and a host of other idiosyncratic factors.

Effect of level and type of election

The level and type of election has a great impact on the rate of electoral participation. Electoral turnout is greater in national than in state or provincial elections, and greater in the latter than in local elections. Because of this, scheduling has an impact on the turnout for local elections. If local elections are held concurrently with provincial or national elections, a higher voter turnout is achieved than for nonconcurrent elections.

Whether an election is partisan or nonpartisan also affects voter turnout—lighter participation occurring in the nonpartisan elections. Participation is also greater in candidate elections than in noncandidate elections such as referenda. There is some evidence that elections based on proportional representation, in which in some cases every vote counts, have higher electoral turnouts than majority or plurality elections. Noncompetitive or safe electoral districts tend to depress the level of voter turnout, whereas marginal elections increase the turnout.

Technicalities in electoral law involuntarily disenfranchise many potential voters. As a result of moving, people may temporarily lose their vote because of residence requirements for voters in their new electoral district. Complicated voter registration procedures, combined with a high level of geographical mobility, significantly reduce the size of the active electorate in the United States, whereas in Canada and Great Britain the size of the electorate is maximized by government-initiated registration immediately prior to elections. Voter registration in the United States is left to the initiative of individuals and political parties.

Group differences

Relatively low levels of electoral participation are associated with rural residence, and also with low levels of education, occupational status, and income. The groups in society that have been most recently enfranchised also tend to vote less. Hence women vote less than men, American blacks less than whites, and working class people less than those of the middle class. Young people who have just turned of age, and have therefore just been enfranchised, are less likely to vote than people who are middle-aged. It is important to note that the nonparticipation of certain groups in elections has important implications; if everyone were to vote, the balance of electoral power would shift toward the recently enfranchised and less privileged members of the society.

Some people are conscientious nonvoters, although such people are rare. Others, perceiving the vote more as an instrument of censure than of support, may not vote because they are satisfied with the present government. This group of voluntary nonvoters is also small, however. In fact, nonvoters have been shown to be generally less satisfied with the political status quo than are voters. The vote is a rather blunt and ineffectual instrument for expressing dissatisfaction, and nonvoting is more likely to be symptomatic of alienation from, than of satisfaction with, the political system.

Supporters of political parties vote more frequently than nonsupporters; to party supporters the vote becomes a pledge of their party loyalty as well as a political instrument.

Those who feel that government policies have some direct relevance to their lives are more likely to vote than those who are disinterested or who sense the government as being more remote.

Finally, a great number of random factors may determine individual participation in specific elections. Election campaigns vary in their intensity. A crisis atmosphere may induce an unusually large number of people to vote, on one occasion, whereas on another the chance to vote for an extremist candidate may increase the participation of the normally disinterested. Even the weather has a substantial impact on election turnout.

INFLUENCES ON THE DIRECTION OF VOTING

The electoral choice of voters is largely influenced by their membership in social groups or by the social groups with which they identify, and, to a much lesser extent, by campaign issues and party promises. The determinants of party membership or party identification are the most crucial factors behind the voting decision. Membership in social and occupational classes, regional, religious or ethnic groupings, and so on, are thus important influences on the direction of the vote.

Because the electoral choice appears to be largely determined by long-standing group and party identifications, the voting decision may appear to be irrational, in that it is not based on a careful calculation of the immediate benefits expected to accrue to the voter from each electoral outcome. Group-determined voting does not preclude rationality, however, for individual self-interest may well be tied to the destiny of the social groups to which the voter belongs.

Independent voting

The independent voter, contrary to much common belief, is likely to be poorly informed politically and relatively uninterested and uninvolved in politics. Paradoxically, however, by switching their support, independent voters represent an important factor in democratic politics.

In most countries the relatively permanent party alignments are based on social and cultural cleavages. But if elections are to be competitive, and if control of the government is to alternate between parties or coalitions of parties, then some voters must switch party support from election to election. New voters and independent voters, therefore, provide a vital source of change in democratic politics.

BIBLIOGRAPHY

Theory and concepts: DUNCAN BLACK, *The Theory of Committees and Elections* (1958), a rigorous analysis of the logic of voting based on elementary arithmetic and geometry; C.F. CNUDDE and D.E. NEUBAUER (eds.), *Empirical Democratic Theory* (1969), a useful collection of theoretical and empirical studies of modern voting behaviour and electoral processes.

Historical development: H.F. GOSNELL, *Democracy: The Threshold of Freedom* (1948), a survey of problems connected with electoral processes, in a historical perspective; J.A.O. LARSEN, *Representative Government in Greek and Roman History* (1955), an historical study of the relationship between representative institutions and electoral processes; CHARLES SEYMOUR and D.P. FRARY, *How the World Votes: The Story of Democratic Development in Elections*, 2 vol. (1918), still the most complete review of the history of elections in the English language.

Comparative electoral systems: R.R. ALFORD, *Party and Society: The Anglo-American Democracies* (1963), a comparative study of the impact of elections on party formations in the United States, Canada, Great Britain, and Australia; S.M. LIPSET and STEIN ROKKAN (eds.), *Party Systems and Voter Alignments: Cross-National Perspectives* (1967), studies of electoral outcomes in Italy, France, Spain, West Germany, Finland, Norway, Japan, Brazil, and West Africa; W.J.M. MACKENZIE, *Free Elections: An Elementary Textbook* (1958), a simple introduction to electoral politics in a cross-national perspective; M.E. MOTE, *Soviet Local and Republic Elections* (1965), a useful survey in an area in which reliable information is difficult to obtain; T.E. SMITH, *Elections in Developing Countries* (1960), an examination of many of the problems besetting the introduction of elections in countries with limited democratic experience.

Institutional aspects: D.W. RAE, *The Political Consequences of Electoral Laws* (1967), an examination of the implications of differences in electoral laws for political outcomes.

Voting behaviour: DAVID BUTLER and D.E. STOKES, *Political Change in Britain* (1969), a recent study of voting behaviour in Great Britain based on sample surveys in 1963, 1964, and 1966; ANGUS CAMPBELL *et al.*, *The American Voter* (1960), an authoritative study of numerous aspects of individual voting behaviour in the United States, and *Elections and the Political Order* (1966), studies of voting behaviour in the United States, Norway, and France.

(H.E./R.G.)

Electrical and Electronics Engineering

Electrical and electronics engineering originally comprised electrical power, illumination, and the telephone and telegraph but has now expanded into other areas including radio communication, digital computation, automatic control of machines and systems, navigational systems as for radar, sonar, loran and shoran equipment, laser measurements, and the science of information as a basic understanding of the communication process. Frequently included in this field is the analysis of large interacting systems ranging from transportation to biological, ecological, and economic systems. These are logically related to electrical engineering because the mathematical processes employed in the solution of interconnected electrical systems can be applied to other complex interrelated systems.

HISTORY OF THE PROFESSION

Electrical phenomena attracted the attention of European thinkers as early as the 17th century. Beginning as a mathematically oriented science, the field has remained primarily in that form; mathematical prediction often precedes laboratory demonstration. The most noteworthy pioneers included Ludwig Wilhelm Gilbert and Georg Simon Ohm of Germany, Hans Christian Ørsted of Denmark, André-Marie Ampère of France, Alessandro Volta of Italy, Joseph Henry of the U.S., and Michael Faraday of England. Electrical engineering may be said to have emerged as a discipline in 1864 when the Scottish physicist and mathematician James Clerk Maxwell summarized the basic laws of electricity in mathematical form and predicted that radiation of electromagnetic energy would occur in a form that later became known as radio waves. In 1887 Heinrich Hertz experimentally demonstrated the existence of radio waves in his laboratory at Karlsruhe, East Germany.

Early theorists

First practical application

The first practical application of electricity was the telegraph, invented by Samuel F.B. Morse in 1837. The need for electrical engineers was not felt until the inventions of Alexander Graham Bell and Thomas A. Edison some 40 years later; Bell's telephone (1876) and Edison's incandescent lamp (1878) and first central generating plant in New York (1882) overnight created a large demand for men trained to work with electricity.

In the early years of the 20th century a whole new field was opened up for the young profession. Edison's discovery of the "Edison effect," a flow of current through the vacuum of one of his lamps, was the first observation of current in space. Hendrick Anton Lorentz of The Netherlands predicted the electron theory of electrical charge in 1895, and two years later Sir J.J. Thomson of England showed that the Edison effect current was indeed caused by negatively charged particles (electrons). This led to the work of Guglielmo Marconi of Italy, Lee De Forest of the U.S., and many others that laid the foundations of electronics engineering.

ELECTRICAL AND ELECTRONICS ENGINEERING FUNCTIONS

Research. In 1871 the University of Cambridge in England established the Cavendish professorship of experimental physics and built the Cavendish Laboratory, beginning a development that in the 20th century grew into a major trend. With James Clerk Maxwell as its first professor, this laboratory led in research in the 19th century. In the U.S., research and industrial investigation was organized in universities and in industrial research and development laboratories. Many advances, such as ductile tungsten for lamp filaments, practical radio tubes and equipment, long-distance telephony, high-voltage transmission of power, television, the transistor, and satellite communications have come from such organizations. Employing teams of engineers with varied backgrounds and working in well-equipped laboratories, they are able to approach problems fortified with knowledge and equipment not available to the individual inventor.

It is estimated that nearly half of the engineers in the electrical and electronics profession in the early 1970s contribute to research and development activities. Education for such work almost always requires study beyond the usual four-year degree in electrical engineering, mathematics, or physics.

Development. Following the work of the research team is the development and design engineer, who transforms the relatively incomplete models of the research laboratory into marketable products. The developer must plan the product with current production methods and facilities in mind. Often automated production machines are employed, in part to replace expensive hand labour and in part to perform operations beyond the capabilities of human workers. Thus, the design and development engineer must possess a theoretical knowledge sufficient to understand the results of the laboratory, plus a practical understanding of manufacturing operations.

Other functions. Production supervision, monitoring the product, and improving production methods are carried on by technicians, trained in applied electrical engineering. They work with the production force and may modify standard products to meet special needs.

Other engineers work in the field, providing technical guidance to customers, devising systems to meet specialized needs, and, in many cases, emphasizing the advantages of their products in a sales effort.

Opportunities in management constantly arise for those who demonstrate ability to administer, to supervise, to lead other workers in handling technical and corporate problems. Here, additional education in management, finance, and administration adds to the technical skills previously demonstrated by the young engineer.

Consulting or offering one's technical abilities to industries, municipalities, or other governmental bodies on a client–engineer basis is frequently undertaken by power generation and distribution engineers.

SUBDIVISIONS OF ELECTRICAL AND ELECTRONICS ENGINEERING

Expansion of field has been rapid

Electrical and electronics engineering has repeatedly expanded as fresh areas of knowledge were introduced following basic research efforts. Engineering had hardly begun in the field before the relatively simple Edison direct-current system of generation and distribution was challenged by the polyphase alternating-current system. Alternating current possessed considerable economic and industrial advantages but was complex for its day.

Advances in power-system design continued to the mid-1920s. Then radio broadcasting began, along with high-frequency radio transmission; engineers' knowledge was extended to cover the mysteries of the vacuum tube. Television, ready for commercialization after World War II, brought new electron-optical devices such as the kinescope and the scanning camera tube into use and required further broadening of knowledge and skills.

Advances in World War II

During World War II electronics engineers worked with still higher frequencies, such as radar and microwave radio. These systems, in which electromagnetic energy is transmitted through metal pipes, called wave guides, required engineers to extend their knowledge of electric fields and mathematics. As other detection systems were developed, such as sonar for underwater ranging with acoustic waves, electrical engineers became sound experts. Infrared light was used for invisible illumination of night scenes, calling for at least peripheral knowledge and skills in this area. Also developed were loran and shoran, systems for long- and short-range navigation. Equipment developed for these systems was later used in telephone and television microwave relay systems.

Even more sensational was the revolution brought about by the development of the transistor, first announced in 1948. This device led in the early 1970s to the era of integrated circuits or microcircuits, in which transistors, resistors, capacitors, and connectors are deposited by metallic evaporation or by diffusion of one material into another on silicon chips, often using only a fraction of a square millimetre per circuit. New techniques emerged, with whole circuits deposited in a monolithic array visible only with a microscope.

In the 1950s the development of the electronic digital computer brought a revival of the on–off coded signals, first used in telegraph, in a wholly new application. The computer does its work with small electrical circuits, carrying out the "and," "or," "not" operations of conventional logic. Besides use in scientific computation and in business systems, computer elements and functions appear in navigational equipment, such as that used to measure distances and calculate course in airliners and to provide inertial guidance for submarines and spaceships.

Able to build a circuit of almost any complexity, often with the help of a computer, the electrical engineer has now turned his attention to the overall performance of his systems more than to a study of the individual components. His ability to analyze complex interconnected systems has broad potential for application to social, environmental, and economic problems in the important field of systems engineering.

Power systems, a backbone element of all modern industry, continue as an area of importance. Power engi-

neers, planning the huge systems of modern metropolitan areas, must constantly bear in mind the numerous related social and environmental problems.

ELECTRICAL AND ELECTRONICS ENGINEERING EDUCATION

University education in electrical engineering began soon after the development of incandescent lighting and central power generation in 1882. During the 1890s engineering schools all over the world added studies in electrical engineering, often as optional courses within the mechanical engineering field. Academic departments of electrical engineering emerged around the turn of the century; over 190 departments of electrical engineering now exist in the U.S. In the U.K. the formal electrical education usually began within the physics area. In Germany, the field was often divided into "heavy current" and "light current," for power and electronics, respectively, and some schools established separate academic departments. This was also undertaken in India and Japan. In the U.S., there were educational leaders who felt that a fundamental electrical science background was needed for both areas and that specific applicational study should follow entrance into an industrial position. Splitting of academic departments thus did not occur widely, and most U.S. schools offer electrical education based on electrical theory, bolstered by considerable mathematics. Specialization follows at the postgraduate level or after acceptance of a post in industry.

Heavy- and light-current education

Departments usually offer work in the basic areas of circuit theory, electronic circuits, transistor and semiconductor theory, electromagnetic field theory, communication or information theory, and automatic control. This basic technical core is supported by extensive mathematical study and by work in economics, management science, social science, and the humanities.

It may safely be predicted that electrical and electronics study will continue to stress the mathematical and physical background of the field. Use of the computer in circuit and system design will grow. Graduate study will be undertaken by larger numbers. Interdisciplinary knowledge and skills will attain increased importance.

Modern electrical industries customarily offer opportunities for continued study, either in technical or management areas, after employment of the graduate, usually financing part or all of the expenses.

Technical institutes or community colleges offer programs of two-year duration, leading to Associate degrees. These programs are directed toward application, involve more direct use of equipment than do college degree programs, and usually employ somewhat less mathematics. Technician graduates directly or indirectly support engineers in developing, designing, testing, and installing equipment; they serve in a liaison function between engineers and skilled craftsmen. The Associate often moves into supervisory positions in production management or into technical sales work.

PROFESSIONAL ORGANIZATIONS IN THE FIELD

Early electrical industrial activities in England led to organization of the Institution of Electrical Engineers (IEE) in 1871. The American Institute of Electrical Engineers (AIEE) was founded in 1884. It undertook the dissemination of technical information concerning electrical power, illumination, and telephony through meetings and publications. The early wireless operators and engineers organized the Institute of Radio Engineers (IRE) in New York in 1912.

The expansion of the electronics field that occurred during and after World War II caused the IRE to grow phenomenally in size and importance.

The power field of electrical engineering and the radio and electronics fields overlap in electronic control of equipment in industry and in many military and aerospace fields. Leaders in the field long recognized that the members of the profession would gain if one organization represented all, and in 1962 the AIEE and the IRE merged into the Institute of Electrical and Electronics Engineers (IEEE). With headquarters in New York City, the organization serves members throughout the world. It fosters the development of electrical and electronics engineering and the allied sciences and operates an extensive publishing program with over 30 journals.

In Japan, Australia, and India there are additional organizations serving members in electrical power and in telecommunications. National organizations serve members in the countries of Europe as well. In the U.S.S.R. the radio and electronic interests of the profession are advanced by the A.S. Popov Scientific and Engineering Society of Radio Engineering and Electrical Communications, Moscow, so named to commemorate the achievements of Aleksandr Stepanovich Popov (1859–1905) in early radio research. The IEEE, in its worldwide activities, cooperates with all these national organizations.

Other societies covering portions of the field include the Illuminating Engineering Society (IES) and the Audio Engineering Society (AES), both in New York City.

THE FUTURE

With continuing new developments in the semiconductor field, with the computer industry having many opportunities for further expansion of use, with global communication by satellite now expanding, and with the power industry facing challenges raised by rapid growth in demand coupled with environmental encroachment, the electrical field does not lack unsolved problems. In the future it will continue in the development of solutions for its problems by application of well-founded theory.

BIBLIOGRAPHY. R.S. KIRBY *et al.*, *Engineering in History* (1956), a history of the engineering profession from the Greeks to the 1950s, subdivided by fields of work; J.R. WHINNERY (ed.), *The World of Engineering* (1965), a report on the achievements of the engineer; *Your Challenge in Electrical Engineering* (n.d.), a booklet prepared by the IEEE describing the work and career requirements of the electrical and electronics engineer; *Journal of Engineering Education* (monthly), a journal reporting on progress in engineering education at the university and technical institute level; S.H. RHINE and D.B. CREAMER, *The Technical Manpower Shortage: How Acute?* (1969), a survey of manpower supply and projected needs.

(J.D.R.)

Electricity

Electricity is the phenomenon associated with positively and negatively charged particles of matter at rest and in motion, individually or in great numbers. The properties of electrical charges, of moving charges that constitute currents, and of complex circuits of currents are understood in two complementary ways. The older view is in terms of theories and laws that have developed from experiments during the past two centuries in a steadily evolving knowledge. The other view is concerned with the space around the charged particle called the electric field, and knowledge of fields flowered in the latter half of the 19th century. Because every atom contains both positive and negative charged particles, electricity cannot be divorced from the physical properties of matter, which include such diverse phenomena as magnetism, conductivity, and crystal structure; from chemical reaction; or from the waves of electromagnetic radiation, which include visible light. The study of electricity thus involves the structure of matter and is intimately related to the apparatus with which electricity is studied. No one has ever seen an electrical charge; and until this century it was impossible to isolate single fundamental charges and study their behaviour in certain ways that take account of the charges on the probing instrument. Some of the information subsumed under electricity, therefore, was obtained through thought experiments; *i.e.*, through questions such as "What is the magnitude of the force acting between two charges?"

Electric and magnetic phenomena are inextricably connected with each other; moreover, most, though not all, of the devices employed daily rely on a combination of electric and magnetic effects. Thus, although this article is concerned principally with the subject of electricity, a complete and satisfactory treatment must also include topics on magnetism (*q.v.*).

The main reasoning of the article will be conducted in verbal terms, using only elementary mathematical no-

tions and a minimum of the differential and integral calculus; this will, however, be augmented by using vectors (quantities having direction and magnitude) whenever necessary and appropriate.

The most readily observable electrical phenomenon in a modern household is likely to be the illumination that follows the activation of a light switch. The simplest, or the most easily described, phenomenon of electrical origin is, however, static electrification, which occurs in a dry atmosphere when one walks across a carpet or removes clothing. The phenomena that accompany static electrification are diverse: the occurrence of sparks, minor sensory shocks, and, occasionally, the smell of ozone; but from the point of view of the fundamental properties of electricity, the most significant is the attractive force between two bodies electrified by friction. The whole science of electrostatics is founded on a quantitative investigation of this force and will be covered in the sections *Development of concepts and theories* and *Static electricity*. It will suffice here to note that static electrification is now interpreted as resulting from the transfer, by abrasion, of charged particles from one physical body to another. If the bodies are initially electrically neutral, they acquire equal and opposite charges and attract each other. As the two bodies are pulled apart, the work done against the attractive force increases their potential, or stored, energy; in some circumstances, this may be released later as a spark or an electric shock. This process works on a grand scale in the atmosphere, and the resulting discharge of electricity and dissipation of potential energy is manifest in lightning. It is possible, though some ingenuity is required, to devise machines to accumulate static electric charge and raise it to a high enough potential energy to produce other more useful and different electrical effects such as heating, chemical change, or magnetism. Before these effects are dealt with, however, there is one further aspect of static electricity that requires mention. If two bodies are electrified by friction, then placing other bodies between them alters and generally reduces the force that they exert on each other. In particular, if they are joined by a metal wire, a column of water or any wet object, a filament of graphite, or human tissue, by simultaneously touching each body with the fingers, the two bodies soon lose their electrification and no longer exert forces on each other. In the last case, an electric shock is experienced and the sensation is confined to the part of the body in the direct path between the two electrified bodies. This suggests that something flows between the two electrified bodies. The flow is now identified with the flow of electric charge.

Flow of charge between two electrified bodies

Charge can be transferred from a point A to a point B by moving a charged body initially at A to B; but the charge also can flow through a material body from a charged body at A to another body at B without any apparent motion of any of the bodies. Material media that have this property of transferring electrical charge are called electrical conductors, or, more usually, when spoken of in the context of electricity, simply conductors. All metals are conductors, and so are a number of nonmetallic media; *e.g.*, silicon, graphite but not diamond, water and a few other liquids, a number of organic and inorganic compounds (*e.g.*, lead sulfide), and, by virtue of their water content, most animal and plant tissues. Media that do not conduct electricity are known as insulators. Diamond, glass, dry paper, most hydrocarbons and stable inorganic and simple organic compounds, and, up to a point, most gases, are insulators. In the sections *Direct electrical currents* and *The conduction of electricity* below, the quantitative characterization of conductors and insulators is expounded by means of the resistance they offer to the flow of electricity. Resistance is described by their specific resistivity and, in the units used here, is expressed in ohm-metres. The specific resistivity of a superconducting metal is zero, that of a normal metal at room temperature about 10^{-7} ohm-metre and that of a good insulator 10^{17} ohm-metres. The enormous range of 10^{24} ohm-metres between a metal and a good insulator, or even the range of 10^{16} or so between a poor conductor and a good insulator makes it rather easy to classify media simply as conductors, perfect, good, fair, or bad and as insulators good, fair, or bad.

Other aspects of electrical phenomena are those associated with the flow of electric charge or currents. The familiar phenomena here are the production of heat as in an electric oven, the production of light either by heating a filament or through an electric discharge in a gas, the production of chemical change as in electrolysis, the production of mechanical work, either by the motion of an armature in a relay or in an electric motor, or the production of magnetic effects. All of these effects have converse effects, but whereas the conversion of heat or light directly to electrical energy in a thermocouple or photocell is usually not a practical process for producing electricity, the various forms of chemical battery and electromagnetic dynamo are prime means for generating electrical effects.

By not going too deeply into the fundamental nature of the processes involved, all of these phenomena can be described without reference to electrostatics. Setting aside the electromagnetic effects for now, the other effects can all be described in terms of the motion of an electric fluid, possibly against viscous friction from a region of higher potential energy to a region of lower potential energy. The name electric charge, or charge, is now given to this fluid; but it is a matter for experimental verification that charge in this context has the same properties and nature as charge in an electrostatic context. The verification rests on the observation that the same media conduct electricity in both contexts, and the same media are insulators in both contexts. It can then be demonstrated experimentally that electrostatic charge produces all the effects attributed to this electric fluid and that similarly the fluid produces all the effects attributed to electrostatic charge.

Magnetism was originally discovered as a phenomenon quite apart from electricity. One of the great turning points in the development of the physical sciences was the experimental discovery that electrical currents produce magnetic effects and that changing magnetic fields produce electrical effects. The whole structure of theoretical physics and most of modern electrotechnology rests on these observations. The magnetic effects of electric currents are seen at their simplest in the electromagnet. The forces exerted on conductors carrying currents, for example, drive the electric motor and the cone of a loudspeaker; the electric effects of changing magnetic fields comprise the basis for the operation of many types of dynamo; and both the magnetic effects of currents and the electric effects of changing magnetic fields are responsible for the operation of the transformer. The propagation of radio waves, radiant heat, light, X-rays, and gamma rays alike depends on this connection between magnetism and electricity.

Although electricity is generally associated with manmade devices such as electric motors, radio receivers, and so forth, electrical and electromagnetic effects play an important role throughout physics, chemistry, and biology. Two of the senses, sight and the perception of heat, depend either entirely or in part on the interaction of the eyes or skin with electromagnetic waves. Thus, much personal knowledge of the outside world is mediated by electromagnetic phenomena. The range and delicacy of perception can be increased by the use of optical or electronic instruments such as the microscope, telescope, radio telescope, infrared detector, gamma ray counter, and the simple galvanometer, thus extending man's knowledge of external phenomena. Electromagnetic laws also play an important role in observed phenomena. The behaviour of astronomic objects and the laws that determine cosmic structure depend to some extent on electric and magnetic fields. Biological processes (*e.g.*, muscle action, nerve transmission, and the activity of the brain) are accompanied by electrical effects; and these can often, as in electroencephalography, be detected directly and furnish useful information about the processes themselves. The nature and stability of the Earth's atmosphere, too, depend crucially on electric and magnetic effects, especially in the upper layers.

Sensory perception of electromagnetic waves

There is, however, a more fundamental way in which electromagnetism influences the whole of the material world. It is now known that matter consists of positively charged heavy nuclei and negatively charged electrons of small mass bound together as atoms. In some circumstances—*e.g.*, in a metal, a semiconductor, or an ionized gas—electrons are detached from atoms, leaving positive ions. These latter may either remain fixed in the material structure, as in a metal, or be free to move, as in a gas. The conduction of electricity in metals is caused by the motion of free electrons and in a gas by the motion of both electrons and ions. In some liquids (*e.g.*, water) one or more of the atoms in a molecule may lose an electron to another part of the molecule, yielding a mobile positive and a mobile negative ion.

Electrons detached from atoms can, in some circumstances, be drawn out of a material medium into vacuum and their motion controlled by electric or magnetic forces; the whole technology of electron tubes is based on this phenomenon, and much knowledge of the structure of matter is based on observations of the behaviour of free electrons in vacuum. The motion of electrons can also be controlled within matter. Technically, this control is easiest in a semiconductor such as silicon; a whole family of electronic devices from the simple rectifier to the most complex integrated circuits used in computers is based on this principle.

The force that binds electrons to nuclei in atoms is the electrostatic attraction manifest between two oppositely charged particles. Electrostatic force also plays some part in determining the structure of nuclei; but its most significant role is in the structure of the outer electronic part of atoms. The way the electrostatic force determines this structure is governed by the laws of quantum mechanics, which are used in working out much of theoretical atomic physics and theoretical chemistry. In the end, however, the crucial fact is that most of the physical properties and the whole of the chemical behaviour of atoms depend on the simple laws of electrostatic force acting between charged particles.

Electric effects determine the structure of matter, and, to an almost equally significant degree, electromagnetic theory determines thinking about the physical universe. The perfection and apparently almost universal validity of the electromagnetic laws and equations leads to theories of other phenomena by analogy with the theory of electromagnetism. Thus, for example, when Enrico Fermi, a physicist, came to consider the laws governing the beta decay of nuclei—*i.e.*, the emission of electrons from atomic nuclei—he set them up in a form as close as possible to the laws governing the emission of gamma rays, or electromagnetic radiation, by nuclei. There is one further way in which the highly developed theoretical structure of electromagnetism has influenced thinking and this is through the theory of relativity that first arose from the theoretical electromagnetic studies of physicists Henri Poincaré in France and Hendrik Antoon Lorentz in The Netherlands and the experimental observations of electromagnetic phenomena of numerous investigators around the turn of the century, of which the most important was the work of A.A. Michelson, a physicist, and Edward Williams Morley, a chemist. Both theory and experiment indicated that the velocity of light, an electromagnetic phenomenon, is a universal constant independent of any velocity possessed by either the observer or the source. Lorentz and Poincaré worked out many of the consequences of these facts, but it was left to Einstein to enunciate the special theory of relativity in a general form independent of electromagnetism. The fact that the velocity of light plays a unique and central role in the theory of relativity, which applies to all phenomena, indicates either that electromagnetism plays a crucial role in the structure of the universe or, at the very least, that electromagnetic phenomena reflect this structure in a particularly simple and direct fashion. Speculations about the nature of this connection extending well back beyond the formulation of the relativity theory have played, and still play, a dominant part in determining the development of fundamental physics.

Electromagnetism and cosmic structure

Of the sections of this article, the first is concerned qualitatively and nonmathematically with the general background to the study of electricity. The middle sections—II, III, IV, and V—develop the basic theoretical structure of electricity, beginning with the fundamental laws of electrostatics, and the final sections VI and VII, deal again in a mainly qualitative way with the most important electric effects. The middle sections necessarily involve a certain amount of mathematical manipulation and a knowledge of elementary algebra, and some familiarity with the basic notions of calculus will be required to assimilate parts of these sections. Vector notation is used in sections II and VI. The article as a whole is complementary to that on MAGNETISM, and the reader will find it useful to consult that article at several points. Because electricity is a subject with applications in almost every branch of science and technology, ELECTRICITY covers only a fraction of the topics that could be included under this subject heading. Only material that sets out or illustrates general principles has been included.

The article is divided as follows:

I. Development of concepts and theories
- Origin of concepts
- The rise of electrical theory
- Development of electrical technology
- Modern views of the nature of electricity and magnetism

II. Electrostatics
- Static electricity
- Dielectrics
- Electrostatic potential

III. Direct electrical currents
- Basic principles
- Steady currents, magnetic fields, and electromagnetic units
- Electromotive force
- Direct-current circuits

IV. The conduction of electricity
- The motion of charged particles in an electric field
- Conduction in vacuum
- Conduction in gases
- Electrolytic conduction
- Conduction in metals
- Conduction in semiconductors
- Contact potential and thermoelectricity
- Electron emissions

V. Alternating electric currents
- Faraday's law of electromagnetic induction
- The mathematical and graphical representation of sinusoidal phenomena
- The basic alternating-current circuit laws
- The detection and measurement of alternating currents and voltages
- Resonant circuits
- Alternating current bridges
- Nonlinear elements
- Electric waves in cables
- Filters
- Transient phenomena
- Eddy currents and the skin effect
- Power generation and distribution

VI. The primary effects and properties of electric fields and currents
- Magnetic effects of steady currents
- Magnetic effects of changing currents
- Force, energy, and power associated with fields
- Electromagnetic radiation

VII. The effects of electricity on matter
- Piezoelectricity
- Electro-optic phenomena
- Thermal effects
- Electrical production of light
- Chemical effects
- Bioelectric effects

I. Development of concepts and theories

ORIGIN OF CONCEPTS

Electrical, magnetic, and optical effects were known to the ancients, though little systematic investigation of them occurred until the late Middle Ages. Curiously enough, because electrostatics tends to be regarded as both the simplest and most fundamental branch of the whole subject, the first important investigations were made in magnetism, no doubt stimulated by the useful

and curious properties of the magnetic compass. There is some doubt as to where this device originated. It was known in England in the 12th century and was not, apparently, regarded at that time as a novelty. The first observation that has any bearing on modern notions of magnetism was made in the 13th century by Petrus Peregrinus de Maricourt, a French crusader, in Picardie who took a natural magnet, or lodestone, rounded into globular form and laid on it a magnetic needle and marked on the globule the line along which the needle set itself as it was placed on different parts of the globule. He found that when the whole stone had thus been marked the lines formed a set of meridians of longitude passing through two points at opposite ends of the stone, just as meridians of longitude on the Earth's surface pass through its north and south poles. De Maricourt called these points the magnetic poles and observed that the way a magnetic needle set itself when brought near the body depended only on the position relative to these poles. Thus, he is given credit for the notions of poles and polarization that were to dominate magnetic and electric theories for 700 years, and still play an important role in electromagnetism.

Lodestone investigation in the 13th century

The father of the modern subject of electromagnetism was William Gilbert of England. In his book *De Magnete magneticisque corporibus et de magno magnete tellure*, published in 1600, three years before his death, he not only gave a systematic discussion of magnetism, especially the magnetism of the Earth, but also considered the force between two objects charged by friction. To this force he gave the name electric. He showed that frictional electricity occurs in many common materials and he also noted what is still regarded as one of the primary distinctions between magnetism and electricity: the force between magnetic objects tends to align the objects relative to each other and is little affected by most intervening objects; the force between electrified objects is primarily a force of attraction or repulsion between the objects and is grossly affected by intervening matter.

Gilbert attributed the electrification of a body by friction to the removal of a fluid, or "humour," which then left an "effluvium," or atmosphere, around the body. The language is quaint but if the "humour" is renamed "charge" and the "effluvium" renamed "electric field," Gilbert's notions closely approach modern ideas. The 17th and early 18th centuries saw an increased interest in electrostatic effects. Electrical conductors were discovered by Stephen Gray in England in 1729, and in 1733, Charles François de Cisternay Dufay, gardener to the king of France, announced the existence of two types of electric force: an attractive force, already known, and a repulsive force. These observations were organized and systematized by Benjamin Franklin, who essentially attributed the plus sign (+) and the minus sign (—) to the two types of electricity and enunciated the law now known as the conservation of charge (the net sum of the charges within a region is always constant). Sir Isaac Newton's work on gravity had surprisingly little influence on the development of electrical theory before the middle of the 18th century. There had been some speculation as to the law of force between charged bodies, but this was not elucidated until 1767, when Joseph Priestley, an English chemist, acting on a suggestion of Franklin's, showed that when a hollow metal (conducting) vessel is electrified no electric force acts on objects in its interior. He inferred that "the attraction of electricity is subject to the same laws with that of gravitation and is therefore according to the (inverse) squares of the distances." With this discovery, subsequently verified in other ways by many experimenters, the laws of electrostatics were complete and have not further been changed to this day, although the years between Priestley's discovery and the middle of the 19th century saw a great clarification of the expression of the laws and substantial mathematical advances in handling them. The advances are mainly associated with two physicists, Henry Cavendish and Charles-Augustin Coulomb, and a mathematician, Poisson, during the 18th and early part of the 19th centuries. Poisson's equation, published in 1813, and the law of charge conservation virtually contain in two lines the whole of the laws of electrostatics.

While the science of electrostatics had been progressing and its fundamental laws were becoming clear and codified, the study of current electricity, or "galvanism," had begun. Though Luigi Galvani was a professor of anatomy in Bologna, he also had an interest in electrical (*i.e.*, electrostatic) phenomena and acquired several electrical machines that apparently he kept and investigated in his dissection room. The sequence of events that followed from this coincidence is not entirely clear, but certainly Galvani noticed that electrical sparks produced by a nearby machine caused muscular contraction in a dissected frog. He went on to investigate this effect and at first considered that it was the result of atmospheric electricity because similar effects could be observed during lightning storms. In this he was almost certainly correct, but a second happy chance led him to an altogether more important discovery—that whenever a connection was made by a metal connector between the frog's muscle and the frog's nerve the same muscular contraction was observed, although some metals appeared to be more effective than others. Galvani concluded that a fluid that he identified with electric charge was transported by the metal from the nerve to the muscle and caused the contraction. Galvani's results, published in 1791, excited considerable controversy and speculation.

Galvani's experiment with frogs

A more serious attitude to Galvani's discoveries was taken by a physicist, Alessandro Volta of Pavia, who in 1792 correctly conjectured that the effect was caused by the contact between metal and a moist body. By 1800, Volta, pursuing this line of thought, had constructed what is now known as a voltaic pile consisting of layers of copper, moist pasteboard, and zinc (repeated in that order) and had discovered that, with an extended pile, he could obtain an electric shock by touching the first copper plate and the last zinc plate. Volta confirmed that the effects of his pile were in every way equivalent to those of static electricity. Within 20 years, galvanism had been unequivocally related to static electricity.

Heating effects caused by electrostatic discharges had been known for some time, and there had even been sporadic reports of iron objects becoming magnetized when struck by lightning. But the advent of a sustained and reliable source of electric current made more systematic investigation possible.

THE RISE OF ELECTRICAL THEORY

In 1820 a Danish physicist named Hans Christian Ørsted announced his discovery that a magnetic compass needle was influenced by currents in neighbouring conductors; and by the end of the year a French physicist, André-Marie Ampère, had shown that two parallel wires carrying current attracted each other if the currents flowed in the same direction and repelled each other if the currents were opposed. The first quantitative analysis of Ørsted's results was made by two French physicists, Jean-Baptiste Biot and Félix Savart, also in 1820, and in 1825 Ampère essentially formulated in mathematical terms the laws that govern the interaction of steady currents with static magnetic fields. The final simplification of these laws and their expression in the form now used, however, was not achieved until much later in the century.

The power of a material to conduct static electricity had been investigated earlier (1775) by Cavendish, but the final clarification of the laws was effected by a physicist in Germany, Georg Simon Ohm, and Ohm's law (see below), published in 1827, still forms part of the conceptual structure—and perhaps the most widely used part—employed in practical electrical design. The relation between electric currents and heat, although known to earlier investigators, was not investigated quantitatively until somewhat later (1841) by James Prescott Joule, a physicist who formulated in more or less its final form the theory of the heating effects accompanying the flow of electricity in conductors.

In 1813 the chemist Sir Humphry Davy appointed Michael Faraday as his assistant at the Royal Institution

of Great Britain, London. Faraday is primarily remembered for his discovery of electromagnetic induction and the introduction of the concept of a line of electric force or magnetic force. The discovery of electromagnetic induction (the production of a transient current in a conductor when the current in a nearby conductor is changed) was announced in 1831. But the concept of lines of force that permeated Faraday's thinking has influenced physics since Faraday's time and profoundly impressed the Scot James Clerk Maxwell, born the same year (1831), when he began his researches that culminated in 1864 with the formulation of his electromagnetic equations.

Discovery of electromagnetic induction

While Faraday was engaged in his experimental researches, a parallel development in the mathematical theory of electricity and magnetism was proceeding on the continent of Europe, mainly in Germany. To this period belongs the work of the German physicists Franz Ernst Neumann, Wilhelm Eduard Weber, and H.F.E. Lenz. At the same time, Joule and two other physicists, Sir William Thomson (later Lord Kelvin) of England and Hermann von Helmholtz of Germany, were clarifying and developing the theory of the relation between electricity and other forms of energy; and Helmholtz, Thomson, Joseph Henry, a United States physicist, Gustav Kirchhoff, a German physicist, and Sir George Gabriel Stokes, a British physicist, were extending the theory of the conduction and propagation of electrical effects in conductors. In 1856, Weber and his German colleague, Rudolf Kohlrausch, determined the ratio of the electric and magnetic units and found that it is almost exactly equal to the velocity of light and of the same dimensions. In 1857, Kirchhoff used this result to demonstrate that electrical disturbances propagated on a perfectly conducting wire with the speed of light. Thus by 1860 the connection between electromagnetism and optics was apparent. The final steps in establishing the connection between light and electromagnetism were made by Maxwell and, with the formulation of his celebrated equations, the history of classical electromagnetism comes to an end. Maxwell's equations have not yet been superseded, and subsequent developments in electromagnetic theory have either been concerned with the relation between electromagnetism and the atomic, or submicroscopic structure, of matter or with exploring the practical and theoretical consequences of Maxwell's equations.

Maxwell's greatest distinction is that he finally synthesized electricity, magnetism, and optics into one coherent whole. His electromagnetic equations will be treated in detail below, and it need only be noted here that whereas Faraday had discovered experimentally that a changing magnetic field produces an electric field, Maxwell hypothesized that a changing electric field will produce a magnetic field. With this hypothesis, Maxwell rounded out the structure of electromagnetism and brought optics and radiant heat within its purview. More recent developments, such as relativity and quantum mechanics, have done little to upset the fundamental correctness of Maxwell's formulation. Indeed, as experiments confirm the predictions of quantum electrodynamics based on Maxwell's equations with ever-increasing accuracy, confidence in their validity increases rather than diminishes. The notion of a field, introduced by Faraday and expressed succinctly in mathematical form by Maxwell, continues to dominate 20th-century thinking.

With the formulation of Maxwell's equations, physicists were equipped with a complete description of the electromagnetic field, although, as might be expected, Maxwell's ideas did not win general acceptance until the end of the 19th century. The primary focus of interest in physics shifted from electromagnetic effects to their origin in the structure of matter.

Electrical discharge in gases

By the middle of the 18th century, luminous electrical discharges in rarefied gases had been observed; but little notice of the phenomenon was taken until, in 1838, Faraday gave the first exact description of the effects that occurred and noted a dark space between the glowing negative electrode and the main bulk of the discharge. Faraday remarked on the significance of the difference between negative and positive electricity, but little further work was done on the subject until the invention of a mercurial air pump made low pressure investigations easier. Between 1850 and 1895 the properties of cathode ray discharge were subjected to almost continuous investigation, and a mass of apparently contradictory and obscure evidence accumulated. Between 1896 and 1898 two English physicists, Sir J.J. Thomson and Sir J.S.E. Townsend, identified a cathode ray as a stream of negatively charged particles now called electrons, much lighter than the lightest atoms of chemistry.

From the discovery of the electron to the present day the theory of electromagnetism has been inextricably entangled with theories of the atomic, subatomic, and subnuclear structure of matter, one theory influencing the other and vice versa.

The thermodynamics of thermal radiation had been investigated in Germany by the physicist Wilhelm Wien between 1890 and 1900. Wien had virtually exhausted the resources of thermodynamics in dealing with this problem. Two British scientists, Lord Rayleigh (John William Strutt) and Sir James Hopwood Jeans, had by 1900 applied the newly developed science of statistical mechanics to the same problem. They obtained results that, though in agreement with Wien's thermodynamic conclusions (as distinct from his speculative extensions of thermodynamics), were only in partial agreement with experimental observations. The problem had also attracted the attention of Max Planck, a German physicist, who attempted to combine both the statistical approach and a thermodynamic approach. By concentrating on the necessity of fitting together the experimental data he was led to the formulation of an empirical law that satisfied Wien's thermodynamic criteria and fitted the data. When Planck interpreted this law in terms of Rayleigh's statistical concepts, he came to the conclusion that radiation of frequency (ν) exists only in quanta of energy, $nh\nu$, in which n is a whole number and h is a new universal constant. Planck's result, announced in December 1900, marked the foundation of quantum mechanics and initiated a change in the basic structure of physical theory that was as profound as anything that had happened in the previous century.

By 1900 it was apparent that Thomson's electrons were a universal constituent of matter, and in the early years of the century many physicists attempted to construct theories of the electromagnetic properties of metals, insulators, and magnetic materials in terms of electrons. The most notable landmark in this period was the publication of *The Theory of Electrons and its Applications to the Phenomena of Light and Radiant Heat* by Lorentz in 1909 (based on his lectures in 1906). Some of Lorentz's ideas have been modified in the light of quantum theory, and some of his ideas have been put into more precise form in the decades that have elapsed since 1909; but just as Maxwell unified the theories of electricity, magnetism, and optics, Lorentz unified their interpretation in atomic terms.

The one other major conceptual advance in electromagnetic theory that remains to be discussed is the theory of special relativity. In Maxwell's time a mechanistic view of the universe was popular; and just as sound was interpreted as an undulatory motion of the air, light and other electromagnetic waves were regarded as an undulatory motion of an intangible medium, the ether. The question then arose as to the velocity of the Earth through the ether and whether or not the velocity of light measured by an observer moving relative to the ether would be affected by his motion. Michelson and Morley in 1887 had shown that no such effect occurred, and Lorentz and Poincaré in the period 1900–04 had shown that this was consistent with Maxwell's equations. On this basis they built a theoretical construct, the theory of relativity, in which the absolute motion of a body relative to a hypothetical ether is no longer significant. The theory was first given the name the Principle of Relativity by Poincaré in a lecture at St. Louis, Missouri, in September 1904; in 1906, Planck gave the first formulation of relativistic dynamics. The most general formulation of the principle of

Relativity

relativity was, however, given by Einstein in 1905; and it is with his name that the principle is usually associated.

Maxwell's equations, the special theory of relativity, the discovery of the electronic structure of matter, and the formulation of quantum mechanics had all occurred by 1930. The triumphs of quantum electrodynamics in the era 1945–55 and later added little to a fundamental understanding even if they immeasurably increased confidence in the general validity of electromagnetic, atomic, quantum, and relativity theory.

DEVELOPMENT OF ELECTRICAL TECHNOLOGY

Some attention must now be given to the development of electrical technology. This period may be said to have begun with Faraday's discovery of electromagnetic induction, which gave a possible means of turning mechanical energy into electrical energy on a large scale.

The first moderately efficient battery (the Daniell cell) dates from 1836; and the forerunner of the modern dry battery (the Leclanché cell) was first announced in 1866. The lead–acid accumulator did not appear until 1881, six years after arc lighting energized by a motor generator had been installed in the Gare du Nord.

Within a year of Faraday's discovery a small hand generator was demonstrated in Paris, and by 1850 generators were being manufactured in several countries and experiments with carbon arc lights for lighthouses were afoot in England and France. The first dynamos were little more than an assembly of coils and permanent magnets that could be maintained in relative motion. The slotted armature appeared only in 1880, although between 1860 and 1870 the principle of the self-excited generator was discovered. The early years of the electrical industry were thus dominated by the problems of generation, and the only serious use of electric power considered was the arc light. Faraday's discovery of the transformer was not put to practical use until the 1880s, and at about this time the question of whether alternating current (ac) or direct current (dc) should be used for power transmission became a live issue.

The carbon arc lamp was never suitable for domestic use and required regular attention, so that it could hardly ever have led to a large scale demand for electricity. The incandescent filament lamp, originating in the 1840s, became a practical device (1878) with a carbon filament lamp invented by an English physicist, Sir Joseph Wilson Swan, and was independently discovered within the year by Thomas A. Edison of the United States. (Litigation between Swan and Edison was finally resolved by the formation of a joint company.) The tungsten filament lamp introduced in 1905 has become the main form of electric lamp, although in the last 20 years it has gradually begun to give way to the more efficient and longer-lived, but more expensive, fluorescent gas discharge lamp.

Simple and robust instruments for measuring electric charge, current voltage, and power have been in an almost continuous state of development since Faraday's time. At the present time a considerable fraction of the output of the electrical industry, and especially of the scientific instrument industry, is concerned with the production of electrical measuring instruments of ever-decreasing size and ever-increasing reliability and accuracy. Most major countries maintain a central standards institute, one of the major activities of which is the maintenance and establishing of electrical standards.

Apart from lighting and, in some countries, space heating, the other major use of electric power is in electric motors. The electric motor is the standard domestic prime mover in hand tools, cleaners, kitchen equipment, fans, boiler pumps, in some modes of transportation, and the like; in manufacturing, it has virtually eliminated all other forms of prime mover.

First commercial electric motors

The principle of the electric motor was devised by Faraday in 1821, but no motor of commercial significance was produced until 1873; and the first important ac motor, the creation of a United States inventor, Nikola Tesla, was not demonstrated in the United States until 1888. Tesla produced his first motors in association with the Westinghouse Electric Company (now Westinghouse Corporation) a few years after direct current motors had been installed in trains in Germany and Ireland. By the end of the 19th century the electric motor had taken a recognizably modern form. Subsequent improvements have rarely involved radically new ideas; but as design experience has accumulated and new and better bearing, armature, magnetic, and contact materials have become available, motors have become smaller, cheaper, more efficient, and more reliable.

The Leclanché cell of 1866 and the lead–acid accumulator of 1881 still dominate the battery market, although improved out of all recognition both in size, efficiency, and convenience. Other chemical cells are used for special purposes, and in recent years there has been some resurgence of interest in batteries. It has been prompted partly by the increased use of portable electronic equipment and partly by the search for a fuel cell that will convert the chemical energy of a readily available fuel, such as oil, directly to electricity without going through an intermediate, and necessarily inefficient, steam-raising thermal process.

Although not necessarily a major user of electric power in terms of quantity, the electronics and communications industries and services are perhaps one of the most spectacular products of the electrical age. By 1866 telegraphy over wires—using simple electrochemical or electromechanical receivers—was a well-established technique, and in 1850 a submarine cable was laid across the English Channel. This became unserviceable in a matter of weeks, and not until 1865 was a further and successful attempt made. Alexander Graham Bell's telephone was patented in the United States in 1876, and within a few years the first public telephone services began to take root. In 1895 Sir Ernest Rutherford, a British physicist, transmitted radio signals for three-quarters of a mile at Cambridge, England, and by 1901 Guglielmo Marconi had established communication across the Atlantic. Radio telephony, the electrical recording and reproduction of sound, and television had to await the development of the thermionic valve, or electron tube; and public radio transmissions were not established before the 1920s.

In a remarkably short time the electron tube had grown from Thomson's discovery of the electron. An English physicist, Sir John Ambrose Fleming, invented the electron diode in 1904 and Lee De Forest, an inventor in the United States, the triode in 1907, with which weak electrical signals could be amplified. By the end of World War I electronics based on the triode was an established industry. The first radio telephony transmission—from Arlington, Virginia, to the Eiffel Tower in Paris—was made in 1915; and a commercial radio telephone service opened between New York City and London in 1927. The years between the wars were mainly concerned with the expansion of the radio and the phonograph entertainment industries and the sound film industry; but, especially in England, rapid progress was being made toward the transmission of moving pictures; and just before the outbreak of World War II the British Broadcasting Corporation inaugurated the first public television service.

Until 1939 the electronics industry had been almost solely concerned with communications and broadcast entertainment, but during the 1930s work had started in England on a device for detecting aircraft now known as radar. Although the system is important in itself, for both its military and its civil uses, it also marked the beginning of an electronics industry not entirely dependent on communications and entertainment. During the war and in the years that followed, the electronics industry made strides paralleled only by those of the chemical industry. Television has become a commonplace; and new electronic gadgets and systems, of which the most significant is the electronic digital computer, have appeared. For many purposes the transistor or semiconductor triode, invented in the 1940s at Bell Telephone Laboratories, has replaced the electron tube and has led to a notable diminution in the size, cost, and failure rate of electronic equipment.

In the next section, the historical approach is abandoned and a brief summary is given of the modern view of electricity, outlining some of the ideas to be developed at greater length in later sections.

MODERN VIEWS OF THE NATURE OF ELECTRICITY AND MAGNETISM

The question "What is electricity?" like the question "What is matter?" really lies outside the realm of physics and belongs to that of metaphysics; but the questions "How does electricity fit into the general understanding of the physical universe?" or "How is electricity related to other physical phenomena?" are proper in the context of physics and must be considered.

Importance of the notion of a moving electric charge

Electromagnetism. Electromagnetism, the theory dealing with electric and magnetic fields, is based on the concept of a moving electric charge. The way this arises is as follows: If a charged macroscopic body is moved, the charge is transferred from one point in space to another, and the motion leads to a current. Conductors can also transfer charge without apparent bodily motion. It is logically possible for charge to flow without mass motion, but the study of processes on an atomic or sub-atomic scale shows that charge transfer is always accompanied by the motion of particles of finite mass. Thus, charge transfer is associated with the motion of electrons, protons, nuclei, ions, or a variety of unstable particles such as mesons. A charge may therefore be assigned to these particles. Experimentally, each particular type of particle is always associated with the same charge; and this charge is always an exact integral multiple of a fundamental unit e, about 1.6×10^{-19} coulomb (the unit of electrical charge). The only two unconditionally stable elementary particles with a finite mass are a proton with charge $+e$ and an electron of charge $-e$. Neutrons (elementary particles of about the same mass as protons but with zero charge) are unstable except when bound in nuclei. Composite systems of neutrons (*i.e.*, nuclei, atoms, and ions) may have charges that are other multiples of e. Thus, for example, the oxygen nucleus has a charge $+8e$, the chloride ion a charge $-e$, and the phosphate ion (PO_4) a charge $-3e$. The general dimensions of nuclei, electrons, and protons are 10^{-13} centimetre; but atoms and ions are usually 100,000 times larger (10^{15} times greater volume) and about 10^{-8} centimetre in linear dimensions.

Today the flow of an electric current is considered to be caused by the motion of one or another of these types of charged particles. Because the electron, of mass about 10^{-30} kilogram, is 2,000 times lighter than any ion, it responds more rapidly to electric fields; and most of the phenomena grouped together as conduction are caused primarily by the motion of electrons. In metals and semiconductors this is exclusively the case; but in a gas discharge both electrons and ions are important, and in an electrolyte (an electrically conducting compound) such as water only ions are responsible for conduction. Both electrons and ions also lead to important effects in insulators, and electronic motion is entirely responsible for the phenomenon of magnetism.

A conductor always contains mobile charges, usually electrons, but in an insulator the charges are bound to fixed positions in the lattice (the atomic sites in a crystal) if it is a solid, or in exactly neutral groups, atoms, or molecules, if the insulator is a liquid or gas. Thus, in air, each oxygen molecule contains 16 protons and 16 electrons and has no net charge, whereas in a sodium chloride crystal, although the individual sodium (Na^+) and chlorine (Cl^-) ions are charged, they are tightly bound in position in the crystal lattice. When a charge is brought near an insulator and its electric field exerts a force on the charges in the insulator, the most these charges can do is to perform minute excursions of less than atomic dimensions from their equilibrium positions. Thus, if the external charge is positive, the electrons in atoms will move slightly toward it and the positive nuclei will move away from it. In sodium chloride the motion of the ions as units is superimposed on this effect. The displacements may be small, but the large number of displaceable charges in unit volume leads to gross and easily observable effects. In the theory of dielectrics, to be dealt with in more detail in a section below, these displacements lead to an effect called polarization of the medium, so that if two charges of opposite sign are placed on either side of an insulating body, the net effect is a reduction in the force acting between the two charges. The general theory that emerges from this picture of the behaviour of insulating bodies is in accord with facts.

Originally, magnetism was regarded as an effect completely distinct from electricity, but the discoveries of Ørsted, Ampère, and Faraday showed that in fact the two effects are closely connected. Further, all the effects associated with magnetized bodies can be reproduced exactly by systems containing circulating electric currents. Thus, there is no need to postulate a separate magnetic source. This is the modern view, and all magnetic effects are now attributed either to actual macroscopic currents in circuits or to microscopic currents circulating within atoms. In its most naïve form this theory would identify the atomic currents with the effects of electrons in atoms rotating in their orbits about the nuclei. Neither this picture nor its simple transposition into quantum mechanical terms leads to complete agreement with experiment, and, in the early days of quantum mechanics, this was evaded by introducing the concept of the spinning electron. The electron itself was given an internal degree of freedom associated with internal rotation, angular momentum, and a circulation of internal current. This led to exactly the additional magnetic properties actually observed. The origin of electron spin became apparent when P.A.M. Dirac, an English physicist, gave the first complete relativistic and quantum mechanical treatment of the electron.

Relation of electric and magnetic effects

Charged particles and Maxwell's equations. If the phenomena of electromagnetism are explained as the consequences of the motion of atomic or sub-atomic charged particles, it is appropriate to look briefly at the nature of the charged particles themselves.

The electron is characterized by its electric charge and mass, which are regarded as its most significant properties. Its spin about its own axis and magnetic moment, the same magnetic property possessed by a bar magnet, are connected with these properties; but that they are not its only properties can be seen by considering the basic reaction of beta decay in which a neutron (n) of zero charge decays, or transforms, to a proton (p), an electron (e), and an antineutrino ($\bar{\nu}$), a massless particle that moves with the speed of light, symbolized $n \rightarrow p + e + \bar{\nu}$. The electron, by virtue of its negative charge, generally interacts with a positively charged proton; but here it interacts in a quite different way, with not only a proton but with two uncharged particles, the neutron and the massless antineutrino. This is not the only additional interaction in which electrons engage, and the proton has a whole series of even more complex interactions. Thus, although electric charge fits into a general scheme of fundamental interactions, the relation is unlikely to be simple. At present, the nature of this scheme is somewhat of a mystery. The only hard facts are that nature provides a multitude of charged and uncharged particles, some of which have a finite mass and others of which are, like the neutrino, both chargeless and massless. There are only two stable particles with mass, the proton and the electron; and these both have the same spin, one-half Planck's constant divided by 2π (symbolized $\frac{1}{2}\hbar$), and charges plus and minus e ($\pm e$). The proton, however, has some 2,000 times the mass of the electron.

It is known with enormous precision that the proton and electron charges are numerically equal, for otherwise normal matter could not be electrically neutral, but no more fundamental reason why this should be so is known; nor, despite almost continuous speculation for the last 50 years, is it known why the fundamental unit of charge should have exactly the value it does. Thus, although electromagnetic phenomena fit into the same general scheme as the structure of matter, the ultimate laws that determine this scheme are as yet undiscovered.

Maxwell's equations of the electromagnetic field relate

two vectors (quantities characterized by both magnitude and direction), the electric field E and the magnetic field B, to electric charges and currents, which act as the sources of the field much as a running tap acts as a source of flow lines in a bathtub, or an open drain acts as a sink of flow lines. Theorists typically seek symmetrical theories, and in recent years there has been some discussion as to whether magnetic charges and currents could not also appear in the field equations. Dirac, the founder of elementary particle theory, originated the idea in 1931, and certainly it has considerable aesthetic appeal. Calculations have indicated that if free magnetic charges are to exist they must be associated with massive particles and could be created only in particle reactions at energies well above anything at present available. Thus, to date, there is no experimental evidence whatsoever that supports the inclusion of magnetic charge in the field equations. As far as is now known, Maxwell's equations are a complete description of the electromagnetic field and its interaction with matter. There are problems in combining the field equations with the laws of relativistic quantum mechanics. Contemporary quantum electrodynamics is a scheme that allows the calculation of many experimentally measurable effects with extreme accuracy but leaves the general problem of combining a continuous field theory with particle mechanics unresolved.

Significance of Maxwell's equations

II. Electrostatics

Electrostatics is concerned with the relation between three quantities: electric field, charge, and force. One of these quantities, charge, has magnitude but no direction associated with it; the other two involve both magnitude and direction and are called vectors, being represented by boldface characters (*e.g.*, E for electric field and F for force). Vector notation is a compact, simple language that expresses important physical results in a form brief enough to be memorable and also taken in at a glance. As an example, an important electromagnetic law, Faraday's law of electromagnetic induction, states that the total electromotive force (potential difference) acting around any closed path in space is equal to the rate of decrease of the magnetic flux (a measure of the magnetic field taken as lines per unit area) linking the path. This law can be expressed in mathematical terms without using vectors. A simple formulation is one that relates the partial derivatives (symbolized by the delta sign: ∂) of the Cartesian components (a component is the value of a quantity along one of three possible axes, usually denoted x, y, or z) E_x, E_y, and E_z of the electric field to the time derivatives of the Cartesian components B_x, B_y, and B_z of the magnetic field. This gives the three equations, each having three partial derivative terms:

Law of electromagnetic induction

$$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = -\frac{\partial B_x}{\partial t}$$

$$\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} = -\frac{\partial B_y}{\partial t}$$

and

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -\frac{\partial B_z}{\partial t}.$$

In vector notation, however, the law is completely expressed by the equation

$$\nabla \times E = -\frac{\partial B}{\partial t}.$$

In this equation, $\nabla\times$ represents a mathematical operator, called the curl, which, when acting on the vector E, produces the three partial differential equations above. Clearly, this vector notation, called curl E, is both memorable and more easily recognizable. It is also more useful than the purely verbal statement because it can be manipulated, using mathematical techniques, and can be used in conjunction with vector statements of the other electromagnetic laws to derive useful practical results.

STATIC ELECTRICITY

The basic phenomena and laws. If two objects are rubbed together and then drawn apart it sometimes happens, especially in a dry atmosphere and with dissimilar materials, that the bodies appear to exert an attractive force on each other. They are said to be electrified. The force between the bodies decreases rapidly as they are pulled apart and generally appears to decay with the passage of time. Further, if two similar bodies (*e.g.*, two plastic rods) are electrified by rubbing against the same body (*e.g.*, a dry silk handkerchief) both rods attract the handkerchief but the two rods repel each other. Electrification leads to both attractive and repulsive forces. The electrification of a body can be transferred to another body by contact, and two bodies electrified by contact with the same body repel each other. A further important qualitative observation is that electrification can be transferred from an electrified to an unelectrified body by connecting the bodies by a metal wire. By themselves, these observations are not enough to lead to a quantitative theory but they suggest that, in frictional electrification, a "fluid" can either be transferred by actual contact or can flow through metals. Historically the elucidation of the laws of electrostatics, the interaction of charged bodies at rest, was achieved indirectly, and following the historical path does not help to explain these laws as well as might be hoped. The basic laws are therefore better stated in terms of a set of hypothetical experimental observations. These observations can be made with reasonable precision, though ultimately the validity of the electrostatic laws rests on the precision with which they predict the results of less direct experiments. The experimental facts may be set out in the form of five laws.

Electrification

1. If A and B are two electrified bodies, the force acting between A and B is along the line joining A and B, and its magnitude varies inversely with the square of the distance between A and B.
2. Action and reaction balance. Body A exerts a force on body B equal in magnitude to the force B exerts on A but in the opposite direction.
3. If body A exerts a force F_{AC} and body B exerts a force F_{BC} on a test body C at the same distance from A or B, and F_{AD} and F_{BD} are the forces exerted on a second test body D at the same distance, then $F_{AC}/F_{BC} = F_{AD}/F_{BD}$. These three observations make it possible to assign to each body a numerical measure of electrification, the charge q, which determines both the force exerted on, and the force experienced by, a body. If repulsive forces are defined as positive and attractive forces as negative, the force (F) acting between bodies of charge $q(A)$ and $q(B)$ at a distance r apart, in accordance with experimental fact (1) above, can be written as equal to a universal constant k times the product of the two charges divided by the square of their separation distance, or,

$$F = k\frac{q(A)q(B)}{r^2}.$$

If the charges have the same sign the force is repulsive ($+F$ in the convention used here). The interaction of two bodies with a third leads to a definition of equality of charge. The force between two bodies of the same charge is equal to the constant k times the ratio of the charge squared to the separation squared, or, $F = k\,q^2/r^2$. If r is expressed in metres and F in newtons, the unit of force, the choice of the constant k determines a unit for charge q. In the mks (metre-kilogram-second) or SI (International System of Units) scheme of units the constant k is equal to $\frac{1}{4\pi\epsilon_0}$ in which the constant ϵ_0 is defined as $\frac{10^7}{4\pi c^2}$, and the constant c is the experimental value for the speed of light. Further, ϵ_0, called the dielectric constant *in vacuo*, is regarded as having dimensions (*i.e.*, not a pure number but is expressed in units). The reason for this choice of constant k is connected with the desirability of defining the units of charge q in terms of precise magnetic measurements rather than less precise electrostatic measurements. As for other systems of units, in electrostatic units (esu), or Gaussian units (named after

Other systems of units

a German mathematician, Carl Friedrich Gauss) the constant k is equal to unity (one), with r and F of the force equation in centimetres and dynes. In Heaviside–Lorentz units $k = \frac{1}{4\pi}$, and in electromagnetic units (emu) $k = \frac{1}{c^2}$ with r in centimetres and F in dynes (a unit of force, 10^{-5} newton).

Force, because it is a directed quantity, by definition is a vector. If there is a charge $q(1)$ at a point specified by the vector $\boldsymbol{r}(1)$ and another, $q(2)$ at point $\boldsymbol{r}(2)$ (see Figure 1), the force $\boldsymbol{F}(1,2)$ acting on $q(1)$ and arising from $q(2)$

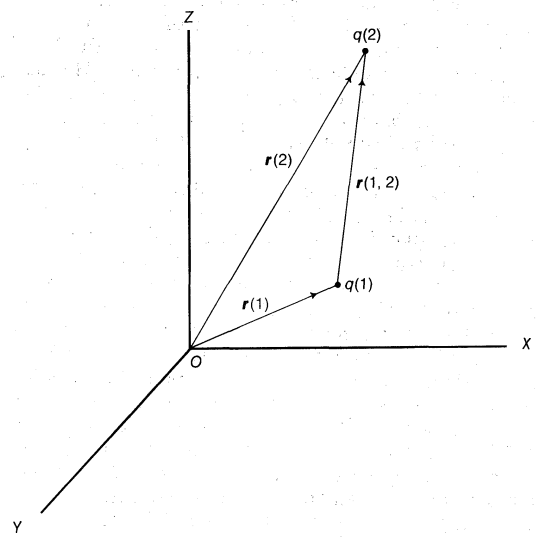


Figure 1: Vector addition $\boldsymbol{r}(2) = \boldsymbol{r}(1) + \boldsymbol{r}(1, 2)$ in Cartesian coordinate system OX–OY–OZ, showing that vector $\boldsymbol{r}(2)$ is the resultant of the sum of vectors $\boldsymbol{r}(1)$ and $\boldsymbol{r}(1, 2)$ (see text).

can be expressed in terms of the vector separation $\boldsymbol{r}(1,2) = \boldsymbol{r}(2) - \boldsymbol{r}(1)$ of the charges as the equation

$$\boldsymbol{F}(1,2) = \frac{q(1)q(2)}{4\pi\epsilon_0 r^3(1,2)}\, r(1,2),$$

in which $r(1,2)$ is the magnitude, or length, of the vector $\boldsymbol{r}(1,2)$. (In this case, length r in the denominator is cubed because it is also introduced into the numerator as a first power in order to make a vector equation.)

4. All forces acting on a single charge are additive. If charges $q(1)$ at $\boldsymbol{r}(1)$, $q(2)$ at $\boldsymbol{r}(2)$ etc., up to $q(n)$ at $\boldsymbol{r}(n)$, act on a charge $q(0)$ at $\boldsymbol{r}(0)$ the force on this charge is

$$\boldsymbol{F}(0) = \boldsymbol{F}(0,1) + \boldsymbol{F}(0,2) + \text{etc.}$$

and this can be expressed as

$$\boldsymbol{F}(0) = q(0) \sum_n \frac{q(n)\boldsymbol{r}(0,n)}{4\pi\epsilon_0 r^3(0,n)}, \qquad (1)$$

in which $\sum_n$ means the sum of terms, each of which has n equal to a whole number 1,2,3, etc. The results so far give a means of defining and measuring charge.

5. The final law of electrostatics is that charge is conserved. Before investigating this law it will be convenient to develop a little further the ideas implicit in equation (1) and to examine the way in which electrical conductors fit into the scheme of electrostatics.

As a beginning, the electric field $\boldsymbol{E}$, work, and electrostatic potential will be defined. The sum on the right hand side of equation (1) is a vector. If the electric field $\boldsymbol{E}$ at the point $\boldsymbol{r}(0)$ occupied by the charge $q(0)$ is defined so that the force is equal to the product of charge and field, $\boldsymbol{F}(0) = q(0)\boldsymbol{E}(0)$, then the electric field $\boldsymbol{E}(0)$ at $\boldsymbol{r}(0)$ generated by the other charges $q(n)$ at points $\boldsymbol{r}(n)$ is equal to the sum

$$\boldsymbol{E}(0) = \sum_n \frac{q(n)\boldsymbol{r}(0,n)}{4\pi\epsilon_0 r^3(0,n)}$$

If there is only one charge q and it is at the origin, $\boldsymbol{r} = 0$, the field at a general point specified by $\boldsymbol{r}$ will be

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{q\boldsymbol{r}}{4\pi\epsilon_0 r^3}. \qquad (2)$$

It can be shown that the work $W(1,2)$ done in this field *against* the force $q(0)\boldsymbol{E}(\boldsymbol{r})$ in taking a test charge $q(0)$ from $\boldsymbol{r}(1)$ to $\boldsymbol{r}(2)$ is

$$W(1,2) = q(0)\frac{q}{4\pi\epsilon_0}\left(\frac{1}{r(2)} - \frac{1}{r(1)}\right).$$

This result has two important consequences. First the work done depends only on the initial and final positions $\boldsymbol{r}(1)$ and $\boldsymbol{r}(2)$ and is independent of the path (*i.e.*, work is not a vector quantity; it is called a scalar quantity because it has only magnitude), and secondly the work $W(1,2)$ can be expressed as $q(0)$ times the change in the function $\phi(\boldsymbol{r})$, *i.e.*, $q(0)\phi(\boldsymbol{r})$, in which

$$\phi(\boldsymbol{r}) = \frac{q}{4\pi\epsilon_0 r}.$$

This equation defines the electrostatic potential.

Because the work is independent of the path, the work done against or by the electrostatic field in taking a charge round a closed path is zero. This is in conformity with the conservation of energy on the assumption that the motion of the charge does not, of itself, cause work to be done elsewhere in the universe, true only if the motion of the charge is slow (*i.e.*, quasistatic). It also means that the vector, curl $\boldsymbol{E}$, or $\nabla \times \boldsymbol{E}$, must be zero. Thus, the electrostatic field $\boldsymbol{E}$ may also be defined as the gradient (symbolized ∇ and called del) of a scalar potential; *i.e.*, expressed in vector notation as

$$\boldsymbol{E} = -\nabla\phi, \qquad (3)$$

another way of writing equation (2), a most important vector relationship. Because the effects of several charges are additive the potential of several charges is the summation

$$\phi(\boldsymbol{r}[0]) = \sum_n \frac{q(n)}{4\pi\epsilon_0 r(0,n)}. \qquad (4)$$

Definition of conservative field

Writing the properties of the vector electrostatic field in terms of a simple additive scalar potential is an enormous simplification. A field of any kind with this relationship to potential is known as a conservative field.

The electric field $\boldsymbol{E}$ has another important property. If the electric field $\boldsymbol{E}$ is designated E_n when it is perpendicular to a closed surface S, and E_n is known (*i.e.*, can be calculated) for each point of the surface, then the expression $\oint_S E_n dS$ is a surface integral, meaning that E_n is summed over the whole surface. Gauss's theorem states that such a surface integral is proportional to the sum of the electric charges that lie within the closed surface, which can be expressed in equation form as

$$\oint_S E_n dS = q/\epsilon_0, \qquad (5)$$

in which q is the total charge inside S and ϵ_0 is the dielectric constant *in vacuo*.

The results above can be generalized to the case in which the charge is continuously distributed in space with a density ρ (*e.g.*, in the space between the cathode and anode of a vacuum tube). In this case the total charge q is found by integrating over the volume (V) enclosed by surface S: $q = \int_V \rho dV$. Substituting this value of q in equation (5) gives the relation

$$\oint_S E_n dS = \frac{1}{\epsilon_0}\oint_V \rho dV.$$

If V and S are allowed to become infinitesimal in this equation, then upon integration, the result

$$\frac{1}{V}\oint_S E_n dS = \rho/\epsilon_0$$

follows. The expression on the left side of the equation has an equivalent form, namely, the divergence of E (div E), denoted $\nabla \cdot E$; so that

$$\nabla \cdot E = \rho/\epsilon_0. \tag{6}$$

In non-vector form, the divergence of E is a partial differential equation in terms of the components E_x, E_y, and E_z of the field E:

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = \rho/\epsilon_0. \tag{7}$$

This is generally regarded as the basic equation of electrostatics and is sometimes called Poisson's equation although this name is usually reserved for the equation that results from the combination of equations (3) and (6):

$$\nabla^2\phi = -\rho/\epsilon_0,$$

Stated explicitly in terms of the coordinates x, y, and z, it is

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2} = -\rho/\epsilon_0. \tag{8}$$

As may be seen, both of these equations are expressed in terms of the scalar potential ϕ.

If in the equation (4) the charges are replaced by a continuous density $\rho(r[1])$ in a volume element dV at $r(1)$, it yields the relation

$$\phi(r[0]) = \int_V \frac{\rho(r[1])\, dV(r[1])}{4\pi\epsilon_0 r(0,1)}. \tag{9}$$

If the potential ϕ is given as a function of position, equation (8) yields the charge density ρ directly. On the other hand, if the charge density is given, equation (9) yields the potential ϕ and thus the electrostatic field E.

A conductor contains mobile electric charges, usually electrons, and, in the presence of a field E, these charges move. The motion continues until the redistribution of charge in the conductor produces a field that just cancels E in the conductor. Thus in electrostatics a conductor is a body that cannot support (contain) an electric field. Because the total field in the conductor is zero the gradient of the potential is zero by equation (3), $\nabla\phi = 0$. For the gradient of a quantity to be zero the quantity itself must be zero or a constant. Thus the potential must be constant in a conductor. In particular, the surface of a conductor must be everywhere at equal potential, an equipotential surface. If there is no electric field E inside the conductor, its components E_x, E_y, and E_z do not change and therefore the left side of equation (7) is zero and (7) reduces to $\rho = 0$. Charge density zero means there is no volume charge within the conductor. Thus all the charge associated with a conducting body must reside on its surface. Surface charge per unit area is symbolized by the Greek letter σ (sigma). Furthermore, at the surface of a conductor E must be normal (*i.e.*, perpendicular) to the surface, and the intensity of the normal component of E is related to surface charge σ by $\sigma = \epsilon_0 E_n$. The original observations that led to the law of charge conservation were made by Benjamin Franklin who performed the following experiment. One person A standing on an insulated block rubs a glass tube with silk. A second person B, also on an insulated block, touches the glass tube and receives a shock. If now either A or B touches a third person C standing on the floor, C will receive a shock. If, however, A and B first touch each other they receive a shock and subsequently, when either A or B touches C, C will feel no effect. Franklin interpreted this as follows. A initially, by friction, removed negative electricity from the glass leaving it with a net positive charge. B, on touching the glass, transfers negative electricity to it and is left with a net positive charge. When A touches C he (C) receives a negative shock and when B touches C he (C) receives a positive shock; but if A touches B first their opposite charges are cancelled and therefore neither one can affect C. In the whole series of processes the algebraic sum of the charges present remains zero.

Franklin's charge conservation experiment

The first quantitative investigation of electric charge was carried out by Michael Faraday in his famous ice pail experiment. He began by charging a metal ball suspended on an insulating silk thread. He then connected a metal ice pail resting on an insulating block to a gold leaf electroscope (for the operating principle of this device see below, *Electrostatic induction*), and lowered the charged ball into the pail. The electroscope reading increased as the ball was lowered into the pail, finally reaching a steady reading once the ball was within the pail. If the ball was withdrawn without touching the pail the electroscope reading fell to zero but if the ball touched the bottom of the pail the reading remained at its steady value and the ball on removal was found to be completely discharged. Faraday concluded that the electric charge induced on the outside of the pail, when the ball was inside the pail (but not in contact with it) was exactly equal to the charge initially on the ball. He then inserted other bodies in the pail, such as a set of concentric pails insulated from one another, space between these pails being filled with various insulating media; *e.g.*, sulfur. In each case, the electroscope reading was the same once the ball was completely within the pail. From this he concluded that the total charge of the system was an invariable quantity equal to the initial charge of the ball. In other words, charge is a conserved quantity. As remarked above, this belief that conservation is a fundamental property of charge rests not only on Franklin's and Faraday's experiments but on its complete agreement with all the observations in electrical engineering, quantum electrodynamics, and experimental electricity. It might be expressed as follows: "In an isolated system the algebraic sum of all the charges present remains constant." If an isolated system with a closed surface is divided into an inner part A and an outer part B by a second closed surface S, any charge transferred from region A to region B must cross S and its rate of flow is the outward current I through S. If q is the charge in S and dq/dt its rate of increase, then $dq/dt + I = 0$. If ρ is the charge density in the volume V enclosed by S, and J is the vector current density, it can be expressed as the sum of two integrals,

$$\int_V (d\rho/dt)dV + \oint_S J_n dS = 0,$$

in which J_n is the outward normal component of J at the surface S. This holds for any volume, however small, and letting V become infinitesimal this gives

$$\frac{d\rho}{dt} + \frac{1}{V}\oint_S J_n dS = 0,$$

which can be expressed in vector notation using the divergence (symbolized $\nabla\cdot$) of the vector J as

$$\frac{d\rho}{dt} + \nabla\cdot J = 0.$$

This equation without doubt can be regarded as the simplest and most economical statement of the conservation law.

Recapitulation of the basic laws

The basic laws of electrostatics have now been stated, but it is convenient to collect these laws and to express them in a number of different forms.

The most basic statement is concerned with the definition of charge.

1. To any electrified body can be assigned a numerical quantity, capable of taking either sign, its charge q. The forces experienced and exerted by the body are proportional to q.

2. The sum of all the charges in an isolated system is constant.

3. If a body of charge q at r experiences a force F the electric field (E) at r is equal to F/q.

4. The field at a point located at r relative to a body of charge q is

$$E = \frac{qr}{4\pi\epsilon_0 r^3}.$$

5. The effects of several charged bodies are linearly

additive. From these five laws a number of subsidiary laws can be derived.

6. The line integral of the electrostatic field $\boldsymbol{E}$ around any closed path is zero. This is a mathematical way of saying that although an electric charge may do work in moving through an electric field, or work may be done on it, the net work will be zero if the charge is returned to its starting position. In vector notation

$$\nabla \times \boldsymbol{E} = 0.$$

7. The field $\boldsymbol{E}$ can be derived as the negative gradient of a scalar potential by

$$\boldsymbol{E} = -\nabla\phi.$$

8. The scalar potential at $\boldsymbol{R}$ resulting from charges $q(n)$ at $\boldsymbol{r}(n)$ can be expressed as

$$\phi(\boldsymbol{R}) = \sum_n \frac{q(n)}{4\pi\epsilon_0|\boldsymbol{R} - \boldsymbol{r}(n)|}$$

(here the magnitude of a vector sum is expressed by parallel brackets); or, if the charge distribution is continuous with density $\rho(\boldsymbol{r})$, as

$$\phi(\boldsymbol{R}) = \int_V \frac{\rho(\boldsymbol{r})dV(\boldsymbol{r})}{4\pi\epsilon_0|\boldsymbol{R} - \boldsymbol{r}|}.$$

9. If S is the closed surface of a volume V in which there is a net charge q,

$$\oint_S E_n dS = q/\epsilon_0.$$

If the charge distribution is continuous this leads to

$$\nabla \cdot E = \rho/\epsilon_0.$$

10. As a result the electrostatic potential ϕ satisfies Poisson's equation

$$\nabla^2\phi = -\rho/\epsilon_0,$$

and in empty space, Laplace's equation, which is

$$\nabla^2\phi = 0.$$

11. If $\boldsymbol{J}$ is the current density vector, the conservation of charge law is expressed by

$$\nabla \cdot \boldsymbol{J} + \frac{\partial\rho}{\partial t} = 0.$$

12. Lines of electric field are normal to equipotential surfaces.

To these laws some special laws relating to conductors can be added. Of these the most basic is:

13. The electrostatic field within a conductor is zero. From this it follows that

14. The surface of a conductor is an equipotential.

15. Charge resides only on the surface of conductors.

16. The electric field just outside a conductor is normal to its surface, and

17. If E_n is the normal electric field away from a conducting surface the surface charge density on the conductor is $\sigma = \epsilon_0 E_n$. Of all these, the basic laws are (1), (2), (4), (5), and (13). The result (3) is a definition and the remaining laws are restatements or consequences of the basic laws.

Given definitions of q, ρ, $\boldsymbol{E}$, and $\boldsymbol{J}$ the simplest statement of the laws, in vector notation, is the set of equations

$$\begin{aligned} \boldsymbol{F} &= q\boldsymbol{E} \\ \nabla \times \boldsymbol{E} &= 0 \\ \nabla \cdot \boldsymbol{E} &= \rho/\epsilon_0 \\ \nabla \cdot \boldsymbol{J} + \frac{\partial\rho}{\partial t} &= 0. \end{aligned}$$

Except for the third equation, which is valid only in vacuum, these laws are universally valid in any medium. Alternatively, the third and fourth laws are universally valid in vacuum, but as will be seen, the first and second have to be modified when considering magnetic forces and electromagnetic induction. With these reservations the above set of four equations summarizes the theoretical structure of electrostatics. The complete scheme of electrostatics also includes a treatment of the effects of insulating bodies, or dielectrics, which is detailed below.

Properties of the electrostatic field. The force acting on unit charge in a field $\boldsymbol{E}$ is equal to $q\boldsymbol{E}$ (from the equation $\boldsymbol{F} = q\boldsymbol{E}$) and lines of force are traced by starting at one point in space and proceeding along the direction of $\boldsymbol{E}$ following any changes in $\boldsymbol{E}$. In this process equipotential surfaces are crossed and the path is normal to these surfaces. An example is given in Figure 2

From (top) W.R. Smyth, *Static and Dynamic Electricity* copyright 1939; used with permission of McGraw-Hill Book Co.

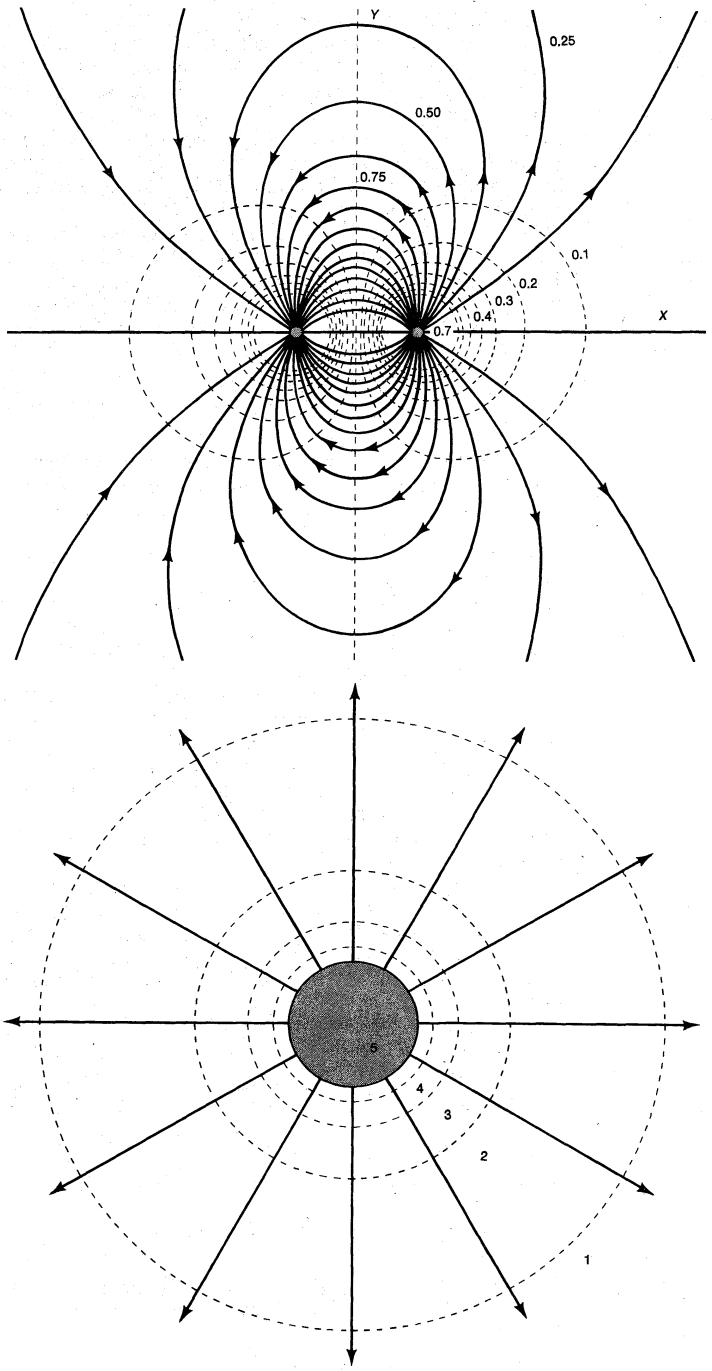


Figure 2: Lines of force (solid lines) and equipotentials (broken lines). (Top) Near two equal charges of opposite sign; (bottom) near a charged sphere (see text).

(top) which shows the lines of force and equipotentials (*i.e.*, equipotential lines) in the vicinity of two charges of opposite sign. Lines of force are seen crowded together near the charges where the field is most intense, and where the field is strongest the equipotentials are closest. Figure 2 (bottom) shows the lines of force and equi-

Lines of force and equipotentials

potential associated with a uniformly charged conducting sphere. These equipotentials get further apart with distance according to the law

$$\phi = \frac{q}{4\pi\epsilon_0 r}.$$

The density of lines of force decreases quadratically with distance (as $1/r^2$), exemplifying the inverse square law; but the total number crossing each spherical surface remains constant, as required by Gauss's law

$$\oint E_n dS = q/\epsilon_0.$$

A diagram such as this powerfully suggests that lines of force emanate from charge, and explains why charge is often referred to as the source of the electric field. In some cases general notions of symmetry and an intuitive idea of the behaviour of field lines based on this approach allow the solution of electrostatic problems without engaging in mathematical calculations. In other cases, Gauss's law often provides enough insight to lead to a numerical solution.

The presence of an electric field in space is associated with the presence of charge or charges, which suggests that work must have been expended to bring them together. It can be shown with the help of Gauss's law that the work (W) is expressed by the integral equation

$$W = \frac{1}{2}\epsilon_0 \int_{V\to\infty} E^2 dV, \qquad (10)$$

in which the range of integration is extended to include the whole volume V. This equation expresses the energy of the system entirely in terms of the electrostatic field.

The energy stored per unit volume in a compressed gas or a distorted solid is proportional to the square of the stress and equation (10) suggests that in some ways the "ether" is analogous to an elastic medium. This idea was pursued with great vigour by the theoretical electricians of the 19th century and, although the search for mechanical explanations of electrical phenomena is now regarded as pointless, the theoretical structure developed by Maxwell on this basis remains an important component of electromagnetic theory. The presence of an electric field in a region of space does, in fact, confer on it many of the properties of an elastic medium even if the concept of an elastic ether is no longer used.

Electrostatic induction. Both Cavendish's and Faraday's experiments showed that when a charged body is placed inside a hollow conductor the outside of the conductor acquires a charge equal to that of the body. This is an important example of the phenomenon of electrostatic induction, and it is instructive to see how it comes about in terms of the laws of electrostatics that have just been formulated.

Figure 3 shows a body with a charge q inside a spherical conducting vessel. Inside the conductor on the surface S, shown as a broken line, the electric field is zero. Thus $\oint_S E_n dS$ over the surface S is zero. Gauss's theorem therefore reveals that there is no net charge inside S. There must therefore be a surface charge density σ' induced on the inner surface S' of the vessel so that

$$q + \oint_{S'} \sigma' \, dS' = 0.$$

But the vessel was originally electrically neutral and so there must be an opposite surface charge σ'' on the outer surface S'' and hence

$$\oint_{S''} \sigma'' \, dS'' = -\oint_{S'} \sigma' dS' = q.$$

In the absence of any outside disturbing influence this charge will be evenly distributed over S''. If the radius of the outer surface is a, then integration of this equation gives $4\pi a^2\sigma'' = q$. Lines of electric field will leave the outer sphere radially and the field at the surface of the sphere is $\frac{\sigma''}{\epsilon_0}$. The application of Gauss's theorem over any spherical surface at a radius r, r being greater than a, $r > a$, leads to

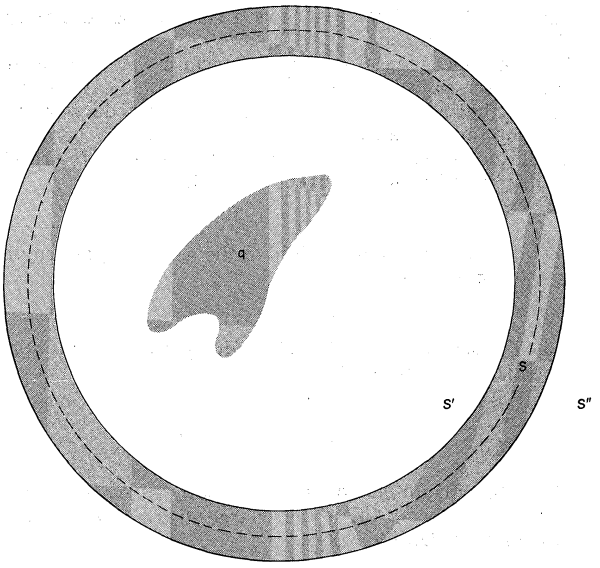


Figure 3: A charged body q inside a spherical conducting vessel S with walls S' and S'' (see text).

$$E = \frac{q}{4\pi\epsilon_0 r^2},$$

and the whole system behaves externally as though there were a point charge q at the centre. If the charged body is allowed to touch the interior of the sphere it neutralizes the inner surface charge but does not affect the outer charge. When the body is removed it is uncharged and the sphere has a net charge q on its outer surface.

The spherical capacitor

Next to be considered is a spherical capacitor, a device for storing electric charge, consisting of two concentric metal spheres. The inner conducting sphere initially has a charge q that must reside on its outer surface. If the outer conducting sphere is initially uncharged and isolated there will be an induced charge $-q$ spread over its inner surface and therefore $+q$ on its outer surface. Next the outside of this sphere is connected to ground or earth, so that the outer charge leaks away, and the contact is then broken. If later the inner sphere is removed (*e.g.*, by imagining the outer sphere cut in two) without touching the outer sphere, the outer sphere will be left with a net charge $-q$. This charge could be transferred to a third body and the process repeated. Thus starting with a single charge q after n steps a charge $-nq$ is transferred to the third body and $+nq$ to ground (the total charge $+q$ being conserved at all times). Many continuously acting electrostatic machines work on essentially this principle.

Again, when the outer sphere is still grounded, the inner sphere has a plus charge $(+q)$ on its outer surface and the outer sphere a negative charge $(-q)$ on its inner surface. If now the radii of both spheres are increased keeping their separation constant, then the surfaces flatten out and local conditions approach more and more closely to the geometry of a capacitor consisting of two plane plates parallel to each other. Therefore, if a plate carrying a charge q is placed above and near a grounded plate, the charge will concentrate on the inner surface and a charge $-q$ will be induced on the upper surface of the grounded (earthed) plate. If the earthed contact is broken and the lower plate moved away it will have a net charge $-q$.

The first and most obvious use to which these induction effects can be put is in electrostatic shielding. Many pieces of laboratory apparatus, and some pieces of domestic equipment (*e.g.*, piezoelectric phonograph pickups, transistor amplifiers, and electrostatic microphones) are adversely affected by stray electric fields. These can be shielded completely from stray fields simply by placing them inside a metallic enclosure usually referred to as a screen or screening can.

A second application of induction is the measurement of charge. In Faraday's time the most usual indicator of the electrical state of a body was the gold leaf electroscope, which consists of two thin leaves of gold hanging from an insulated metal rod, and mounted inside a metal box. When the rod is charged the leaves repel each other and the deflection indicates the size of the charge. Faraday's ice pail experiment

Faraday's electroscope incorporated an ice pail or metal vessel (also conducting) standing on a plate attached to the insulated metal rod. When a charged body was placed in the pail, even if it did not touch the pail, induction (see below) insured that the electroscope indicated the charge on the body. Further, once the charge was measured the body could be removed with charge unchanged. A modern experiment would retain the pail, or something like it called a Faraday cage, with a more sophisticated indicator than a gold leaf electroscope.

In the same paper in which Faraday described the phenomenon of electrostatic induction he speculated on its possible connection with atmospheric electricity and lightning. The modern view would probably be that although induction may play a limited role in the production of charged volumes of air or clouds the main mechanisms involve the mechanical separation of positive and electrical charge much as in a Van de Graaff machine used in generating a high potential (named after the inventor, Robert J. Van de Graaff, a physicist in the United States). For a more detailed account see the articles on LIGHTNING and THUNDERSTORMS.

The electric dipole. To obtain a general understanding of electrostatics it is not necessary to study in detail the fields caused by particular arrangements of charges, with one exception, that of the so-called dipole. An electric dipole is a system of two equal but opposite charges $+q$ and $-q$ at a fixed separation l from each other, which can be specified as a vector moment as shown in Figure 4A. The vector moment (p) of the dipole is defined as a vector quantity directed from the negative charge to the positive charge and equal in magnitude to the product of charge and separation distance:

$$p = ql.$$

Importance of the dipole

The dipole is important because it is the simplest system of overall electrical neutrality (*i.e.*, has no net charge) that produces external effects. Aspects of interest in connection with the dipole are: the form of the potential and field distribution it generates, the effect of an impressed field on a dipole, and the generalization of the concept in connection with atomic physics.

The energy of a dipole in a field E may be derived from a potential ϕ in which the electric field is equal to minus its gradient, $E = -\nabla\phi$. As may be shown, the energy of a charge situated in an electric field is equal to the product of charge times electric potential. The energy of the dipole thus is equal to the sum of its energies, that is, $+q$ of the positive charge times its potential $\phi(r + l)$ at position $r + l$ and $-q$ of the negative charge times its potential $\phi(r)$ at r:

$$U = q\phi(r + l) - q\phi(r),$$

r being the position of the negative charge. If it is assumed that l is small or that ϕ varies only slowly over the length of the dipole the energy (U) can be approximated by minus the scalar product of dipole moment (p) and electric field (E), or $U = -p \cdot E$. It can be shown that if E is a constant field and θ the angle between the dipole moment (p) and (E), the dipole energy (U) becomes $U = -pE \cos\theta$.

In a constant field the dipole energy (U) is independent of the position of the dipole but depends on its orientation. There is therefore no net force on the dipole, but a twisting force that tends to turn p towards the direction of E (*i.e.*, align the two charges along the field direction). The magnitude (Γ) of this couple is

$$\Gamma = -\frac{\partial U}{\partial\theta} = pE \sin\theta;$$

or, in vector notation, the vector product of the dipole moment (p) and the electric field (E),

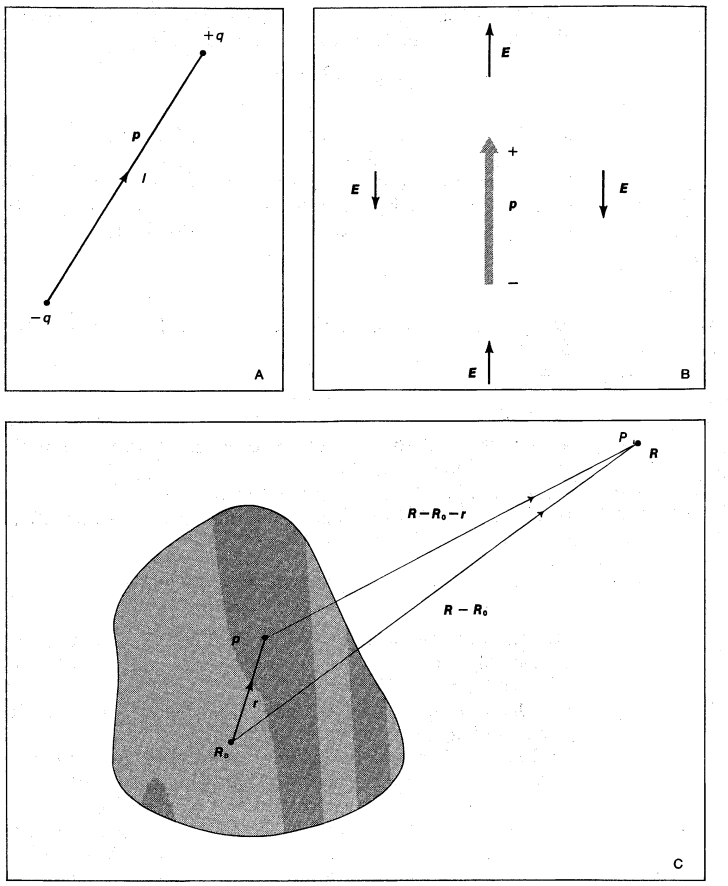


Figure 4: *Electric dipole.*
(A) Dipole of two equal and opposite charges $-q$ and $+q$, separated by distance l. The dipole moment $p = ql$ is directed from $-q$ to $+q$. (B) Fields E near a dipole p. (C) Vector notation for the position (R) of a point (P) in the vicinity of a dipole (p).

$$\Gamma = p \times E.$$

If E is non-uniform, the force acting on the dipole is $-\nabla U$, and so

$$F = -\nabla U = \nabla(p \cdot E).$$

If p is a constant vector both in magnitude and direction this result can be simplified to

$$F = (p \cdot \nabla)E; \qquad (11)$$

i.e., the x component of the force is

$$F_x = p_x\frac{\partial E_x}{\partial x} + p_y\frac{\partial E_x}{\partial y} + p_z\frac{\partial E_x}{\partial z}.$$

It can be shown that (11) is also the correct expression even when dipole moment (p) is not constant but is related to E. This is important because all atoms and molecules acquire a dipole moment in an electric field and in most cases $p = \alpha\epsilon_0 E$, in which α is a positive constant (of the order of the volume of the atom). In this case,

$$F = \alpha\epsilon_0(E \cdot \nabla)E,$$

and this may be shown to be

$$F = \tfrac{1}{2}\alpha\epsilon_0\nabla(E^2).$$

This means that neutral atoms and molecules are attracted towards regions in which the electric field is the strongest. The potential $\phi(R)$ at point R resulting from a dipole p at the origin may be expressed by the equation

$$\phi(R) = \frac{p \cdot R}{4\pi\epsilon_0 R^3}.$$

As may be seen from this equation, the potential decreases quadratically with distance, whereas, as was seen above, the potential for a point charge decreases with the first power. The corresponding electric field pro-

duced by the dipole is $E = -\nabla\phi$ and the final result obtained after much vector manipulation is

$$E(R) = \frac{3(p \cdot R)R - R^2 p}{4\pi\epsilon_0 R^5},$$

and decays as the cube of the distance. Some insight into the form of the field lines can be gained by considering the value of E along the axis of the dipole in which $p \cdot R = pR$, and in the equatorial plane in which $p \cdot R$ is equal to zero. Along the axis, the dipole field is

$$E = \frac{2p}{4\pi\epsilon_0 R^3}$$

and is parallel to p. In the equatorial plane, the dipole field is

$$E = -\frac{p}{4\pi\epsilon_0 R^3}$$

and is opposed to p. These results are shown schematically in Figure 4B.

Energy of interaction between two dipoles

The energy of interaction between a dipole p_1 at the origin and a dipole p_2 at R is obtained from the above expression for $E(R)$ and $U = -p \cdot E$ as

$$U = -\frac{3(p_1 \cdot R)(p_2 \cdot R) - R^2 p_1 \cdot p_2}{4\pi\epsilon_0 R^5}$$

and falls off as the cube of the distance R. Because the energy of interaction of two charges q_1 and q_2 falls off only as $1/R$, dipole interactions between systems that also have a net charge are negligible at any appreciable distance. On the other hand, because most of the systems studied in physics (*e.g.*, atoms, molecules, piezoelectric crystals, or complete systems of batteries and conductors) are usually neutral when considered as a whole, the distant interaction between these systems is often solely caused by their dipole moments. In magnetism, in which no free "poles" exist, the dipole effects are always dominant.

If there is a distribution of charge, of density ρ, confined to a small volume V of linear dimensions l, the potential at a great distance ($R \gg l$), which is

$$\phi(R) = \int_V \frac{\rho(r)dV(r)}{4\pi\epsilon_0|R - r|}$$

can be approximated as follows. An origin at R_0 is taken somewhere in the charge distribution and r is defined relative to R_0 as shown in Figure 4C; the denominator of the quantity being integrated is then expressed as

$$\frac{1}{|R - R_0 - r|} = \frac{1}{|R - R_0|} - r \cdot \nabla\left(\frac{1}{|R - R_0|}\right) + \text{etc.};$$

so that

$$\phi(R) = \frac{1}{4\pi\epsilon_0}\frac{1}{|R - R_0|}\int_V \rho dV - \frac{1}{4\pi\epsilon_0}\nabla\left(\frac{1}{|R - R_0|}\right) \cdot \int_V \rho r dV.$$

But $\int_V \rho dV = q$ is the total charge and, if the dipole moment is defined as

$$p = \int_V \rho r dV, \tag{12}$$

it produces a potential $\phi(R)$ according to

$$\phi(R) = \frac{q}{4\pi\epsilon_0|R - R_0|} - p \cdot \nabla\left(\frac{1}{4\pi\epsilon_0|R - R_0|}\right).$$

Thus the effects of the charge distribution have been approximated in terms of a point charge q at R_0 and a dipole p at R_0. If it happens that q equals zero, this approximation is unique and the choice of R_0 is arbitrary as long as it is within or near the distribution. If, however, q is not equal to zero, the choice of R_0 is critical and determines the correct value of p. Changing R_0 to R'_0 changes p by $-q(R'_0 - R_0)$. When equation (12) is applied to atoms, R_0 is always taken to be at the nucleus and no ambiguity arises. When nuclei are considered, R_0 is usually taken as the centre of mass. In a general consideration of media that are continuous on a macroscopic scale, no unique origin is specified and the location of R_0 becomes quite complicated, leading to false conclusions if R_0 is defined incorrectly (see the section below dealing with dielectrics).

The definition of the dipole moment by equation (12) is a simple generalization of the original definition (see Figure 4C) and is an important concept in atomic physics. Many important phenomena (*e.g.*, dielectric behaviour, radiation and absorption of spectral lines, and intermolecular forces) are directly related to atomic dipole moments.

The electron. The science of electrostatics, so far presented, is the study of fields, potentials, and charges that have been indifferently treated as charges on material bodies, point charges, or continuous volume distributions of charge. The discovery of the essentially electrical constitution of matter in the early years of the 20th century clearly forces a reconsideration of some of these notions. This reassessment has two effects, one clearly beneficial: a new insight is gained into the nature of electrical phenomena, and the other less welcome: everything tends to be interpreted in terms of electrons even before the basic laws of electricity and the nature of the gross phenomena that they describe have been grasped. The electron is treated separately in another article (see ELECTRON) and only its most basic properties and their consequences for the study of electricity need be of concern here.

Millikan's oil drop experiment

The first reasonably precise determination of the charge of the electron ($-e$) was made by Robert A. Millikan, a physicist in the United States, in 1912. Although his method is not now used for precise measurements his result was a landmark in physics. Millikan observed the rate of fall of oil drops between two plates across which he could maintain a known potential difference. In the absence of any field the rate of fall of a drop is determined by gravity, the size of the drop, and the viscosity of air. Because the gravity and the viscosity are known, the rate of fall gives the size and mass of the drop. If a drop acquires a charge q, then the force acting on it is no longer the mass (m) times the acceleration of gravity (g), but $mg + qE$, and its velocity is changed in the ratio $(mg + qE)/qE$. In the circumstances of Millikan's experiment, in which the air was slightly ionized, he found that charge q was always a multiple, generally -1 times a basic unit $e = 1.60 \times 10^{-19}$ coulomb, which he identified as the electronic charge.

The important determinations were (*i*) the order of magnitude of the charge and (*ii*) the fact that the product of Avogadro's number (N, the number of atoms in a gram atom of a substance, 6.02×10^{23}) and the electronic charge, Ne, is 9.6×10^4 coulombs, which is just one Faraday; *i.e.*, the charge required to liberate one gram (one gram atom) of hydrogen. Further, because the value of e/m for electrons measured by Thomson, and with increased precision by later workers, is 1.76×10^8 coulomb per gram, the mass of the electron is 9.11×10^{-28} gram or 9.11×10^{-31} kilogram. (A good round number for the mass is 10^{-30} kilogram.) These numbers give some idea about the numbers of electrons that actually take part in macroscopic electrostatic phenomena. A few examples are:

The normal density of solid matter is around five grams per cubic centimetre or 5,000 kilograms per cubic metre and a gram molecule of a simple compound occupies something between five and 100 cubic centimetres. Thus the number of atoms per cubic centimetre is in the region of 10^{22} to 10^{23}. The number of electrons per atom varies from one to more than 92 in the periodic table, but the number of relatively loosely bound valence electrons per atom that can be appreciably influenced by electric fields is usually small, in many cases there is only one per atom. As a general rule these are of the order of 10^{23} effective

electrons per cubic centimetre or 10^{29} per cubic metre. The total charge associated with these electrons is thus some 10^{10} coulombs per cubic metre. The number of electrons lying within one atomic diameter of the surface of a solid is $(10^{29})^{2/3}$ per square metre (*i.e.*, about 10^{20}) and so the available charge per square metre is of the order of 10 coulombs per square metre. Now a surface charge density of 10 coulombs per square metre is associated with an electric field normal to the surface of $10/\epsilon_0 \simeq 10^{12}$ volts per metre. Even in extreme cases fields rarely exceed 10^8 volts per metre, and so, in general, the surface charges that appear in electrostatics correspond to fractional changes in the internal charge distribution. Furthermore, a field of 10^{12} volts per metre corresponds to a field of 100 volts per angstrom (the angstrom is 10^{-10} metre and is the typical radial dimension of atoms). A field of this strength would be sufficient to disrupt and destroy the surface atoms, and certainly at this magnitude the simple notions about conductors and insulators would be useless.

A conducting sphere of radius a metres with a charge q has a potential (ϕ): $\phi = q/4\pi\epsilon_0 a \simeq 10^{10} q/a$ so that $q \simeq 10^{-10}\, a\phi$. The displaceable electronic charge (Q) is of the order of

$$Q \simeq 10^{10} \frac{4}{3}\pi a^3 \simeq 4 \times 10^{10} a^3;$$

and so,

$$q/Q \simeq 2.5 \times 10^{-21} \phi/a^2.$$

If, for example, $\phi = 10^4$ volts and $a = 10^{-6}$ metre this ratio is only 2.5×10^{-5}; in most practical cases it is much smaller. It may be concluded that when a body of normal dimensions is charged there is an utterly negligible change in the number of electrons it contains. Only in the most extreme cases (the channel in a field effect transistor is an important example) is the change in electronic content enough to change the properties of the material appreciably.

If it were possible to give a conducting sphere of one metre radius a charge of 10^{10} coulombs, certainly the addition of the required electrons would change its mass by less than one part in 2,000. Because in practice it would be difficult to give the sphere a charge of much more than $\frac{1}{100}$ coulomb, the fractional change in mass would be more like one part in 10^{15} and quite unobservable. From the point of view of practical electricity, the electric charge behaves like a massless fluid. It is only in electronic devices, and especially vacuum electronic devices in which electrons are detached from matter, that the finite mass of the charge carriers is significant.

Flow rate of electrons in one ampere

A current of one ampere corresponds to a flow of about 10^{19} electrons per second. In a metal with some 10^{29} mobile electrons per cubic metre a current density of one ampere per square metre corresponds to a mean electronic velocity of 10^{-10} metre per second; *i.e.*, electrons cover about one atomic diameter per second. Even at 10^{10} amperes per square metre, which is about the maximum safe current density in a conductor, the velocity is only one metre per second. Electrons in metals have random velocities as a result of thermal energy of about 10^6 metres per second, and thus the effects of an electric field driving a current produce only a minuscule change in the motion of the electrons. Although the discovery of the electron gives a new understanding of electricity, for most practical purposes it makes no difference to the description of macroscopic phenomena. Electricity in this context can be regarded as a massless and continuous fluid. It is only in microscopic physics (*i.e.*, atomic physics) and in electronics in which the motion of rather small numbers of electrons is controlled, that the mass and discrete charge of the electron become significant.

From the point of view of macroscopic practical electricity, one of the most important consequences of the discovery of the electron and the electronic structure of matter is that it relates the electrical properties of a medium to its structure. Thus, for example, it can be understood why carbon in the form of diamond is an insulator and in the form of graphite is a conductor, or why the relatively few insulating elements tend to be clustered towards the top right-hand corner of the chemists' periodic table. Another important practical consequence of this knowledge is, of course, the whole technology of electronics.

DIELECTRICS

Capacitors. If two conductors are placed in proximity with each other and one conductor is charged, the other acquires a surface charge by induction. If, however, the conductors are connected to the terminals of a battery to maintain a potential difference between the two separated conductors, the overall effect will be that one conductor will acquire a net positive charge and the other conductor an equal negative charge. The charge (q) will be proportional to the potential difference (ϕ), the coefficient of proportionality—*i.e.*, C—in the equation relating them, $q = C\,\phi$, is called the capacitance of the system. The system itself is referred to as a capacitor or condenser. The unit of capacitance in the International System of Units (SI), or metre–kilogram–second (mks) system, is the farad (F). It has the dimensions of coulomb per volt. The electrostatic unit of capacitance has no particular name and is usually written esu. One farad is equal to nearly 9×10^{11} esu. In practical electronics the μF, or microfarad, which is 10^{-6} F, and the picofarad, or pF, which is 10^{-12} F, are normally used. A useful order of magnitude conversion is 1 esu is approximately equal to 1 pF.

The potential (ϕ) associated with a point charge q at a distance r is $q/4\pi\epsilon_0 r$; and so

$$\epsilon_0 = \frac{1}{4\pi}\frac{q}{\phi}\frac{1}{r}.$$

The dimensions (physical units) of the fundamental electrical constant ϵ_0, the dielectric constant for empty space, are therefore most easily expressed in practical terms as farads per metre, or Fm^{-1}. The corresponding magnetic fundamental constant μ_0 in the expression for the magnetic field dB resulting from a current element Idl is

$$dB = \mu_0 Idl \times r/4\pi r^3,$$

and is arbitrarily given the exact value $4\pi \times 10^{-7}$ henry per metre, in which one henry (H) is the unit of inductance, in order to define the unit of current, ampere (A). This is connected with the practical use of the magnetic current balance for maintaining an exact standard of current related to mechanical forces. Maxwell's field equations lead to an expression for the velocity of light (c) in vacuum as $(\mu_0\epsilon_0)^{-1/2}$. Inasmuch as c can be measured with great accuracy and μ_0 is defined, ϵ_0 is defined by the relation

Velocity of light defined by Maxwell's equations

$$\epsilon_0 = \frac{1}{\mu_0 c^2} = \frac{10^7}{4\pi c^2},$$

in which c is the agreed experimental value of the velocity of light. This expression leads to the result

$$\epsilon_0 = 8.854\ldots \times 10^{-12}\mathrm{Fm}^{-1}.$$

For order of magnitude calculations, ϵ_0 may be taken to be approximately equal to $10^{-11}\mathrm{Fm}^{-1}$.

The charge q on a conductor can be expressed as the surface integral of its surface charge density, and, since by Gauss's theorem the surface charge density (σ) is equal to the product of the dielectric constant and the outward normal component of the electric field (E_n) at the surface, or $\sigma = \epsilon_0 E_n$, the capacitance (C) can be expressed as

$$C = \epsilon_0 \frac{\int E_n dS}{\phi}.$$

This relationship gives a formal method of calculating capacitance from the geometry of the system. It is, however, often simpler to proceed by a more direct method, as shown in the following three examples.

The first example to be considered is a spherical capacitor consisting of two concentric spheres of radii a and b. For simplicity the outer sphere of radius b is taken to be at zero potential and the inner at $\phi(a)$ with a charge q. By symmetry the electric field is radial (outward from

centre), and Gauss's theorem yields, for E at radius r,

$$4\pi r^2\epsilon_0 E(r) = q.$$

Thus,

$$E(r) = q/4\pi\epsilon_0 r^2.$$

The potential at r is obtained by integrating this expression with respect to r:

$$\phi(r) = \phi(a) - \int_a^r E(r)dr,$$

and this gives

$$\phi(r) = \phi(a) - \frac{q}{4\pi\epsilon_0}\left(\frac{1}{a} - \frac{1}{r}\right);$$

but where r equals b, the potential $\phi(b)$ must be zero and so from the preceding equations,

$$\phi(a) = \frac{q}{4\pi\epsilon_0}\left(\frac{1}{a} - \frac{1}{b}\right).$$

Thus, the capacitance of the system is

$$C = \frac{q}{\phi(a)} = 4\pi\epsilon_0\frac{ab}{b-a}.$$

If b is much greater than a, this is approximately

$$C \simeq 4\pi\epsilon_0 a.$$

The result, therefore, is the capacitance of an isolated sphere. If the radius a is one centimetre, the capacitance is roughly 10^{-12} F, or one picofarad; in fact, it is also exactly one esu.

The second example is a long coaxial cylindrical capacitor consisting of two cylinders of radii a and b ($b > a$). In this case τ is taken to be the charge per unit length on the inner cylinder. Again the field is radial, but now Gauss's theorem yields

$$2\pi r\epsilon_0 E(r) = \tau,$$

and

$$E(r) = \frac{\tau}{2\pi\epsilon_0 r},$$

which gives

$$\phi(r) = \phi(a) - \frac{\tau}{2\pi\epsilon_0}\log\frac{r}{a}.$$

Because $\phi = 0$ when $r = b$, this yields

$$\phi(a) = \frac{\tau}{2\pi\epsilon_0}\log\frac{b}{a},$$

and the capacitance per unit length is

$$C' = \frac{2\pi\epsilon_0}{\log\frac{b}{a}} \text{ farad m}^{-1}.$$

In this case if b becomes infinitely large, C' approaches zero, but this has no physical meaning unless the length of the system is infinite; otherwise the calculation is based on a fallacy, that the field is radial over most of the length.

The final example is that of the plane parallel plate capacitor composed of two large parallel conducting plates separated by a distance d. Near the middle of the plates the field E between them is uniform and normal to the plane of the plates. The surface charge (σ) is $\epsilon_0 E$ and the potential difference is Ed; thus the capacitance per unit area (C'') in this region is $C'' = \sigma/\phi = \epsilon_0/d$. Near the edges of the plates there is a fringing effect of the fields, and the calculation is more complicated; but if the transverse linear dimensions of the system are large in comparison with d, it is easy to see that the total capacity (C) for plates of area A is approximately $C \simeq \epsilon_0 A/d$. The fringing fields or edge effects make C slightly larger. If the plate separation is one millimetre and the area is 100 square metres, the formula gives for the capacitance approximately 10^{-6} farad, or one microfarad. Because radio capacitors of up to $10^4\mu$F are readily available it can be concluded that they are plates of large area (folded) and small separation. This can be confirmed rather easily by taking a capacitor apart.

Fringing fields

If a capacitor of capacitance C is already charged to a potential difference ϕ, and an element of charge dq is added to the system, external sources (*e.g.*, batteries, electrostatic machines, etc.) will have to do an amount of work ϕdq to place the charge on the plates. The stored energy (W) can thus be calculated as

$$W = \int_0^q \phi dq = \frac{1}{C}\int_0^q q dq = \frac{1}{2C}q^2.$$

This work can also be written as

$$W = \tfrac{1}{2}C\phi^2 = \tfrac{1}{2}q\phi.$$

Neglecting edge effects, in a parallel plate capacitor this gives $q = A\epsilon_0 E$ and $\phi = Ed$; so that $W = \frac{1}{2}Ad\epsilon_0 E^2$.

In a parallel plate condenser the energy can be written either as

$$W = \frac{q^2}{2C} = \frac{q^2 d}{2\epsilon_0 A}, \tag{13}$$

or as

$$W = \tfrac{1}{2}C\phi^2 = \frac{\epsilon_0\phi^2 A}{2d}. \tag{14}$$

The force (F) pulling the plates together can be obtained from either result using the definition of work, change of potential energy with change of distance:

$$F = \frac{\partial W}{\partial d}.$$

Taking the partial differential, equation (13) gives

$$F = \frac{q^2}{2\epsilon_0 A},$$

whereas (14) gives

$$F = -\frac{\epsilon_0\phi^2 A}{2d^2} = -\frac{q^2}{2\epsilon_0 A},$$

which has the opposite sign. Only one result can be correct. Now the relation $F = \partial W/\partial d$ is only correct if, as d changes, no other external source does work on the system. This is the case if the plates are isolated and charge q is a constant, but is not the case if the plates remain connected to a battery and ϕ is held constant as the plates are moved. In this case, q decreases as the plates move apart and charge is forced back into the battery, increasing its internal energy. The correct expression for the force (F) is therefore

$$F = \frac{\partial W}{\partial d} - \phi\frac{\partial q}{\partial d},$$

because $-\phi\delta q$ is the work done on the battery for a change of d. Also, because

$$q = \frac{\epsilon_0 A\phi}{d}$$

and

$$-\phi\frac{\partial q}{\partial d} = +\frac{\epsilon_0 A\phi^2}{d^2} = \frac{q^2}{\epsilon_0 A},$$

this gives

$$F = -\frac{1}{2}\frac{q^2}{\epsilon_0 A} + \frac{q^2}{\epsilon_0 A} = +\frac{1}{2}\frac{q^2}{\epsilon_0 A}$$

as before.

This is not the only puzzling feature of this result. The charge on the surface of the positive plate is q, and a field (E) equal to ϕ/d acts on this plate. At first inspection it would seem that the force ought to be $Eq = \frac{q\phi}{d} = \frac{q^2}{\epsilon_0 A}$ and not $\frac{1}{2}\frac{q}{\epsilon_0 A}$. To see why this is not so, the conditions at one plate surface must be considered in more detail. At this surface the field E changes from zero in

the bulk of the metal to E_0 in vacuum on passing through the (thin) layer of surface charge. This layer has an actual charge density (ρ) per unit volume so that the charge density (σ) per unit area is

$$\sigma = \int_{layer} \rho dx.$$

The force (f) per unit area is

$$f = \int_{layer} \rho E dx;$$

but, from either Gauss's theorem or $\nabla \cdot \boldsymbol{E} = \rho/\epsilon_0$, the equation $\rho = \epsilon_0 \frac{\partial E}{\partial x}$ is obtained, and so

$$f = \epsilon_0 \int_{layer} E \frac{\partial E}{\partial x} dx = \epsilon_0 \int_0^{E_0} E \, dE = \frac{1}{2}\epsilon_0 E^2.$$

Thus, the force (F) for the total area (A) is

$$F = Af = \frac{1}{2}A\epsilon_0 E^2 = \frac{1}{2}\frac{q^2}{\epsilon_0 A}.$$

The factor ½ arises because, however thin the surface layer of charge is, the field drops to zero on passing into the metal through the layer and the charge in the layer experiences only the average field $\frac{1}{2}(E_0 + 0) = \frac{1}{2}E_0$.

Capacitors connected together

Capacitors are important circuit elements in electronics, and they are dealt with below in the section on alternating current circuits. This section treats only one related result, series and parallel connections. Two capacitors C_1 and C_2 are connected in parallel when two plates of each are connected together as in Figure 5 (left). For a

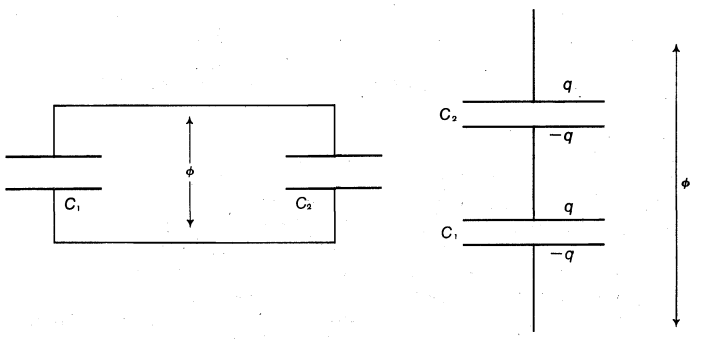


Figure 5: Two ways of connecting capacitors: (left) two capacitors C_1 and C_2 in parallel; (right) two capacitors C_1 and C_2 in series (see text).

given potential difference the total charge is $C_1\phi + C_2\phi$, and so the effective capacitance is $C_1 + C_2$. Two capacitors are connected in series when alternate plates are connected together (Figure 5 right). The charges are both the same and the total potential difference (ϕ) is the sum of the potential of each

$$\phi = \frac{q}{C_1} + \frac{q}{C_2},$$

and thus

$$\frac{1}{C} = \frac{\phi}{q} = \frac{1}{C_1} + \frac{1}{C_2}$$

or

$$C = \frac{C_1 C_2}{C_1 + C_2}.$$

The capacitance of the series combination is less than the capacitance of either capacitor alone.

Although only the capacitance of systems of conductors of especially simple geometry has been considered, any two conductors of any shape will have some capacitance, and, in addition, any isolated conductor will also behave as a capacitor; its capacitance will be of the order $4\pi\epsilon_0 l$, l being its linear dimension. The human body, for example, has dimensions of the order of one metre and therefore a capacitance of the order of 10^{-10} farad, or 100 picofarads. It is notable that the capacitance between two conductors of 100 metres of standard power cable is of the order of 5,000 picofarads.

Dielectric concepts. *Dielectric media.* Only one type of material body (*i.e.*, the electrical conductor) has been considered so far in relation to the electrostatic field. All other materials have been classified in a contrary way as insulators—*i.e.*, nonconductors. Here other kinds of materials will be investigated.

If any material body is inserted between the plates of a capacitor the potential difference (ϕ) between the plates will drop, even if the material is an insulator. Equally, if the potential difference is held constant the charge (q) increases. In both cases, then, the capacitance (C) is increased, and it becomes immediately apparent that insulators have electrical properties that are not completely specified simply by saying that they do not permit the passage of an electric current.

Experimentally, it is found that when a capacitor *in vacuo* is immersed in an insulating medium, the capacitance is increased by a numerical factor ϵ that is characteristic of the medium. The constant ϵ is referred to as the dielectric constant of the medium or, more precisely, the relative dielectric constant (relative to vacuum for which $\epsilon = 1$, symbolized ϵ_0). It is found that ϵ for a particular medium (*e.g.*, oil or mica), except in extremely high fields, is almost independent of the electric field and varies slightly with temperature and the state of strain of the medium. The value of ϵ, however, can vary considerably with the frequency of the electric field. For example, the dielectric constant of glass is about seven in static fields but falls to 2.3 in fields with frequencies in the range of 10^{14} hertz (abbreviated Hz; cycles per second); *i.e.*, the fields associated with optical phenomena. If the material is a single crystal, it is also found that the factor ϵ depends quite appreciably on the relative directions of the electric field $\boldsymbol{E}$ and the crystal axis. Materials that have this property are said to be electrically anisotropic. If the anisotropy also occurs at optical frequencies it leads to birefringence, or double refraction, the phenomenon in which the velocity of light in some media depends on the vibration plane of the light waves. The effects of anisotropy are usually small, but in some cases the value of ϵ varies with direction by a ratio of more than 2:1.

Two fundamental laws of electrostatics

The expression for the capacitance C of a plane parallel capacitor ($C = \epsilon_0 A/d$) depends on two fundamental laws of electrostatics. One of these is essentially that potential ϕ and electric field $\boldsymbol{E}$ are related by $\boldsymbol{E} = -\nabla\phi$; the other is that the surface charge σ on the plates is equal to $\epsilon_0 E_n$, and this is an expression of Gauss's theorem or equivalently of $\nabla \cdot \boldsymbol{E} = \rho/\epsilon_0$. The capacitance C is larger than predicted by this formula when the space between the plates is filled with a dielectric medium, and so one at least of these laws must be modified. Because both these laws originate in the expression for the force between charges, namely,

$$\boldsymbol{F} = \frac{q_1 q_2 \boldsymbol{r}}{4\pi\epsilon_0 r^3},$$

this law also needs modification. Indeed, starting with this law it might be claimed that it is an experimental fact that when two charges are immersed in a medium the force $\boldsymbol{F}$ is reduced by a factor $1/\epsilon$. Such an experimental fact is hard to verify if the medium is a liquid and impossible if it is a solid. It is simpler to proceed in a rather less direct way and maintain the validity of $\boldsymbol{E} = -\nabla\phi$ and the defining equation $\boldsymbol{F} = q\boldsymbol{E}$, but alter the equation $\nabla \cdot \boldsymbol{E} = \rho/\epsilon_0$. This expression is replaced by a completely new equation involving a new vector $\boldsymbol{D}$, called Maxwell's displacement vector and written $\nabla \cdot \boldsymbol{D} = \rho$. The displacement $\boldsymbol{D}$ is then related to $\boldsymbol{E}$ by $\boldsymbol{D} = \epsilon\epsilon_0\boldsymbol{E}$, in which ϵ is unity in a vacuum but characteristic of the medium in a dielectric. The electric field $\boldsymbol{E}$ now satisfies the equation $\nabla \cdot (\epsilon\epsilon_0\boldsymbol{E}) = \rho$, but not, be it noted, $\epsilon\epsilon_0(\nabla \cdot \boldsymbol{E}) = \rho$, which has quite different conse-

quences in a region in which ϵ varies with position (*e.g.*, at the surface of a dielectric).

Gauss's theorem now becomes $\oint_S D_n dS = \int_V \rho dV$, and, in a homogeneous uniform dielectric,

$$\boldsymbol{D}(r) = \frac{q\boldsymbol{r}}{4\pi r^3},$$

$$\boldsymbol{E}(r) = \frac{q\boldsymbol{r}}{4\pi\epsilon\epsilon_0 r^3}$$

$$\phi(r) = \frac{q}{4\pi\epsilon\epsilon_0 r}.$$

The dielectric constant of a medium can be measured simply by constructing a capacitor using the medium as the insulator and comparing its capacitance with a similar vacuum-spaced capacitor. Maxwell's equations and the theory of electromagnetic waves, however, lead to the result that the inverse ratio of the velocity c' in a dielectric medium to the velocity c of the waves in vacuum is equal to the quantity n and the square root of ϵ; *i.e.*, $n = c/c' = \sqrt{\epsilon}$.

The ratio n is the refractive index of the medium, and so ϵ can be obtained by measuring n. Because refractive index is a function of wavelength, the measurement gives ϵ for a particular frequency.

Numerical values for the dielectric constant

The quantity $\epsilon - 1$ depends on the density of the medium. In gases, ϵ is close to unity—for air at standard temperature and pressure, for example, $\epsilon = 1.0006$—and for many purposes almost any gas can be regarded as very nearly equivalent to vacuum. In solids ($\epsilon - 1$) usually exceeds unity. In most cases ϵ lies in the range 2–10, but several solids have static dielectric constants over 1,000. Dielectric constants in the optical frequency range are usually between two and three. A few media—*e.g.*, rutile (TiO_2)—have values as large as eight and correspondingly high refractive indices.

The following list briefly summarizes the effects of a medium with a dielectric constant ϵ greater than unity:

1. The electric field $\boldsymbol{E}$ from a charge within the medium, and also the potential ϕ, are reduced by $1/\epsilon$.
2. The capacitance of a system of conductors immersed in the medium is increased by ϵ.
3. The velocity of propagation of electromagnetic waves is reduced by $\sqrt{1/\epsilon}$.
4. The field $\boldsymbol{E}$ no longer satisfies the relation $\nabla \cdot \boldsymbol{E} = \rho/\epsilon_0$, but instead $\nabla \cdot (\epsilon\epsilon_0\boldsymbol{E}) = \rho$, which may also be written as $\nabla \cdot \boldsymbol{D} = \rho$.

Because a dielectric medium increases the capacitance, most practical capacitors employ dielectrics. This usage has a further advantage in that the breakdown strength of solid and liquid dielectrics is much higher than that of air. The plate separation for a given voltage can therefore be reduced, and this in turn leads to more capacitance for a given overall size.

Dielectric polarization. It is necessary now to examine how these effects arise, and, in particular, why in a dielectric the field $\boldsymbol{E}$ no longer satisfies the same equations. This amounts, in effect, to determining why $\epsilon - 1$ is not zero, and suggests that the significant vector is not the displacement $\boldsymbol{D}$ but the difference between $\boldsymbol{D}$ and $\epsilon_0\boldsymbol{E}$ that vanishes in vacuum. A new vector $\boldsymbol{P}$ is therefore defined so that

$$\boldsymbol{D} = \epsilon_0\boldsymbol{E} + \boldsymbol{P}.$$

For reasons that will eventually become clear, this quantity $\boldsymbol{P}$ is called the polarization. Since $\nabla \cdot \boldsymbol{D} = \rho$, the result follows,

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0} - \frac{1}{\epsilon_0}\nabla \cdot \boldsymbol{P}.$$

It would, then, appear that variations in polarization from point to point act as sources of electric field in exactly the same way as charge.

It is known that the mobile charges in conductors, and the charges that are transferred in frictional electrification, are described by ρ. It is therefore reasonable to suspect that polarization in some way describes the effects of the immobile or bound charges in insulators. A closer insight into this notion may be derived by considering a long, uniform, cylindrical rod of a dielectric in which there are no free charges and the polarization $\boldsymbol{P}$ is uniform and parallel to the axis of the rod. Because $\boldsymbol{P}$ in the rod is uniform, the divergence $\nabla \cdot \boldsymbol{P}$ is zero inside the rod but not zero at its end faces. If the notation $\rho_b = -\nabla \cdot \boldsymbol{P}$ is introduced so that $\nabla \cdot \boldsymbol{E} = \rho_b/\epsilon_0$, then the field $\boldsymbol{E}$ that arises from the effect of $\boldsymbol{P}$ will be the same as the field that arises from a charge density ρ_b. Now ρ_b is zero except at the end of the rod; so that the field $\boldsymbol{E}$ is the same as a field resulting from two charges, one at each end of the rod. In Figure 6 a coordinate system is defined for the problem, below which are two plots, the variation of $\boldsymbol{P}$ with x and the variation of the divergence $\nabla \cdot \boldsymbol{P} = \frac{\partial \boldsymbol{P}}{\partial x}$ with x. The total "charge" at each end of the rod is obtained by integrating $-A\frac{\partial \boldsymbol{P}}{\partial x}$ through a thin layer δx including the end faces of the rod. At the left end this gives $-AP$ and at the right end $+AP$. The electric field $\boldsymbol{E}$ from this polarized rod is therefore the same as the electric field of two charges $+AP$ and $-AP$ separated by a distance l; *i.e.*, a dipole of moment AlP. But Al is the volume of the rod. This suggests that $\boldsymbol{P}$ might be interpreted as the dipole moment per unit volume of the rod resulting from bound charges.

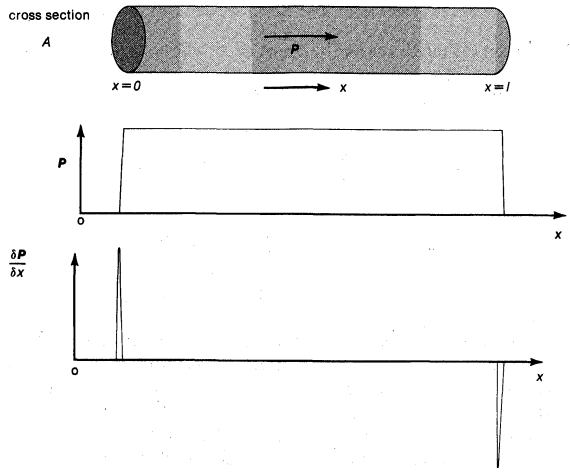


Figure 6: Polarized cylinder and graphs of the variation of $\boldsymbol{P}$ and $\delta\boldsymbol{P}/\delta x$ with distance (see text).

Polarization from the atomic point of view

It is instructive to look at the problem from an atomic point of view. If an electric field is applied to an insulator (*e.g.*, diamond or sulfur), the electrons in the atoms will be pulled in one direction by the field and the nuclei in the other. The centre of gravity of the negative electronic charge and the positive nuclear charge will no longer coincide; in other words, the atoms will have acquired a small electric dipole moment. If the field $\boldsymbol{E}$ points in the direction of increasing x, the positive nuclei will move toward increasing x and the induced moment $\boldsymbol{p}$ is parallel to $\boldsymbol{E}$. The relation between $\boldsymbol{p}$ and $\boldsymbol{E}$ is dealt with below; here it can be assumed that $\boldsymbol{p}$ is proportional to $\boldsymbol{E}$ and is written as $\boldsymbol{p} = \epsilon_0\alpha\boldsymbol{E}$.

The quantity α, called the atomic polarizability, has the dimensions of volume. If there are N atoms in unit volume, the dipole moment density (*i.e.*, the total dipole moment per unit volume) is $N\boldsymbol{p} = \epsilon_0 N\alpha\boldsymbol{E}$. If this were identified with the polarization vector $\boldsymbol{P}$ it would give

$$\boldsymbol{D} = \epsilon_0\boldsymbol{E} + \epsilon_0 N\alpha\boldsymbol{E}$$

and the relative dielectric constant would be

$$\epsilon = 1 + N\alpha.$$

The reasoning is plausible, but the question still remains why, if $\boldsymbol{P} = \epsilon_0 N\alpha\boldsymbol{E}$, or $\boldsymbol{P} = N\boldsymbol{p}$, the quantity $-\nabla \cdot \boldsymbol{P}$ behaves as though it were a source of electric field, just as a charge is a source of the field. To understand this, a simple one-dimensional example can be considered, but first it is necessary to make a clear distinction between

macroscopic and microscopic quantities. The fields and vectors considered here are quantities related to macroscopic—*i.e.*, laboratory-size—bodies. The atomic structure of a medium is a microscopic feature of the medium and on a macroscopic scale (or even on the scale of light waves) is unobservable. Thus, any reference to macroscopic quantities or macroscopic volume elements implies defining neither their position nor their extent with microscopic (*i.e.*, atomic) accuracy. The quantities $\boldsymbol{D}$, $\boldsymbol{E}$, $\boldsymbol{P}$, ρ, and $\rho_b = -\nabla \cdot \boldsymbol{P}$ are, on a microscopic scale, averages over rather vaguely defined regions containing many atoms. It may be assumed that a medium has a polarization $\boldsymbol{P}$, which is increasing with increasing x. In this case, it can be represented in atomic detail by a row of dipoles, increasing in size toward the right. For this representation, the divergence of the polarization,

$$\nabla \cdot \boldsymbol{P} = \frac{\partial \boldsymbol{P}}{\partial x}$$

is positive. Next, the net charge is determined somewhere along this row for a region of macroscopic length L the exact position of which is vague on an atomic scale. If the left end of L is at x and the right end at $x + L$ it must be assumed that x is chosen randomly. If the individual dipole charges are plus and minus q, the dipole lengths are $l(x)$, and the separation between dipoles is d, there will be zero charge in L whenever both ends of L lie between dipoles. The probability of this happening, if x is chosen randomly, is

$$\left(\frac{d - l(x)}{d}\right)\left(\frac{d - l(x + L)}{d}\right).$$

The charge will be $+q$ if the left end at x falls inside a dipole and the right end does not. This probability is

$$\frac{l(x)}{d}\left(\frac{d - l(x + L)}{d}\right).$$

The charge will be $-q$ if the right end intersects a dipole and the left does not. This has a probability

$$\frac{l(x + L)}{d}\left(\frac{d - l(x)}{d}\right).$$

The expected value (Q) of the charge in L is therefore

$$Q = q\left\{\frac{l(x)}{d}\left(\frac{d - l(x + L)}{d}\right) - \frac{l(x + L)}{d}\left(\frac{d - l(x)}{d}\right)\right\},$$

which simplifies to

$$Q = -\frac{q}{d}\{l(x + L) - l(x)\}.$$

Now if L, though large on an atomic scale, is not so large that $l(x + L)$ is appreciably different from $l(x)$, this can be approximated by

$$Q = -L\frac{\partial}{\partial x}\left(\frac{ql(x)}{d}\right).$$

But $\frac{ql(x)}{d}$ is just the dipole moment per unit length and Q/L is the expected macroscopic charge per unit length; so,

$$Q/L = -\delta P/\delta x.$$

This equation is the one-dimensional form of $\rho_b = -\nabla \cdot \boldsymbol{P}$. This treatment is not wholly satisfactory, but it at least avoids the contradictions implicit in the usual elementary accounts. A detailed and general mathematical proof that, if $\boldsymbol{P}$ is defined as the average dipole moment density, then $-\nabla \cdot \boldsymbol{P}$ acts as a source term for the macroscopically observable electric field (in exactly the same way as the free or mobile charge density), is not simple; indeed it has been given only quite recently. The result itself has been known and used since the beginning of the 20th century.

It can be seen, then, that if $\boldsymbol{D}$ is written as $\epsilon_0\boldsymbol{E} + \boldsymbol{P}$, the vector $\boldsymbol{P}$ has a simple relation to the structure and properties of the medium. It is important to realize that, in identifying $\boldsymbol{P}$ with the electric dipole moment density, caused by bound charges in the medium, it is not necessary to invoke any special properties of these charges nor to assume that $\boldsymbol{P}$ is proportional to $\boldsymbol{E}$. The interpretation of $\boldsymbol{P}$ in the equation $\boldsymbol{D} = \epsilon\boldsymbol{E} + \boldsymbol{P}$ is valid, either if $\boldsymbol{P}$ is not proportional to $\boldsymbol{E}$ or if $\boldsymbol{P}$ is not solely the result of the relative displacement of electrons and nuclei. In particular, in a medium such as sodium chloride (NaCl), part of the dipole moment density may be caused by the relative displacement of sodium (Na^+) and chlorine (Cl^-) ions.

In most media the polarization $\boldsymbol{P}$ is in fact almost exactly proportional to electric field $\boldsymbol{E}$. If $\boldsymbol{P} = \epsilon_0\chi\boldsymbol{E}$ the dimensionless constant χ is known as the dielectric or electric susceptibility or, if the context is clearly electric (as distinct from magnetic), as the susceptibility. From the expression $\boldsymbol{D} = \epsilon_0\boldsymbol{E} + \boldsymbol{P}$ and $\boldsymbol{D} = \epsilon\epsilon_0\boldsymbol{E}$ it is seen that $\epsilon = 1 + \chi$.

Electric susceptibility

By far the vast majority of gases, liquids, and solids have a linear constitutive relation; *i.e.*, $\boldsymbol{P}$ is proportional to $\boldsymbol{E}$. In these cases only $\boldsymbol{P} = \epsilon_0\chi\boldsymbol{E}$.

The polarization does not usually follow instantaneously on the application of a field $\boldsymbol{E}$; in a few cases the time lag (t) may be several seconds, although in the majority of media it is much shorter. Values ranging from 10^{-2} second down to 10^{-15} second occur in different types of media. Long times tend to be associated with viscous and highly polarizable liquids, short times with most other media. If the frequency of the electric field is $\omega/2\pi$ and the product ωt is small, the polarization is virtually instantaneous.

At high (optical) frequencies the relation between $\boldsymbol{P}$ and $\boldsymbol{E}$ can be anisotropic (*i.e.*, is dependent on direction) even in gases and liquids, and anisotropy is associated with the phenomenon of natural optical activity (*i.e.*, the rotation of the plane of polarization of light); but in static fields all gases, liquids, amorphous solids, glasses, and cubic crystals are isotropic—that is, the relation between $\boldsymbol{P}$ and $\boldsymbol{E}$ is independent of the direction of $\boldsymbol{E}$ in the medium. In crystals having other than cubic symmetry the relation is almost always direction-dependent.

In a few media (*i.e.*, crystals belonging to symmetry classes that lack a centre of symmetry) the constitutive relation may contain a small quadratic (squared) term. If for simplicity the vector character of $\boldsymbol{P}$ and $\boldsymbol{E}$ is omitted, then the relation is

$$P = \epsilon_0\chi E + \epsilon_0 dE^2.$$

Some media called ferroelectrics, such as barium titanate, exhibit a spontaneous polarization even in zero field. Other media, the pyroelectric crystals with polar symmetry, also have this property, but it is usually masked by other effects; *e.g.*, the accumulation of surface charge. Many crystals acquire a polarization $\boldsymbol{P}$ in zero field when they are subjected to strain. These media are known as piezoelectrics. The nature and origin of these effects are dealt with below. Here, it is enough to know that a medium can have a non-zero value of $\boldsymbol{P}$ even when $\boldsymbol{E} = 0$. The relation $\boldsymbol{P} = \epsilon_0\chi\boldsymbol{E}$, although extremely common, is not a basic law of electricity.

Fields in dielectrics. The basic equations that govern the behaviour of electric fields in dielectrics are four in number:

1. The electric field ($\boldsymbol{E}$) is equal to minus the gradient of the electric potential (ϕ):

$$\boldsymbol{E} = -\nabla\phi. \tag{15}$$

2. The curl of the electric field is zero:

$$\nabla \times \boldsymbol{E} = 0. \tag{16}$$

3. The electric displacement ($\boldsymbol{D}$) is equal to the dielectric constant (ϵ_0) times the electric field, plus the electric polarization ($\boldsymbol{P}$):

$$\boldsymbol{D} = \epsilon_0\boldsymbol{E} + \boldsymbol{P}.$$

4. The divergence of the electric displacement is equal to the density of electric charge (ρ):

$$\nabla \cdot \boldsymbol{D} = \rho. \tag{17}$$

If attention is restricted to regions free from mobile or free charge, then $\rho = 0$ and $\nabla \cdot \boldsymbol{D} = 0$. In simple linear dielectrics $\boldsymbol{P} = \epsilon_0 \chi \boldsymbol{E}$ and $\boldsymbol{D} = \epsilon_0(1 + \chi)\boldsymbol{E} = \epsilon_0 \epsilon \boldsymbol{E}$, but this cannot always be assumed to be the case. Equations (15) and (16) are both equivalent to the statement that the line integral of $\boldsymbol{E}$ around any closed path is zero. When this is applied to a path in Figure 7 (left), with two

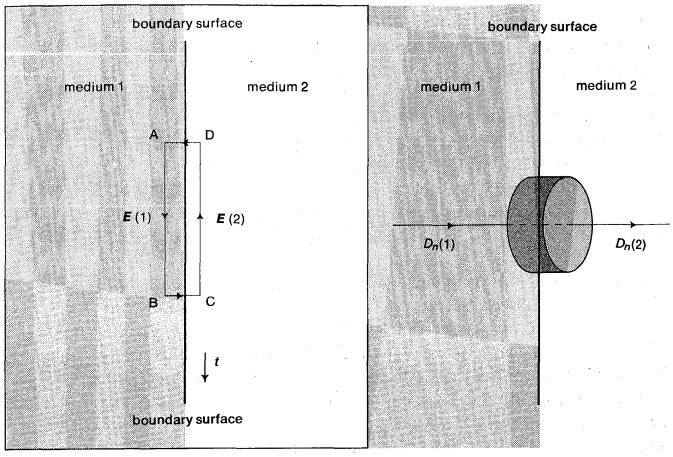


Figure 7: Components of field $\boldsymbol{E}$ and displacement $\boldsymbol{D}$ across a boundary surface. (Left) Tangential component of the electric field is continuous; (right) normal component (D_n) of the displacement vector is continuous (see text).

finite segments of either side of a surface separating two different media (1 and 2) and two infinitesimal segments crossing the surface, and if E_t is the component of $\boldsymbol{E}$ tangential to the surface, the vanishing of the line integral implies that

$$E_t(1) = E_t(2). \tag{18}$$

As the surface of a dielectric is crossed, the tangential component of the electric field is continuous. This implies that ϕ is also continuous.

Similarly, equation (17) implies that the surface integral of the normal component D_n of $\boldsymbol{D}$ over any closed surface in which there is no free charge is zero. If this is applied to the flat "pillbox" shown in Figure 7 (right), with faces in each medium, it may be seen that this requires $D_n(1) = D_n(2)$; for as the sides of the "pillbox" shrink to zero, all the contributions to the surface integral come from the faces, and this condition leads to the conclusion that at the boundary between two different dielectric media the normal component of $\boldsymbol{D}$ is continuous. Therefore,

$$D_n(1) = D_n(2). \tag{19}$$

Equations (18) and (19) are the key to the treatment of problems involving dielectric bodies. Before these relations are applied, two further results that follow from the basic equations can be written, namely,

$$\nabla \cdot \boldsymbol{E} = -\frac{1}{\epsilon_0} \nabla \cdot \boldsymbol{P} \tag{20}$$

and

$$\nabla \times \boldsymbol{D} = \nabla \times \boldsymbol{P}. \tag{21}$$

The significance of equation (20) has been referred to above and will be used below. Equation (21) is of no real use; it is given here only because its omission could suggest that the curl of $\boldsymbol{D}$ is always zero, which could lead to confusion.

A particularly simple problem, that of the field surrounding a dielectric sphere with a definite spontaneous uniform polarization $\boldsymbol{P}$, may now be analyzed. Because the system has a spherical surface of discontinuity, Cartesian coordinates are not really suitable, and it is useful to introduce a spherical polar coordinate system. The azimuthal or longitudinal angular coordinate in this set of coordinates is irrelevant, since the system has axial symmetry about the polar axis parallel to $\boldsymbol{P}$. If ϕ_i is the potential inside the sphere and ϕ_0 that outside the sphere, both potentials satisfy Laplace's equation (proposed by the French mathematician, Pierre-Simon Laplace), since $\nabla^2 \phi = -\nabla \cdot \boldsymbol{E} = \frac{1}{\epsilon_0} \nabla \cdot \boldsymbol{P}$ is zero in both regions ($\boldsymbol{P}$ is zero outside and constant inside). Further, the solutions of the equation must have axial symmetry and, in addition, $\phi(r, \theta = 0) = -\phi(r, \theta = \pi)$. There are only two such solutions of Laplace's equation, $\phi = Ar \cos\theta$ and $\phi = B \cos\theta / r^2$. The first solution is inapplicable outside the sphere (for $\phi \to \infty$ as $r \to \infty$). The second is inapplicable inside (for $\phi \to \infty$ as $r \to 0$). The appropriate solutions are, therefore, $\phi_i = Ar \cos\theta$ and $\phi_0 = B \cos\theta / r^2$, in which A and B are constants. The potential is to be continuous (because E_t is continuous) at the surface of the sphere (arbitrary radius r equal to sphere radius b) and so

$$Aa \cos\theta = B \cos\theta / a^2, \text{ or } B = a^3 A.$$

The further condition $D_n(i) = D_n(0)$ must also be satisfied. Now

$$E_n = E_r = -\frac{\partial \phi}{\partial r}, \text{ and so}$$

$$E_n(i) = -A \cos\theta$$

$$E_n(0) = \frac{2B \cos\theta}{r^3} = \frac{2B \cos\theta}{a^3} = 2A \cos\theta.$$

But also $D_n(i) = \epsilon_0 E_n(i) + P \cos\theta$

and $D_n(0) = \epsilon_0 E_n(0)$;

so that $\epsilon_0 A \cos\theta = \epsilon_0 A \cos\theta + P \cos\theta A$

which gives $A = \frac{P}{\epsilon_0}$.

This leads to $\phi_0 = \frac{1}{3} a^3 \frac{P \cos\theta}{\epsilon_0 r^3}$

and $\phi_i = \frac{1}{3\epsilon_0} Pr \cos\theta.$

In vector notation, if $\boldsymbol{r}$ is a vector displacement from the centre of the sphere, these results are

$$\phi_0 = \frac{4\pi}{3} a^3 \frac{\boldsymbol{P} \cdot \boldsymbol{r}}{4\pi \epsilon_0 r^3}$$

$$\phi_i = \frac{1}{3\epsilon_0} \boldsymbol{P} \cdot \boldsymbol{r}.$$

Thus ϕ_0 is the potential arising from a dipole of moment $\frac{4\pi}{3} a^3 \boldsymbol{P}$, but more interestingly, the internal field $\boldsymbol{E}(i)$ derived from ϕ is $-\frac{1}{3\epsilon_0} \boldsymbol{P}$. The field inside the sphere is opposed to $\boldsymbol{P}$ and therefore

$$\boldsymbol{D}(i) = \epsilon_0 \boldsymbol{E}(i) + \boldsymbol{P} = \frac{2}{3} \boldsymbol{P} = -2\epsilon_0 \boldsymbol{E}(i).$$

If now, in addition to the effects of $\boldsymbol{P}$, there is an external impressed field $\boldsymbol{E}(ext)$ parallel to $\boldsymbol{P}$, then fields are linearly additive

$$\boldsymbol{E}(i) \to \boldsymbol{E}(ext) - \frac{1}{3\epsilon_0} \boldsymbol{P}.$$

Further, if instead of there being a spontaneous polarization $\boldsymbol{P}$, the polarization is given by the linear relation $\boldsymbol{P} = \epsilon_0 \chi \boldsymbol{E}(i)$, there results the relation

$$\boldsymbol{P} = \epsilon_0 \chi \left(\boldsymbol{E}[ext] - \frac{1}{3\epsilon_0} \boldsymbol{P} \right),$$

$$\boldsymbol{P} = \frac{\chi}{1 + \frac{1}{3}\chi} \epsilon_0 \boldsymbol{E}(ext).$$

Even if χ is extremely large, $\boldsymbol{P}$ cannot exceed $3\epsilon_0 \boldsymbol{E}(ext)$. Also, it will be noticed that

$$\boldsymbol{E}(i) = \frac{\boldsymbol{E}(ext)}{1 + \frac{1}{3}\chi}$$

Spherical coordinates

and is less than $E(ext)$. Calculations such as this can be performed only for a limited class of bodies in which the polarization and the electric field are uniform and the surface of the body is an ellipsoid; *i.e.*, the Cartesian equation of the surface is

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1.$$

If the field is applied along a principal axis (*i.e.*, x, y, or z), the general result is that $E(i)=E(ext)-\frac{1}{\epsilon_0}\gamma_\alpha P$, in which γ_α is a dimensionless shape factor appropriate to the axis α (α stands for x, y, or z). These shape factors are generally known as demagnetizing factors because they are of rather greater importance in practical magnetism. The three principal shape factors, γ_x, γ_y, and γ_z, satisfy the equation $\gamma_x+\gamma_y+\gamma_z=1$.

Demagnetizing factors

If the body has axial symmetry about the z-axis ($\gamma_x=\gamma_y$), the shape factors γ_x and γ_y will each be equal to $\frac{1}{2}(1-\gamma_z)$. For a sphere in which x, y, z are equivalent, then $\gamma_x=\gamma_y=\gamma_z=\frac{1}{3}$, as was just seen. A long thin needle parallel to the z-axis approximates an ellipsoid. If E is parallel to the axis the condition E_t is continuous applied to the side of the needle and ensures that $E(i)=E(ext)$. Thus $\gamma_z=0$ and $\gamma_x=\gamma_y=\frac{1}{2}$. For a thin flat disk normal to the z-axis, with E parallel to the z-axis, the condition that D_n is continuous applies to the flat faces and ensures that $D(i)=\epsilon_0 E(ext)$; thus $E(i)=\frac{1}{\epsilon_0}(D[i]-P[i])=E(ext)-\frac{P_i}{\epsilon_0}$ and $\gamma_z=1$ so that $\gamma_x=\gamma_y=0$.

If a long dielectric needle is placed parallel to a uniform external field, E inside the needle is equal to the external field. If a flat disk is placed normal to a field, D, inside the disk, is $\epsilon_0 E(ext)$. In this case if $D=\epsilon\epsilon_0 E$ the result is $E(int)=\frac{1}{\epsilon}E(ext)$.

These results can, so to speak, be turned inside out to give the field inside a hollow cavity in a medium, where the electric field well away from the cavity is E and the polarization is P. If γ is the appropriate shape factor for the cavity then

$$E(cavity)=E+\frac{1}{\epsilon_0}\gamma P.$$

If the cavity is a sphere $\gamma=\frac{1}{3}$ and

$$E(cavity)=E+\frac{1}{3\epsilon_0}P. \tag{22}$$

If it is a long, needle-shaped cavity parallel to E the result is

$$E(cavity)=E; \tag{23}$$

and if it is a disk normal to E,

$$E(cavity)=E+\frac{1}{\epsilon_0}P$$

or

$$E(cavity)=\frac{1}{\epsilon_0}D. \tag{24}$$

In some older texts the results (23) and (24) are taken to explain the significance of the vectors E and D in a dielectric. There is, however, no simple physical case in which they are applicable. The result (22) for a spherical cavity is, however, of some importance. If the field acting on an atom in a dielectric medium is regarded as the field in a spherical cavity surrounding the atom, and if the atom has an atomic polarizability α, the induced dipole moment will not be $\epsilon_0 E$ but $p=\epsilon_0\alpha(E+\frac{1}{3\epsilon_0}P)$.

But because the polarization P is equal to N, the number of atoms per unit volume, times the induced dipole moment p, or $P=Np$, there results

$$P=\epsilon_0 N\alpha\left(E+\frac{1}{3\epsilon_0}P\right);$$

so that

$$P=\epsilon_0\frac{N\alpha}{1-\frac{1}{3}N\alpha}E.$$

Also,

$$\chi=P/\epsilon_0 E,$$

giving

$$\chi=\frac{N\alpha}{1-\frac{1}{3}N\alpha}. \tag{25}$$

Substituting $\chi=\epsilon-1$ in the last equation, the following relationship is found:

$$\frac{\epsilon-1}{\epsilon+2}=\frac{1}{3}N\alpha, \tag{26}$$

a result known as the Clausius-Mossotti equation. The argument leading to equations (25) and (26) has been expressed in much more cogent form and treated almost continuously under the general heading of "local field corrections" since around 1920. It is probably fair to say that there is still no general agreement about the range of validity of these results.

If equation (26) is believed to be correct and α to be a constant atomic parameter, then, in a gas in which N can be varied by varying the pressure, $\frac{\epsilon-1}{\epsilon+2}$ should be proportional to pressure. This is found to be the case experimentally. On the other hand, $\epsilon-1$ is small for gases and $\frac{\epsilon-1}{\epsilon+2}\simeq\frac{1}{3}(\epsilon-1)$, and in practice the simple expression $\epsilon-1=N\alpha$ gives almost as good a fit to the experimental data.

The next step is to consider the energy associated with the field in a dielectric. The work done to change a distribution of free charge ρ by an increment $d\rho$ is

$$dW=\int(\phi d\rho)dV;$$

but now $d\rho=\nabla\cdot(dD)$, and so

$$\phi d\rho=\phi\nabla\cdot(dD)=\nabla(\phi dD)-dD\cdot\nabla\phi.$$

The term $\nabla\cdot(\phi dD)$, when integrated over a volume V with a closed surface S, gives a surface integral that vanishes as S recedes to infinity, and therefore

$$dW=-\int_V dD\cdot\nabla\phi dV=\int_V E\cdot(dD)\,dV.$$

It will be noticed, however, that the volume of integration is the whole of space, and so dW includes contributions from outside the dielectric when $D=\epsilon_0 E$. If the dielectric is linear and $D=\epsilon\epsilon_0 E$ in the body, dW can be integrated and gives $W=\frac{1}{2}\int E\cdot D\,dV$, so that it is possible to regard the scalar product $\frac{1}{2}E\cdot D$ as the field energy density. In a dielectric with $\epsilon=10$ and a breakdown strength of 5×10^7 volts per metre, the maximum energy density is about 10^8 joules per cubic metre. This amount is negligible compared with the thermal energy density of a solid, about 10^8 joules per cubic metre at room temperature, or the energy of combustion of fuel oil, about 10^{10} joules per cubic metre. Energy storage in dielectrics is not an attractive practical proposition for most purposes. Further consideration of energy in dielectrics will be found below under *Force, energy, and power associated with fields.*

Storage capacities compared for heat and electrical energy in a capacitor

Polarization. *Dielectric properties and the structure of matter.* There are essentially six ways in which a material body can acquire a dielectric polarization. In some solids—for example, beryllium oxide (BeO)—the positively charged beryllium (Be) atoms and negatively charged oxygen (O) atoms are arranged in a crystal lattice in such a way that the medium as a whole has a built-in polarization that cannot be appreciably changed without completely disrupting the crystal structure. Media of this type may be referred to as pyroelectric media because, as noted below, the only easily observable effects of the built-in polarization occur when the medium is heated. In other solids—*e.g.*, barium titanate ($BaTiO_3$)—the crystal structure also leads to a built-in polariza-

tion, but, in some temperature ranges at least, a sufficiently strongly applied electric field can reverse the polarization without disrupting the crystal. A medium with this property is known as a ferroelectric. Many media, although normally unpolarized, when mechanically strained (which displaces the atoms in the structure) acquire a polarization even in zero field. Media with this property are called piezoelectric. The effect is of some practical importance, especially in electronics.

Many simple molecules—*e.g.*, those of hydrogen chloride (HCl)—have a permanent molecular dipole moment resulting from the positive charge on the hydrogen atom and the negative charge on the chlorine atom. In a gas or liquid the forces of thermal disorder result in a completely random orientation of the individual dipoles, and no net polarization of the medium as a whole is observed. If, however, a field is applied, the dipoles tend to align themselves parallel to the field, and there results a polarization $\boldsymbol{P}$ parallel to $\boldsymbol{E}$, which is proportional to $\boldsymbol{E}$ for moderate fields. This leads to linear dielectric behaviour, but, because the degree of alignment is controlled by thermal disorder, the effects of $\boldsymbol{E}$ are strongly temperature dependent.

When a compound such as hydrogen chloride is condensed to form a solid crystal, the free rotation of the molecules will cease. When a field is applied to the crystal, instead of the molecules being rotated to align their moments parallel to the field, the only effect will be an elastic displacement of the positive and negative ions away from their position of equilibrium. This will result in an induced ionic dipole polarization. In sodium chloride (NaCl), for example, the lattice of alternating, equally spaced Na^+ and Cl^- atoms in each of the cubic directions leads to no net polarization; but if a field is applied, the distortion of the arrangement will lead to a nonvanishing value of polarization parallel and proportional to $\boldsymbol{E}$, the electric field.

In all matter, gaseous, liquid, or solid, ionic or nonionic, the application of a field displaces the atomic electrons relative to the atomic nuclei, and each atom acquires a dipole moment parallel to the field. This universal mechanism is known as electronic polarization. Because it involves only the displacement of electrons, whereas all the other mechanisms involve the displacement of relatively heavy atoms or ions, it gives the fastest response of polarization to electric field. The time delay involved is typically less than 10^{-15} second, whereas all other mechanisms have delays of 10^{-13} second or longer. Thus electronic polarization is the only mechanism operative at frequencies above 10^{14} hertz, the so-called optical frequencies.

Electronic polarization

Pyroelectricity, ferroelectricity, piezoelectricity, and related effects are dealt with in the following section. The next three sections deal in more detail with the three basic mechanisms, electronic polarization, ionic polarization, and polarization by dipole alignment, which lead to linear dielectric behaviour. Electronic polarization is presented first because it is present in all matter and often masks other effects.

Electronic polarization. In an electric field $\boldsymbol{E}$ an atom acquires a dipole moment $\boldsymbol{p} = \epsilon_0\alpha\boldsymbol{E}$ proportional to $\boldsymbol{E}$, because the field tends to pull the nucleus in one direction and the electrons in the other. The force binding the electrons to the nucleus is their ordinary electrostatic attraction. The effects of this attraction are modified by the fact that the atom is a dynamical system obeying the laws of quantum mechanics, but it is still meaningful to talk about the order of magnitude of the binding force in terms of the nuclear electric field acting on the electrons. An electron has a charge (e) equal to 1.6×10^{-19} coulomb, and its field E_a at a distance r_0 is of the order of

$$E_a \simeq \frac{e}{4\pi\epsilon_0 r_0^2}.$$

But in an atom, r_0 is of the order of atomic dimensions, roughly one angstrom, or 10^{-10} metre. This distance gives an E_a equal to 10^{11} volts per metre, or 10 volts per angstrom. The fields (E) normally dealt with in the laboratory are of the order of 10^8 volts per metre at most and, thus, produce only small effect on the actual electronic distribution in atoms. It is not unreasonable to guess that the resulting charge displacement r for a field $\boldsymbol{E}$ is of the order $r_0\boldsymbol{E}/E_a$, and the acquired dipole moment (p), which is the product of the charge times the distance of separation (*i.e.*, er), is therefore

$$er_0\frac{4\pi\epsilon_0 r_0^2}{e}\boldsymbol{E} \simeq 4\pi r_0^3\epsilon_0\boldsymbol{E}.$$

It may be supposed that in a solid the atoms are packed with their centres at the corners of cubes. The number of atoms N per unit volume will be $N \simeq \left(\frac{1}{2r_0}\right)^3$. Thus $\boldsymbol{P} = N\boldsymbol{p}$ is given by $\boldsymbol{P} = \frac{\pi}{2}\epsilon_0\boldsymbol{E}$.

This crude model therefore yields a susceptibility χ of the order of $\frac{\pi}{2} \simeq 1.5$. It follows that the dielectric constant ϵ is $1 + \chi \simeq 2.5$ and the refractive index n is $\epsilon^{1/2} \simeq 1.6$. In fact, the vast majority of simple, transparent, solid media have refractive indices (n) between 1.4 and 1.8; thus this crude model has at least some relation to reality. A quantum mechanical calculation of α yields an expression $\alpha = \eta 4\pi r_0^3$, in which η is a numerical factor not far from but usually less than unity, so that the result also has a more sophisticated theoretical interpretation.

The atoms in liquids are packed less tightly than those in solids, and as a result the quantities χ, ϵ, and n tend to be somewhat smaller. In a gas at standard temperature and pressure the density is typically 1,000 times less than that in a solid. This suggests that ϵ will be of the order of 1.0015. The actual values of ϵ for air and hydrogen are 1.0006 and 1.00026—*i.e.*, rather smaller than expected, because the quantum mechanical factor η is less than unity. The results for solids are better than those for liquids because in a solid, the product $N\alpha$ is of the order of unity and the local field correction cannot be neglected. This gives

Atomic packing in liquids

$$\chi \simeq \frac{N\alpha}{1 - \frac{1}{3}N\alpha},$$

and if, for example, $N\alpha = 1$, χ is increased to 1.5. The second reason is that r_0, the radius that determines α, is not the same as the radius that determines the packing. In many media—especially, for example, diamond and silicon—the electron clouds of neighbouring atoms overlap and r_0 is greater than the packing radius. This tends to increase χ.

It may be suspected that a more elaborate and accurate calculation of α would give better results. This is only partly right, for two reasons. It is remarkably difficult, except in rather trivial cases (*e.g.*, gaseous helium), to calculate α to better than two-figure accuracy; and, in addition, even if α is known to high accuracy, the general uncertainty attached to the local field theory, which replaces the field acting on atoms by $\boldsymbol{E} + \frac{1}{3\epsilon_0}\boldsymbol{P}$, would make the final result rather uncertain. The atomic theory of dielectric properties, especially of solids, is most usefully regarded as a descriptive theory rather than as an exact quantitative theory.

If the displacement of charge in atoms is proportional to the field, it then follows that the force binding electrons in atoms can be approximately treated as harmonic; *i.e.*, the restoring force is proportional to displacement. This means that the approximate classical equation of motion for a bound electron in a field $\boldsymbol{E}$ is $m\frac{d^2\boldsymbol{r}}{dt^2} + k\boldsymbol{r} = -e\boldsymbol{E}$, where m is the electron mass, k the force constant, $-e$ the electron charge, and $\frac{d^2\boldsymbol{r}}{dt^2}$ the acceleration along $\boldsymbol{r}$. (It is assumed that the much heavier nuclei remain almost stationary.) The restoring force is not exactly harmonic, and the residual effects of anharmonic terms in the force are, to a first approximation, equivalent to frictional or viscous damping. Atomic

vibrations and also atomic collisions also contribute to this term. This gives an equation of motion that can be expressed as

$$\frac{d^2r}{dt^2} + \omega_0\eta\frac{dr}{dt} + \omega_0^2 r = -\frac{e}{m}E, \tag{27}$$

in which $\omega_0 = \left(\frac{k}{m}\right)^{\frac{1}{2}}$

is the natural frequency of oscillation of the electronic system (strictly this is $\omega_0/2\pi$) and η measures the strength of the damping term.

If the applied field has a frequency $\omega/2\pi$ and varies as $E_0 \cos \omega t$, and equation (27) is solved for r and P is expressed as Ner, then the result is

$$P = \frac{Ne^2}{m}\frac{(\omega_0^2 - \omega^2)E\cos\omega t + \omega\omega_0\eta E\sin\omega t}{(\omega_0^2 - \omega^2)^2 + (\omega\omega_0\eta)^2}. \tag{28}$$

The polarization contains a term in phase with E and a term lagging by $\pi/2$, or 90°, which, as will be seen below, corresponds to power absorption and generates heat through the frictional term η. Not surprisingly, quantum mechanics leads to a similar expression and identifies ω_0 as the absorption frequency of the atoms. The in-phase part of (28) yields a susceptibility

$$\chi = \frac{Ne^2}{\epsilon_0 m}\frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + (\omega\omega_0\eta)^2},$$

which when plotted rises to a positive peak when ω is about $\omega_0(1 - \frac{1}{2}\eta)$ and falls to a negative peak when ω is about $\omega_0(1 + \frac{1}{2}\eta)$. When $\omega > \omega_0$ χ is negative and therefore

$$\chi \to -\frac{Ne^2}{\epsilon_0 m}\frac{1}{\omega^2} \to 0$$

as ω gets very large. At frequencies below ω_0 and well away from the resonance,

$$\chi = \frac{\frac{Ne^2}{\epsilon_0 m}}{\omega_0^2 - \omega^2} \tag{29}$$

and the static susceptibility is

$$\chi_0 = \frac{Ne^2}{\epsilon_0 m\omega_0^2}.$$

Thus χ at frequencies less than ω_0 can be expressed as

$$\chi(\omega) = \chi_0\frac{\omega_0^2}{\omega_0^2 - \omega^2}.$$

This result is in good agreement with the observed frequency dependence of the refractive index $\eta(\omega) = (1 + \chi[\omega])^{\frac{1}{2}}$. When the refractive index varies appreciably with frequency (wavelength) a medium is said to be dispersive. Equation (29), which gives a useful description of dispersion, is known as Sellmeier's equation.

Sellmeier's equation

Because electronic resonance frequencies generally lie in the visible or ultraviolet region of the spectrum (*i.e.*, above 10^{14} hertz), at ordinary radio frequencies (below 10^{12} hertz) the electronic contribution to the susceptibility of matter is almost constant and equal to the static value.

Ionic polarization. The forces that hold ions in place in a crystal lattice are primarily electrostatic. The reasoning about the electronic susceptibility that leads to the approximate result $\chi \simeq \frac{\pi}{2} \simeq 1.5$ can be repeated almost word for word. This would suggest that in a simple ionic crystal such as sodium chloride (NaCl) the susceptibility χ is approximately 3 with more or less equal electronic and ionic contributions. In practice the ionic term is usually somewhat larger.

If in unit volume there are N ions, each contributing α_i to the ionic susceptibility and α_e to the electronic susceptibility, and the local field polarizing the ions is $E + \frac{1}{\epsilon_0}\gamma P$ in which γ is the shape factor (*i.e.*, about $\frac{1}{3}$) for the particular crystal lattice, then

$$P = \frac{N(\alpha_i + \alpha_e)}{1 - \gamma N(\alpha_i + \alpha_e)}\epsilon_0 E$$

or

$$\chi = \frac{N(\alpha_i + \alpha_e)}{1 - \gamma N(\alpha_i + \alpha_e)}.$$

If $\gamma N(\alpha_i + \alpha_e)$ approaches unity, χ will be extremely large, even though $N(\alpha_i + \alpha_e)$ itself may not be large. Thus it is not astonishing that values of χ as high as 10^4 can be observed in some solids. If, however, $\gamma N(\alpha_i + \alpha_e)$ is greater than, or equal to, unity, the system is unstable and P is nonzero even in zero field. This can be put more logically by saying that any arrangement of the ions that leads to $\gamma N(\alpha_i + \alpha_e) \geq 1$ is unstable and will not occur. The crystal or solid will only be stable in some other crystal structure, which may or may not correspond to P not equal to zero in zero field. Clearly, however, if a structure leads to values of $\gamma N(\alpha_i + \alpha_e)$ nearly equal to unity, small changes in γ or N will lead to instability. Thus crystals with high values of χ are likely to transform to a ferroelectric structure as the temperature is varied.

The forces that bind ions to their equilibrium positions in a crystal lattice are of the same magnitude as the forces that bind electrons in atoms, but the masses of ions are in general 10^4 to 10^5 times greater. The resonance frequencies associated with ionic polarization therefore occur at frequencies 100 to 1,000 times lower and correspond to infrared or even, in some cases, microwave wavelengths. At frequencies well above a resonance the contribution of ionic processes to χ is negligible. Thus at visible wavelengths χ is represented almost entirely by the electronic term and the refractive index $n = \epsilon^{\frac{1}{2}} = (1 + \chi_e)^{\frac{1}{2}}$. This relationship can be used experimentally to sort out the relative importance of electronic and ionic effects.

Dipolar polarization. A molecule such as hydrogen chloride contains charges of order $+e$ and $-e$ separated by a distance of about one angstrom (10^{-10} metre). Its dipole moment is therefore of the order of 1.6×10^{-29} coulomb-metre. The unit used in almost all discussions of molecular dipole moments is the debye (named after a Dutch chemist, Peter Debye). This is 10^{-18} electrostatic unit and is equivalent to 3.3×10^{-30} coulomb-metre. Typically, molecules have either zero dipole moment—*e.g.*, oxygen (O_2)—or a moment of a few debyes. In an electric field E, the energy of a dipole (p) is $-p \cdot E$; thus, in a field of 10^6 volts per metre, the energy of a molecular dipole is of the order of 10^{-24} to 10^{-23} joule. Boltzmann's constant k is 1.38×10^{-23} joule per degree Kelvin, and at room temperature the mean thermal energy of molecules, kT, in which T is the Kelvin temperature, is about 4×10^{-21} joule. Thus, even in strong electric fields, the energy of interaction of a molecular dipole with the field is a small fraction of its thermal energy. It might be expected that the degree of alignment of the dipoles would be proportional to the ratio pE/kT, and, if there are N dipoles per unit volume, the polarization parallel to E would be given by $Np\frac{pE}{kT}$. An exact classical or quantum mechanical calculation gives

$$P = \frac{1}{3}Np\frac{pE}{kT};$$

and so

$$\chi = \frac{Np^2}{3\epsilon_0 kT}. \tag{30}$$

This equation, known as the Langevin–Debye law, shows that χ varies inversely with T. The law has the same form as Curie's law in magnetism. In a gas at standard temperature and pressure, with a dipole moment (p) equal to one debye, this gives $\chi \simeq 2 \times 10^{-3}$. Local field corrections are relatively unimportant, and both the temperature and pressure dependence are found experimentally to fit the Langevin–Debye equation. At high pressures, in which the number N of dipoles is increased, the local field correction becomes more important; and at densities corresponding to densities in the liquid and solid state, χ—predicted by equation (30)—exceeds three;

The Langevin–Debye law

and the simple Lorentz local field calculation (see equation 25) then gives

$$\chi = \frac{\frac{Np^2}{3\epsilon_0 kT}}{1 - \frac{1}{3}\frac{Np^2}{3\epsilon_0 kT}},$$

which is negative. These divergent results are not found experimentally; *e.g.*, for water, which is a dipolar liquid. The proper method of calculating the local field correction is uncertain, but an improvement on the Lorentz calculation by Lars Onsager has at least the merit that it yields a finite value of χ.

Pyroelectricity, piezoelectricity, and ferroelectricity. Many crystals—*e.g.*, zinc oxide (ZnO)—have structures that contain alternate positively and negatively charged layers of atoms. As a result the crystal has a built-in dielectric polarization P that can easily be of the order of one coulomb per square metre. The electric field within the crystal is then of the order of P/ϵ_0; *i.e.*, up to 10^{11} volts per metre. As a result any stray mobile charge within the crystal, on its surface, or in its surroundings migrates to the ends of the crystal in such a way as to neutralize the field and the external effects of P. If, however, the crystal temperature is changed, the resulting change in atomic positions alters P, and transient external effects can be observed in the interval between the change in P and the redistribution of the surface charge. This effect is known as pyroelectricity. It can be used to detect small temperature changes by their electrical effects and is the basis of some forms of detector for infrared radiation. It can occur only in crystals of particular symmetries, the polar classes, and is used by crystallographers to identify crystals as belonging to these classes.

Piezoelectricity

If a crystal with this type of structure is strained (*i.e.*, deformed by an external force) the change can lead to a change in polarization and external electric charge effects. This relationship between an applied force and an external charge is known as piezoelectricity. Piezoelectric effects resulting from simple compression are limited to the polar crystal classes, but piezoelectric effects resulting from more general types of strain (*e.g.*, shear) can occur in a much wider class of crystals, those that lack a centre of symmetry. Again, the presence of a piezoelectric effect is taken by crystallographers to be evidence for the absence of a centre of symmetry in a crystal structure. In the converse piezoelectric effect the application of an electric field produces strain. The two effects are important practically in electromechanical transducers. The crystal microphone, the crystal pickup, and the ultrasonic transducer are important examples. Perhaps the most important application of piezoelectricity is in the quartz crystal resonator, a device in which electrodes attached to a quartz crystal both produce strain in the crystal and detect the electrical effects of strain. If the voltage applied to the crystal coincides in frequency with a natural mechanical resonance of the crystal, the effects on an external circuit attached to the electrodes are large. Because mechanical resonances in quartz are sharp (*i.e.*, limited to a small frequency range) and lightly damped, quartz crystals are much used in electronics as frequency selective elements. The quality factor (see below *Alternating electric currents*) of a quartz crystal can be as high as 10^5 to 10^6 compared with 10^2 to 10^3 for electrical tuned circuits.

In some crystals the interaction between ionic and electronic polarization through the internal or local electric field makes the structure unstable if the atoms are stationary at their mean positions in the crystal lattice. At sufficiently high temperatures, however, the thermal vibration of the atoms about their mean positions tends to average out the internal fields, thus stabilizing the structure in a state of zero net polarization. The interaction is not simple in any real crystal, and the highly simplified and idealized treatment of the behaviour that follows is based on barium titanate ($BaTiO_3$). At high temperatures above 120° C the structure is cubic and unpolarized and the atomic arrangement is based on a repeat of a structure with one barium atom at the body centre of a cube, eight titanium atoms at the corners (each shared with eight cells), and 12 oxygen atoms (each shared with four cells). Thermal vibration stabilizes the structure so that the mean position of the negatively charged oxygen atoms is at the centre of a cube edge. Each cell and each atomic plane is neutral on the average, and the crystal is unpolarized. When the temperature is lowered below 120° C, the thermal vibration can no longer stabilize the structure and the oxygen atoms move spontaneously along one of the cube edge directions. There are six equivalent distortions possible along the $\pm x$, $\pm y$, or $\pm z$ axes. The new form of the crystal no longer has perfect cubic symmetry and exhibits a spontaneous polarization. At temperatures not too far below the ferroelectric transition temperature, or Curie point, the crystal can be switched from one state of polarization to another by an electric field which, though strong, is not strong enough to destroy the crystal structure. The actual behaviour of real crystals, including those of $BaTiO_3$, is far from being as simple as this, partly because the crystal splits up into oppositely polarized domains (see MAGNETISM) and partly because changes in polarization involve mechanical changes in the crystal structure. In general, ferroelectricity is not well understood, and explanations developed for one particular type of crystal (*e.g.*, $BaTiO_3$) fail when applied to another (*e.g.*, potassium dihydrogen phosphate, KH_2PO_4).

The effect is of some practical importance, in part because ferroelectric crystals can be used as memory elements in computers and in part because ferroelectric media generally display large values of linear dielectric susceptibility both above and below the Curie point.

Except for hysteresis effects in ferroelectric media (*i.e.*, the polarization P and electric field E are not in phase for rapid changes), in ferroelectricity the dielectric relation between P and E is almost exactly linear (*i.e.*, proportional to the first power). A few media, crystals that lack a centre of symmetry, do, however, also have a small quadratic dependence of P on E (*i.e.*, P is proportional to the second power of E), which arises from the fact that the forces binding ions in lattices, and especially electrons in atoms, are not exactly harmonic. This has interesting consequences of some technical importance. If, for example, $P = \epsilon_0\chi E + \epsilon_0 dE^2$, then in a field $E\cos\omega t$, because $\cos^2\omega t = \frac{1}{2}(1 + \cos 2\omega t)$, the polarization P will contain a component at the double frequency 2ω. This is known as second harmonic generation. It can, for example, be used to convert infrared radiation of wavelength 10^{-6} metre to visible radiation of wavelength $\frac{1}{2} \times 10^{-6}$ metre. If two fields $E\cos\omega t$ and a static field E_0 are applied, the polarization is $P = \epsilon_0(\chi + 2dE_0)E\cos\omega t +$ etc. The crystal now appears to have a linear polarizability, $\chi + 2dE_0$, which depends on the static field. If ω is an optical frequency, this gives a field-dependent refractive index and is known as the Pockels effect, or the linear electro-optic effect. It can be used to modulate or deflect a beam of light by the application of an electric field. It should not be confused with the much weaker quadratic electro-optic effect, or Kerr effect, which is mentioned briefly in a section below.

Second harmonic generation

Electrets. The built-in polarization that occurs in crystals of polar symmetry is little altered by external influences. Certain materials, especially waxes containing polar molecules, can be made to exhibit similar but smaller and more controllable effects and are known as electrets. If a molten dipolar wax is subjected to a strong electric field it acquires a net polarization as described above. If the wax is cooled well below the melting point in the presence of the field, this polarization can be "frozen in." If, in addition, the wax is a good insulator and hydrophobic, so that it does not collect surface moisture, this relatively weak frozen-in polarization will not only persist but will remain unmasked by surface changes and leakage. This is one form of electret, the electrical equivalent of a permanent magnet; it is not the only, or even the most important, mechanism that produces electrets. In carnauba wax, a common electret material, the polarization is almost certainly also due to separation of

ionic charge in the field applied during solidification. The electret gives a useful method of maintaining a static electric field over long periods. Electrostatic microphones using plastic electrets are commercially available, and there has been some work done on the use of electret membranes in medicine.

ELECTROSTATIC POTENTIAL

High-voltage phenomena. If two electrodes (*i.e.*, conductor terminals) in vacuum are maintained at a high potential difference, a current flows in the gap between them when the electric field at the surface of the negative electrode exceeds a certain threshold value, characteristic of the material of this electrode. The current in the conductors rises rapidly with increasing field, and, at high enough fields, the ordinary resistive heating of the electrode associated with the current can melt or vaporize the electrode. The emission of electrons from a negative electrode subject to a strong field is known as field emission, and is basically a quantum mechanical phenomenon. The essential features of the process are as follows: In Figure 8 the potential energy of an electron is plotted

Field emission

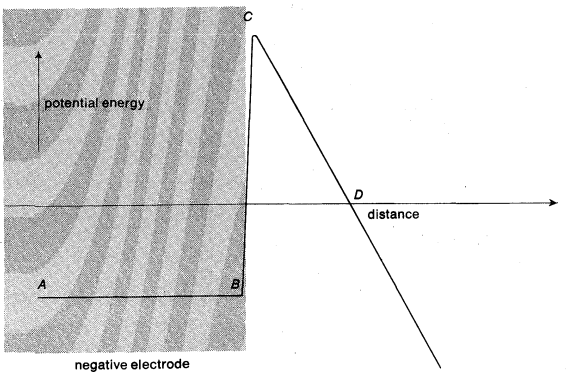


Figure 8: The potential energy of an electron at a metal surface subjected to a strong electric field (see text).

against distance near a metal surface. The potential energy within the metal (region AB of the negative electrode) is negative and rises abruptly at the metal surface. Electrons in the metal lack sufficient kinetic energy to surmount the barrier and thus remain bound in the metal. Outside the metal in the region CD and beyond, the potential energy falls rapidly because of applied field. The slope of the line CD, representing the change of energy with distance, is eE. As E is increased, the distance from D to a point on the metal surface decreases (if the electrons have a negative potential energy) and eventually becomes so small that quantum mechanical tunnelling can take the electrons through the potential barrier and out of the metal. This field emission current depends exponentially on the distance of D from the surface and, as the field increases and the distance decreases, the current rises rapidly.

Though field emission is an unwanted effect in many high-voltage vacuum tubes, it can also be put to good use. If the negative electrode surface has a radius of curvature r and the applied voltage between the electrodes is V, the field at the electrode surface, which is approximately V/r, can be made large, even if V is rather small, by decreasing r. It is possible to draw appreciable field emission currents, even when V is only 100 volts or less, if the electrodes are pointed. The field emission cathode is used as a source of electrons in many practical devices. The effect is also used in the field emission microscope and in a somewhat more complicated way in the field ion microscope, which is capable of resolving single atoms on a metal surface.

In a liquid or solid insulator a field of the order of one volt per angstrom, or 10^{10} volts per metre, is enough to disrupt individual atoms and usually leads to a catastrophic breakdown of the medium. In practice, breakdown usually occurs at fields a factor of 10 lower because microscopic non-uniformities (voids, dislocations, impurity atoms, etc.) cause local field concentrations. The reason that liquids (*e.g.*, oil) are generally used as insulators in high-voltage applications is that they are self-healing, but solids, in general, tend to have somewhat higher intrinsic breakdown fields. The most varied and interesting phenomena associated with breakdown occur in gases and are dealt with in the section *The conduction of electricity*, below.

Fields and potential distributions. *Solutions to Laplace's equation.* Given a definite charge distribution $\rho(\boldsymbol{r})$, the resulting potentials $\phi(\boldsymbol{R})$ and field $\boldsymbol{E}$ can be obtained directly from the relation

$$\phi(\boldsymbol{R}) = \int_V \frac{\rho(\boldsymbol{r})dV(\boldsymbol{r})}{4\pi\epsilon_0|\boldsymbol{R}-\boldsymbol{r}|}.$$

Electrostatic problems do not, however, usually take quite this simple form. More often a set of electrodes at definite potentials is given, and the potential near the electrodes is required; or both a set of electrodes and a specified set of charges are given. Although there are formal mathematical procedures for solving these problems, the techniques required are cumbersome and the results are involved. Instead, the results of a number of special problems are used to construct the desired solution of a new problem. Thus, much of the study of formal electrostatic problems consists of the laborious process of acquiring a series of known solutions; *e.g.*, solutions to Laplace's equation $\nabla^2\phi = 0$.

Method of images. A rather different method of solving electrostatic problems is the so-called method of images. It may be used, for example, where there is a point charge q at a distance $x = h$ above a conducting plane at zero potential, and it is necessary to find the potential at any point x in the region above the plane. This at first appears to be a difficult problem because the potential ϕ is zero in the plane where x is zero and thus $\nabla^2\phi = 0$ for $x > 0$, except at the point $(h,0,0)$, where the charge is located. If, however, an imaginary, or image, charge (denoted by $-q$) were placed at the same distance h *behind* the plane, this would mean that each point on the plane would be equidistant from q and $-q$ and the plane would be at zero potential as required. The potential at any position x,y,z—*i.e.*, $\phi(x,y,z)$—would then be

$$\phi(x, y, z) = \frac{q}{4\pi\epsilon_0}\left(\frac{1}{([x-h])^2+y^2+z^2)^{1/2}} - \frac{1}{([x+h^2])+y^2+z^2)^{1/2}}\right).$$

This equation satisfies the condition $\phi = 0$ on the plane; and for $x > 0$, $\nabla^2\phi = 0$, except at the point $(h,0,0,)$ where a detailed investigation shows that $\nabla^2\phi$ has a singularity corresponding to a charge q. Thus ϕ given by this expression is the solution that is sought. It is not valid below the plane where $\phi = 0$ simply because there is no real charge where x is equal to $-h$. A negative charge $(-q)$, however, actually resides on the surface of the plane.

This is the simplest of image problems, but similar techniques can be applied to more complicated situations and the image method is of considerable practical use.

Other methods. Two completely different techniques for solving electrostatic problems will now be considered, in which the fields are actually measured.

In a homogeneous conducting liquid the current density $(\boldsymbol{J})$ is related to the electric field $\boldsymbol{E} = -\nabla\phi$ by $\boldsymbol{J} = \sigma\boldsymbol{E}$, in which σ is the electrical conductivity of the liquid. The current density $\boldsymbol{J}$ satisfies the equation of charge conservation, $\nabla \cdot \boldsymbol{J} + \partial\rho/\partial t = 0$ ($\partial\rho/\partial t$ is the time rate of change of ρ), and, in a static situation, $\partial\rho/\partial t = 0$ so that $\nabla \cdot \boldsymbol{J} = 0$. Thus $\sigma\,\nabla \cdot \boldsymbol{E} = 0$ and $\nabla^2\phi = 0$. Further, if electrodes of much higher conductivity are inserted in the liquid they will provide equipotential surfaces. Thus the potential distribution in vacuum due to a set of electrodes held at potentials ϕ_1, ϕ_2, etc., can be found by immersing the whole electrode system in a liquid of low but finite conductivity (*e.g.*, saltwater) and picking up the potential point by using a wire probe insulated except at the point. This is known as the electrolytic trough method, and it and related methods are widely used, especially in

Electrolytic trough method

electron optics, to design electrode systems to produce a desired potential distribution.

A second method for determining an electric field is also based on the similarity between Laplace's equation and an equation that describes a different physical system—in this case, a uniformly stretched thin rubber sheet. This method is, however, applicable only to symmetrical systems in which ϕ is known to be independent of coordinate z so that the problem is essentially two-dimensional.

In a horizontal elastic sheet in uniform tension (T) the net downward force on an element of area $dx\,dy$ at a variable height h off the sheet is

$$T\left(\frac{\partial^2 h}{\partial x^2}+\frac{\partial^2 h}{\partial y^2}\right)dx\,dy.$$

If the sheet is to remain stationary this force must vanish. The transverse displacement h of the sheet thus satisfies Laplace's equation in two dimensions

$$\frac{\partial^2 h}{\partial x^2}+\frac{\partial^2 h}{\partial y^2}=0.$$

If the sheet is clamped in a horizontal frame and objects representing positive electrodes are pushed against it from underneath, they physically raise the sheet a distance proportional to their potentials, whereas objects representing negative electrodes push the sheet down; the contours of constant h on the sheet will be the equipotentials of the original problem. Further, if the positive and negative electrodes are interchanged with positive on top and negative below, a small ball rolled on the sheet will experience the same forces in the gravitational field as an electron moving in a corresponding electrostatic field. The rubber sheet model can therefore be used to investigate not only equipotentials but also electron trajectories in practical devices. Many vacuum devices, especially secondary emission multipliers, were originally designed in this way.

Electrostatic problems can be programmed for computer solution. The results so obtained, though possibly accurate, are seldom as useful in practical work or general theoretical work as either an approximate analytic solution in closed form or an analogue solution. Both analytic and analogue solutions give a "feel" for the nature of the problem that is lacking in a purely numerical solution.

III. Direct electrical currents

Although the study of electrostatics provides a basic conceptual structure on which to base the solution of electrical problems, it is, of itself and in practical terms, a relatively insignificant part of the whole subject of electricity.

BASIC PRINCIPLES

In electrostatics, charge, potential difference, or voltage, and current are defined and related to phenomena associated with forces acting on or in electrified bodies. Current is regarded as the rate of flow of charge associated with the motion of charged bodies, which may be generalized to include the flow of charge in conductors. Electrostatics is not concerned with processes that go on while charges are in motion, however, but only with the properties of the equilibrium charge configurations that result when the flow has ceased. The notion of potential is essentially described in terms of the work that would have to be done to produce this final equilibrium configuration by the mechanical transport of charge. When electric currents are dealt with, the electromotive forces (emf) that maintain the flow of charges are of much more concern than the potential concept. These electromotive forces are associated with the steady supply of energy to the system by external agencies; *e.g.*, batteries, electrostatic machines, thermocouples, and electromagnetic generators (dynamos). Most of the common laboratory sources of electromotive force are complex devices involving chemical reactions, rotating machinery, and so on. About the only feature they have in common is the existence of two conducting terminals across which the device maintains a potential difference even when current is drawn from the terminals. Because of this disparity it is sensible to try to relate the study of steady currents to electrostatic concepts by describing a possible, though impracticable, source of a steady electromotive force that involves only the simplest of electrostatic principles. Figure 9 shows a metal case in which a vertical

Sources of electromotive forces

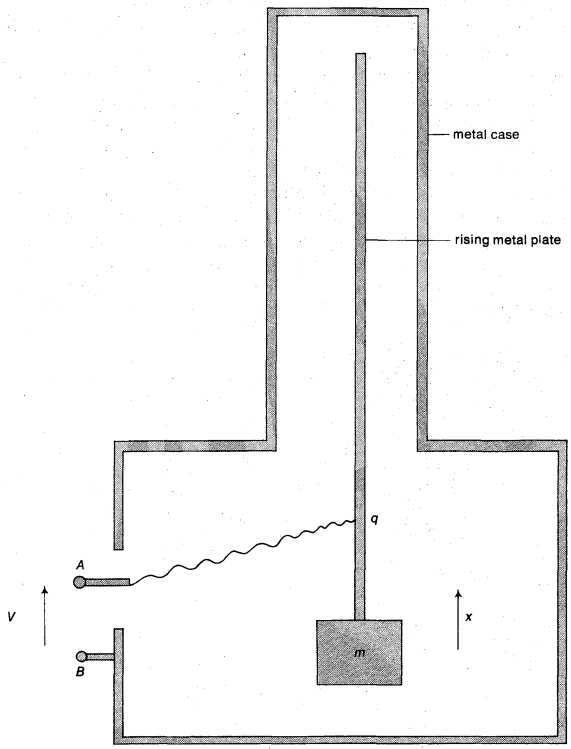


Figure 9: Electrostatic voltage source involving a movable charged plate inside a metal box or case (see text).

plate supporting a mass (m) moves freely in a vertical direction. The metallic case forms one terminal (B) of the device, and the other terminal (A) is connected by a flexible conducting lead to the movable plate. The capacitance C between the plate and the case can be expressed as $C=C_0+\gamma x$, in which C_0 and γ are constants and x is the height of the mass above some fixed reference point. If the movable plate carries a charge q, the electrostatic stored energy of the system is $\frac{1}{2}\frac{q^2}{C_0+\gamma x}$, and the mechanical stored energy is mgx. The total stored (*i.e.*, potential) energy is

$$U=mgx+\frac{1}{2}\frac{q^2}{C_0+\gamma x}.$$

According to the general laws of mechanics, the position of equilibrium of the plate (*i.e.*, the value of x) is the value that minimizes the potential energy, and is found by setting the differential dU/dx of the above equation for U equal to zero; so that

$$mg-\frac{1}{2}\frac{\gamma q^2}{(C_0+\gamma x)^2}=0.$$

Thus

$$C_0+\gamma x=q\left(\frac{\gamma}{2mg}\right)^{1/2}.$$

The potential difference between A and B is $\frac{q}{C_0+\gamma x}$, and if this is denoted by V, then, $V=\left(\frac{2mg}{\gamma}\right)^{1/2}$, which is independent of charge q. This device maintains a constant potential difference between A and B independent of q (until, of course, it is completely discharged

and the movable plate drops to the bottom of the case). When a current I is drawn from the terminals, the charge q decreases at a time rate dq/dt equal to $-I$ and the plate falls at a velocity (dx/dt), in which

$$\frac{dx}{dt} = -I\left(\frac{1}{2mg\gamma}\right)^{1/2}.$$

The mechanical potential energy decreases at a rate $mg(dx/dt)$, which is

$$mg\frac{dx}{dt} = -I\left(\frac{mg}{2\gamma}\right)^{1/2} = -\frac{1}{2}IV,$$

and at the same time the stored electrostatic energy decreases at a rate

$$\frac{1}{2}\frac{d}{dt}\left(\frac{q^2}{C_0 + \gamma x}\right) = \frac{1}{2}\frac{d}{dt}(qV) = \frac{1}{2}V\frac{dq}{dt} = -\frac{1}{2}IV.$$

Thus the total stored energy decreases at a rate $-IV$, and IV is the rate at which energy is delivered to whatever system is drawing current from the terminals.

This thoroughly impracticable device illustrates the basic property of a source of electromotive force. It has two terminals, A and B, across which it maintains a constant voltage or potential difference, and if current I is drawn from these terminals the device delivers energy at a rate IV.

Current measurement

Currents flowing in conductors can be measured by their magnetic or chemical effects, and in general it is found that (except in superconductors at low temperatures) if a current I flows in a conductor the potential V drops with distance along the conductor in the direction in which I is flowing. Thus if a voltage generator is connected to a load (a device that uses electrical energy) the potential difference V_L at the terminals of the load will be less than V_0 at the source. The power, defined as voltage times current, $V_L I$, delivered to the load will be less than the power $V_0 I$ delivered by the source. The difference in power $(V_0 - V_L)I$ has disappeared into the conductors. If the conductors are fixed material, such as wires, plates, or bars that cannot do external mechanical work by motion, the lost energy will be dissipated as heat. The rate of heat production (dH/dt) in the conductors is $dH/dt = I(V_0 - V_L)$, or

$$\frac{dH}{dt} = IV, \tag{31}$$

in which V represents the voltage drop $(V_0 - V)$ in the conductors.

For many conductors (*e.g.*, metals maintained at constant temperature) the voltage drop V is proportional to I. This is Ohm's law. It is written $V = RI$, in which the constant of proportionality, R, is the resistance of the conductor. Although few media other than metals obey Ohm's law exactly over appreciable ranges of I, nevertheless, the voltage drop is often almost linearly proportional to I. Inasmuch as $V = RI$, heat production as given by equation (31) can be expressed as $dH/dt = RI^2$, or, if preferred, as $dH/dt = V^2/R$.

In a conductor of uniform cross-sectional area A and length l, the resistance R of the conductor is found experimentally to be proportional to l and inversely proportional to A. The constant of proportionality, by convention written as ρ, is a property of the material known as its specific resistivity or just its resistivity. The resistance is $R = \rho l/A$. It is often convenient to describe a material in terms of its conductivity, $\sigma = 1/\rho$. In terms of σ this expression gives

$$R = \frac{l}{\sigma A}, \text{ or } G = \frac{1}{R} = \frac{\sigma A}{l}.$$

The quantity $G = 1/R$ is known as the conductance of the conductor.

In a conductor of cross section A carrying a current I the current density is $J = I/A$, and the voltage drop per unit length V/l gives the electric field E; therefore,

$$J = \frac{I}{A} = \frac{V}{RA} = \frac{V}{l}\frac{l}{RA} = \frac{V}{l}\sigma.$$

Thus $J = \sigma E$, or $E = \rho J$. This last statement can be generalized by considering filaments or tubes of current within a conductor parallel to the direction of current and expressed as a law in vector notation:

$$\boldsymbol{J} = \sigma \boldsymbol{E},$$

or

$$\boldsymbol{E} = \rho \boldsymbol{J}.$$

These equations are the most complete and general statement of Ohm's law. The general expression for the heat generated per unit volume then becomes

$$\frac{dH}{dt} = \boldsymbol{E} \cdot \boldsymbol{J} = \sigma E^2 = \rho J^2.$$

If a generator of electromotive force V_0 delivers current I and power $I V_0$, part of this power may be dissipated as ohmic (resistance) heating in a conductor, but part of it may be used to do mechanical work by causing the conductors to move against external forces (*e.g.*, the electric motor), part may be used to store energy in capacitors, or part may be used to store energy in chemical form (*e.g.*, electrolysis). Thus, in general, the power of a system is represented by

$$I V_0 = dH/dt + \sum_n (dW_n/dt),$$

in which $\sum_n (dW_n/dt)$ is the sum of all the various work and energy storage rates.

STEADY CURRENTS, MAGNETIC FIELDS, AND ELECTROMAGNETIC UNITS

Although, historically, magnetism (*q.v.*) developed independently and before the study of electricity as the science describing the behaviour of magnetized bodies, the laws relating magnetism to electricity are now regarded as the fundamental magnetic laws. Magnetized bodies stand in the same relation to these laws as polarized dielectrics to the fundamental laws of electrostatics. In these terms the most fundamental fact about magnetism is that conductors carrying electric currents in the vicinity of magnetized bodies or other conductors experience a force at right angles to the direction of current flow, being proportional to the current. This result is expressed in terms of a magnetic field acting on the current. The equation for the force (dF) acting on an element (dl) of a conductor carrying a current (I), when written in the form

Relation of magnetism to electric currents

$$d\boldsymbol{F} = I d\boldsymbol{l} \times \boldsymbol{B}, \tag{32}$$

serves as a definition of the magnetic field $(\boldsymbol{B})$ in much the same way as the expression $\boldsymbol{F} = q\boldsymbol{E}$ for the force acting on a charge q serves as a definition of the electrostatic field $\boldsymbol{E}$. The vector $\boldsymbol{B}$ will be referred to as the magnetic field, although some texts use magnetic induction for $\boldsymbol{B}$ and reserve the term field for a subsidiary magnetic vector $\boldsymbol{H}$.

In electrostatics the basic law that relates the field $\boldsymbol{E}(\boldsymbol{r})$ at $\boldsymbol{r}$ associated with a charge q at the origin is

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{q\boldsymbol{r}}{4\pi\epsilon_0 r^3},$$

and this expresses the fact that $\boldsymbol{E}$ acts along the line joining the charge to the point $\boldsymbol{r}$ and obeys an inverse square law. The corresponding magnetic result is more complicated. In magnitude, the field also obeys an inverse square law; but in direction, the element $d\boldsymbol{B}(\boldsymbol{r})$ of magnetic field caused by a current I in a line element $d\boldsymbol{l}$ at the origin is at right angles to both $\boldsymbol{r}$ and $d\boldsymbol{l}$. Thus the basic law is

$$d\boldsymbol{B}(\boldsymbol{r}) = \mu_0 \frac{I d\boldsymbol{l} \times \boldsymbol{r}}{4\pi r^3}. \tag{33}$$

Furthermore, the law must be expressed in differential form because no real conductor carrying a current can be straight in entirety; eventually, the conductor has to form a closed loop.

If there are two complete circuits (1) and (2), equations (32) and (33) can be used to express the force $(\boldsymbol{F}_{12})$ acting between the two circuits as

$$F_{12} = \frac{\mu_0}{4\pi} I_1 I_2 \int_1 \int_2 \frac{dl_1 \times (dl_2 \times r_{12})}{r_{12}^3},$$

in which the line integrals run over both circuits and r_{12} is the vector distance from the element dl_1 to the element dl_2. Given the geometric specification of the circuits, the integral can be evaluated with arbitrary accuracy. If, by connecting the circuits in series, the same current I flows in both circuits, then

$$F_{12} = \frac{\mu_0}{4\pi} I^2 G, \tag{34}$$

in which G is a precisely known geometric factor. Equation (34) now forms the basis for the international system of electric units, because, given the constant of proportionality μ_0, a definite value for a current I can always be expressed in terms of the force it exerts. If F is expressed in newtons (one newton is the force required to give one kilogram an acceleration of one metre per second per second) and μ_0 is given the value $\mu_0 = \frac{4\pi}{10^7} = 4\pi \times 10^{-7}$, the corresponding unit of current is the ampere (A). The unit of charge is the coulomb (C), or ampere-second, and the unit of potential is the volt (V), defined so that one ampere flowing for one second across a potential difference of one volt corresponds to an energy of one joule (J). All systems of electromagnetic units are now based on these definitions, but, in the SI or mks system of units the constants μ_0 and ϵ_0 are retained in all the equations. In SI and mks units all electrical quantities are expressed in terms of amperes, coulombs, and volts and all mechanical units in terms of metres, kilograms, and seconds.

The international system of electrical units

Once the constant μ_0 has been defined, the quantity ϵ_0 that appears in electrostatic equations must be obtained by experiment. Direct measurements of electrostatic forces are too difficult and inaccurate to serve for this purpose; however, Maxwell's equations relate the velocity of light in vacuum c to μ_0 and ϵ_0 by

$$c = (\mu_0 \epsilon_0)^{-1/2}; \tag{35}$$

and, because c can be measured with great accuracy, ϵ_0 is taken to be defined by

$$\epsilon_0 = \frac{1}{\mu_0 c^2} = \frac{10^7}{4\pi c^2}.$$

It is clear from equation (35) that μ_0, or ϵ_0, or both, must have dimensions. The conventional choice of dimensions is such that μ_0 is expressed in kilogram-metre-second^{-2}-ampere^{-2} and ϵ_0 in kilogram^{-1}-metre^{-3}-second^4-ampere^2. This gives $(\mu_0\epsilon_0)^{-1/2}$ in metre-second^{-1}, the units of velocity. The most usual expression of these dimensions is through the subsidiary quantities used to measure self-inductance, the henry (H), and capacitance, the farad (F). In these units,

$$\mu_0 = 4\pi \times 10^{-7} \text{ henry metre}^{-1}$$

and

$$\epsilon_0 = \frac{10^7}{4\pi c^2} \text{ farad metre}^{-1}.$$

The farad is the capacitance that stores one coulomb with a potential difference of one volt. The henry, considered in more detail below, is the inductance of a coil that produces a back electromotive force of one volt when the current increases at a rate of one ampere per second.

These definitions give a self-consistent and absolute set of electromagnetic units. The commonest laboratory instrument for measuring electrical quantities, the moving coil meter, relates rather directly to these definitions, for it indicates the value of a current by the force exerted on a conductor in a standardized magnetic field B produced by a permanent magnet system.

In terms of a current density (J), the two magnetic laws can be expressed in a slightly more general form. In a region in which the field is (B), the magnetic force (f_m) acting per unit volume is equal to the vector product of current density and magnetic field,

$$f_m = J \times B,$$

and the field $dB(r_1)$ resulting from currents in a volume element $dV(r_2)$ at r_2 is

$$dB(r_1) = \mu_0 \frac{J(r_2) \times r_{12}}{4\pi r_{12}^3} dV(r_2), \tag{36}$$

with vector $r_{12} = r_1 - r_2$.

If the current density J is produced by a charge of density ρ in motion with a velocity v, then $J = \rho v$ and

$$f_m = \rho v \times B.$$

When this is applied to a volume element containing a charge q, the magnetic force on the moving charge is

$$F_m = qv \times B.$$

If an electric field E is also present, the total force (F) is equal to the sum

$$F = q(E + v \times B).$$

This is Lorentz' force equation, and it finds widespread application in discussing the motion of electrons. If in some region a particle continues in rectilinear (straight-line) motion or experiences no net force, then

Lorentz' force equation

$$E + v \times B = 0,$$

and E must be at right angles to v and B and of magnitude vB.

Just as the equation

$$E = \frac{qr}{4\pi\epsilon_0 r^3}$$

yields the two electrostatic field equations

$$\nabla \times E = 0$$

and

$$\nabla \cdot E = \rho/\epsilon_0,$$

so equation (36) yields two magnetic field equations,

$$\nabla \cdot B = 0$$

and

$$\nabla \times B = \mu_0 J.$$

The equation $\nabla \times E = 0$ applies only to static fields, and similarly the equation $\nabla \times B = \mu_0 J$ applies only to steady fields produced by steady currents. How these equations are modified in time-dependent situations will be explained below.

ELECTROMOTIVE FORCE

A key concept in dealing with the flow of currents in conductors is clearly the idea of an electromotive force that drives the current round the circuit against the frictional forces associated with resistance. This section treats some of the possible sources of electromotive force.

Electrostatic electromotive force. The section on *Basic principles*, above, presented one possible way in which an electrostatic system could be used to obtain a steady potential difference between two terminals. This system, though easy to analyze, is not a practical source of electromotive force. On the other hand, a continuously operating machine such as the Van de Graaff generator can obviously be regarded as a source of electromotive force. This machine can deliver a continuous current in a circuit connected between the lower charging electrodes and the high-voltage terminal as long as the charge on the high-voltage terminal is continuously replenished by charge from the belt. The electrical energy available in the circuit is derived from the mechanical energy expended in driving the belt. In practice, electrostatic sources of emf are not much used, as their operation tends to be somewhat erratic.

Piezoelectric electromotive force. If a piezoelectric (see below *Piezoelectricity*) crystal is equipped with metal electrodes and then strained, a potential difference develops between the electrodes. This device can be regarded as the source of an electromotive force; but it should be noted that a constant current can be drawn from the electrodes, thus keeping a constant potential difference, only if the strain is continuously increased. This procedure is clearly likely to have catastrophic consequences for the crystal. In general, piezoelectric systems

can be used to produce only alternating emf's as the result of oscillating strain.

Thermionic electromotive force. If a conductor is heated, the mobile electrons within the conductor eventually acquire enough thermal kinetic energy to escape from the conductor into vacuum. For most conductors, however, the temperature required to produce any appreciable thermionic emission of electrons is so high that the conductor would be vaporized. A few materials, notably tungsten and barium oxide, are copious emitters of electrons well below the temperature at which the material evaporates. A tungsten cathode heated to about 2,300° K essentially boils off electrons, and these electrons have appreciable kinetic energies. The average energy associated with motion directed away from the metal surface is kT, in which k, Boltzmann's constant, is 1.38×10^{-23} joule-degree^{-1}, and the average energy per electron for the temperature is about 3.2×10^{-20} joule. The charge on an electron being 1.6×10^{-19} coulomb, its energy can be expressed as 0.2 electron volts (eV; one eV is the energy acquired by an electron in moving through a potential difference of one volt). The electrons from the cathode therefore have enough kinetic energy to move against a potential gradient and reach an electrode negative with respect to the hot cathode by an appreciable fraction of a volt. A vacuum diode consisting of a hot filament and negative electrode therefore can be used as a source of emf, converting heat to electrical energy.

Thermoelectric electromotive force. The thermionic electromotive force described in the preceding section is really a thermoelectric emf, and similar effects occur when charge carriers can move from one medium to another at a different temperature. If two different metals are in contact at the same temperature, electrons cross freely from one metal to the other in both directions. If the potential energy of the electrons in one metal, A, happens to be lower than in the other, B, there will be a net transfer of electrons from B to A, and A will acquire a negative charge and potential relative to B until the potential energies in A and B are equalized. This phenomenon is known as contact potential. If a closed circuit consists of a succession of metals $A, B, C, \ldots X, Y, Z, A$, the sum of all the contact potential differences $AB, BC, CD, \ldots YZ, ZA$ is zero, and no net emf operates in the circuit. If, however, the temperature of one junction, say BC, is changed, this will change the contact potential at this junction because it will alter conditions inside B and C; and this will give a net emf that can drive a current around the circuit. Thermoelectric effects are discussed in more detail in the section below, *Contact potential and thermoelectricity*.

Photoelectric electromotive forces. When light or ultraviolet radiation falls on a material body, and especially on a conductor, all the energy $h\nu$ (h is Planck's universal constant) in each incident photon of frequency ν may be transferred to single electrons in the material. These electrons may then have more than enough energy to escape from the conductor, and if they happen to start off near the surface they may, in fact, do so before their energy has been dissipated in collisions. If the conductor is placed in vacuum there will be a current of electrons leaving its surface, and in many cases the electrons may have kinetic energies corresponding to several electron volts. They can therefore move against a potential gradient and reach a negative electrode.

The solar cell

It is also possible to produce a photoelectric effect at the junction between two dissimilar conductors. The solar cell uses a junction between two different semiconductors. In this case, relatively high energy conversion efficiencies from light to electrical energy can be achieved. Again these effects are considered in more detail in the section *The conduction of electricity* below.

Chemical electromotive force. The effects discussed in the above sections on *Direct electrical currents*, though interesting and occasionally useful, are not, with the possible exception of the solar cell, of much practical importance as sources of electromotive force. The two effects described in this and the next section (*i.e.*, chemical batteries and dynamos) are by contrast the main practical devices for producing sustained electromotive forces. The chemical reaction $2HCl + Zn \rightarrow H_2 + ZnCl_2$, hydrogen chloride plus zinc produces hydrogen plus zinc chloride, releases energy. The total energy of the materials on the right-hand side of the reaction is less than that on the left. Normally this energy would be released as heat, but, if the reaction occurs in aqueous solution, it proceeds through stages, some involving charged ions.

The emf available from chemical batteries rarely exceeds 2.5 volts. If n batteries of emf V_0, however, are connected in series (*i.e.*, positive and negative terminals are connected successively), the total emf is increased to nV_0. The batteries used in the laboratory, home, or automobile usually have from two to 100 unit cells connected in series.

Electromagnetic electromotive force. If a material body moves at right angles to a magnetic field $\boldsymbol{B}$ the charged particles (electrons and nuclei) in the body experience a force at right angles to both $\boldsymbol{B}$ and the velocity $\boldsymbol{v}$. If the body is an insulator the charges will not move far under the influence of this force because they are bound to each other and held tightly in position in the body. If, however, the body contains mobile charge and is a conductor, the force will cause a bulk movement of the charge that will continue until the separation of negative and positive charge sets up a field $\boldsymbol{E}$ opposing the magnetic force. The charges then move in a new configuration fixed in the body with the velocity of the body. According to Lorentz' force equation, the field $\boldsymbol{E}$ is given by $\boldsymbol{E} = -\boldsymbol{v} \times \boldsymbol{B}$. This result may be applied to a straight wire of length l, shown in Figure 10 moving

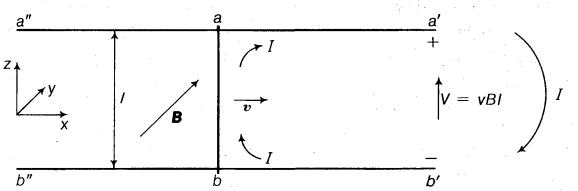


Figure 10: Generation of an electromagnetic emf. The wire ab moves along the rails $a''a'$ and $b''b'$ in the x-direction. The field $\boldsymbol{B}$ is directed into the page in the y-direction (see text).

in a field $\boldsymbol{B}$ that is directed along the y-axis (into the plane of the diagram). If the wire moves with velocity $\boldsymbol{v}$ in the xz plane parallel to the x-axis, then positive mobile charges will be driven by a force F_m toward a and negative charges toward b, and their presence sets up the field $\boldsymbol{E}$, having a magnitude vB and directed from a to b. Furthermore, if the moving wire is forced to slide along two stationary contact wires parallel to $\boldsymbol{v}$, as shown in Figure 10, then there will be a potential difference $V = El = vBl$ between the two wires. This difference can act as an electromotive force in a circuit connected between the terminal points a' and b'. If a current I flows in the external circuit, which therefore absorbs energy at a rate $IV = IvBl$, then the same current I can be said to flow also in the moving wire, and a force IBl will act on the wire in the direction opposite to $\boldsymbol{v}$. To continue in motion, the wire must be driven by an external force $\boldsymbol{F}$ of this magnitude (IBl) parallel to $\boldsymbol{v}$. The force does mechanical work on the system at a rate $\boldsymbol{F} \cdot \boldsymbol{v}$, which is equal to $IBlv = IV$. Thus the device converts mechanical energy to electrical energy. It is a simple dynamo.

The circuit of a simple dynamo

From Figure 10 it can be seen that the expression $V = vBl$ can be expressed as $-B(dA/dt)$, in which $-(dA/dt)$ is the rate at which the area (A) of the loop $b'\ b\ a\ a'\ b'$ is decreasing. This result can be generalized as follows: if there is a conducting loop in a steady magnetic field $\boldsymbol{B}$, and the configuration of the loop changes, either because the loop is rotating, being distorted, or simply translated as a rigid whole, and if the magnetic flux threading the area A of the loop is

$$\psi = \int_A B_n dS,$$

then the emf acting in the loop is

$$V = -\frac{d\psi}{dt}.$$

It is to be noted that the flux ψ can change either because the area of the loop normal to $\boldsymbol{B}$ is changing or because the loop is moving from one region to another where $\boldsymbol{B}$ has a different value. It should be noted also that this is *not* Faraday's law of electromagnetic induction: Faraday's law is concerned with the emf induced in a stationary loop by a changing magnetic field.

The equation $V = -(d\psi/dt)$ is useful in explaining the basic principle of a dynamo. A rectangular wire loop of area A can be imagined rotating with angular velocity ω about an axis perpendicular to the magnetic field $\boldsymbol{B}$. At any instant (t) the flux (ψ) linking the loop is $AB \cos \omega t$, and the emf available at the terminals of the loop is

$$V = -\frac{d\psi}{dt} = -\frac{\partial}{\partial t}(AB \cos \omega t) = \omega AB \sin \omega t.$$

If instead of a single loop there are n turns, the emf is

$$V = n\omega AB \sin \omega t.$$

(The value of sin ωt oscillates between $+1$ and -1.) This device therefore gives an alternating emf, and, by choosing n appropriately, the peak voltage can be raised to any reasonable amount.

Electromotive force and internal resistance. All practical generators of electromotive force involve real conductors of nonzero resistance. For example, in a battery the resistance to the flow of electric current is offered by both the chemical components of the cell (electrolyte) and the electrodes themselves. In a dynamo the windings have a finite resistance. As a result, when a current is drawn from a generator the electromotive force available at the actual physical terminals is equal to the difference between the emf when there is no load and the voltage drop in the internal resistance. The properties of a source of emf are specified in terms of (1) its output emf with no load (open-circuit voltage) and (2) its internal resistance. Thus, if these quantities are denoted by V_0 and R_i, the available emf when a current I is being drawn is $V_0 - IR_i$.

If a load attached to the terminals of a source of emf draws a current I, the power delivered to the load is $IV = I(V_0 - R_iI)$. This has a maximum value $V_0^2/4R_i$ when I is chosen to be $V_0/2R$. The available output power (P_{max}) of a generator is thus $P_{max} = V_0^2/4R_i$. The load that extracts this maximum power has $V/I = R_i$ and is thus equivalent to a resistance equal to the internal resistance of the generator. A typical dry cell used in a flashlight has V_0 equals 1.5 volts and $R_i \simeq \frac{1}{4}$ ohm; thus $P_{max} \simeq 2.25$ watts. As the cell ages, the value of R_i increases and usually reaches an unsatisfactorily high value before the value of V_0 has dropped appreciably. An open-circuit test of V_0 shows little about the state of a battery unless it is completely discharged.

DIRECT-CURRENT CIRCUITS

Ohm's law and the definition of resistance. Ohm's law states that the potential difference V across a conducting body is proportional to the current I flowing through the body; *i.e.*, $V = RI$. Though it is not an invariable law, it is true in enough cases to be useful. Because I has an absolute definition and is measured in an internationally maintained standard unit, whereas the product VI, being power (a mechanical quantity), leads to an equally absolute unit of V, it appears that the unit of R, the standard ohm, would be of little interest. This is not so. The measurement of electrical power with high enough accuracy for use in defining the standard volt is not practicable. The secondary electrical standard is instead the farad, and the standard ohm is obtained by comparing the impedance (see the section *Alternating electric currents*) of a standard capacitor for alternating current with a resistance. Because a resistance is physically easily reproduced whereas a standard capacitor, the capacitance of which can be calculated in geometric terms, is cumbersome and delicate, the ohm, so standardized, is in practice treated as the secondary standard. The volt is then defined as the potential difference across a standard ohm resistance carrying an absolute ampere. This standard is then used to calibrate a standard cell so that its open-circuit electromotive force becomes once removed from the secondary standard, a tertiary standard. Most laboratories base their electrical measurements on a standardized cell and a standard resistance, so that, in practice, currents are measured in terms of the potential difference they cause in a standard resistance. Absolute electrical measurements are undertaken at only a few national standards institutions such as the National Bureau of Standards in Washington, D.C., or the National Physical Laboratory in London. Instruments designed for normal laboratory use have an accuracy of much less than ten parts per million, which is expected of instruments in a standards institution (see CONSTANTS, PHYSICAL).

The standard ohm

The galvanometer, the ammeter, and voltmeter. Apart from a source of electromotive force, the next most important device needed in the experimental investigation of electrical currents is some sort of indicator of the presence of a current or a potential difference, and preferably one that gives a quantitative measure of current or potential difference. Although the device can be quite crude (*e.g.*, a lamp bulb) or can be a sophisticated digital electronic instrument, the commonest and most generally useful instrument is a moving coil galvanometer modified for use as either an ammeter or a voltmeter.

As its name implies, the moving coil galvanometer contains a coil that rotates on an axis that is perpendicular to a magnetic field, and it is the displacement of this coil against a mechanical restoring force, when a current flows through the coil, that indicates the presence, polarity, and strength of a current. The magnetic field is provided by a permanent magnet and is shaped so that the field acting on the coil is as nearly as possible the same for all positions of the coil. The magnet produces a radial magnetic field $\boldsymbol{B}$ in the air gap between the concave poles of a permanent magnet and a cylindrical soft iron core that is inside the coil. A rectangular coil is mounted so that it is free to pivot about the axis of this core, as shown in Figure 11. If the coil carries a current

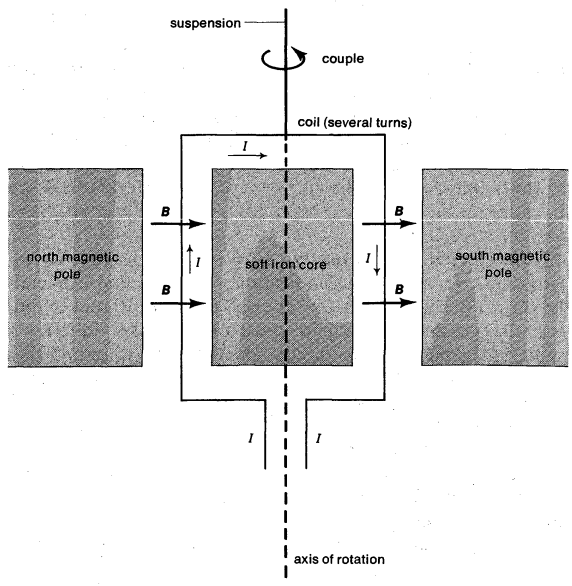


Figure 11: Galvanometer movement; coil with current I is suspended between two magnetic poles of field $\boldsymbol{B}$ (see text).

I in each of n turns, the forces acting on the two vertical sides of the coil produce a couple, which tends to turn the coil about the axis. The magnitude of the couple (Γ) is approximately given by $\Gamma = BAnI$, in which A is the area of the coil. In the most sensitive instruments the coil is suspended by conducting torsion fibres, which both act as current leads and produce a weak couple proportional to angular displacement to restore the coil to its rest position. In less sensitive and more robust instruments, the coil is mounted on pivots, and hair springs provide the restoring couple and act as leads. As long

as the magnetic field remains constant, the deflection is proportional to the current. In a well-designed instrument the deflection is linear with current to about one part in 1,000, and the instrument will maintain its absolute accuracy to a few parts in 1,000. The multimeters found in most laboratories usually have an overall accuracy of 1 or 2 percent. A sensitive torsion galvanometer will give a full-scale deflection for 10^{-8} ampere, but the most usual instruments have full-scale sensitivities of from 50×10^{-6}A to 10^{-2}A.

The ballistic galvanometer

The response of a galvanometer to transient current pulses depends critically on the moment of inertia of the coil and the damping of the motion caused by both the air and currents induced by the motion in the coil and its supporting frame. In the ballistic galvanometer these effects are used to produce an instrument having a maximum transient deflection proportional to the total charge in a current pulse. Today, transient currents are largely measured by electronic instruments.

The coil of a galvanometer necessarily has a finite resistance R_0 and so, if the deflection for a current I is kI, the coil deflection for an applied voltage V is kV/R_0. A galvanometer can be used as either an ammeter or a voltmeter if it will respond to the current or voltage it is to measure. Usually the resistance is small, and so most ammeters and voltmeters consist of a galvanometer with additional resistances.

For example, a galvanometer has an internal resistance R_0 and gives a full-scale deflection with a current I; then by connecting a suitable resistance R in series it will give a full-scale deflection with a voltage V provided $(R + R_0)I = V$. In a typical case, I is 50×10^{-6}A, R_0 is 500 ohms (symbolized by Ω), and if R is $(10^6 - 500)\Omega$ the meter will read full scale when V is 50V. In practice the accuracy of the meter is usually such that it will be adequate if $R = 10^6\Omega$, $\pm$ 0.5 percent.

To convert a galvanometer reading I full scale to an ammeter reading I_0 full scale, a resistance R is added in parallel. The current through R must be $I_0 - I$ and the potential difference across R is $R(I_0 - I)$, which must equal the potential difference, IR_0, across the meter. Thus $R = R_0I/(I_0 - I)$ is the correct value to use.

Standard ammeters and voltmeters of better than 1 percent accuracy are delicate and expensive instruments, and therefore it is often necessary to check the accuracy of a simpler instrument. The procedure using a potentiometer and a standard cell is outlined below under *The potentiometer*. The general theory of direct-current networks will now be developed.

Kirchhoff's laws and direct-current circuits. Kirchhoff's two circuit laws are essentially restatements of two basic electrostatic laws, the conservation of charge and the vanishing of any closed line integral of electric field. If in a complex circuit a number of conductors connected to the same point, or node, carry currents I_1, I_2, etc., toward the node, then charge (q) must be accumulating at the node at a rate (dq/dt) equal to the algebraic sum of the currents.

$$dq/dt = \sum_n I_n = I_1 + I_2 + I_3 + \text{, etc.}$$

But if the system is not changing with time, the potential at the node and the charge located at the node (if any) are constant. This gives the first of Kirchhoff's laws. In a direct-current circuit the sum of all the currents leading to a node is zero

$$\sum_n I_n = 0.$$

If any closed loop is taken in a complex circuit, the line integral of electric field round this loop must be zero.

The formula for the resistance total R of a set of resistances R_1, R_2, etc., in parallel (*i.e.*, a current branches into each resistance), $\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} +$ etc., can, of course, also be derived using Kirchhoff's laws. The idea of a voltage generator and a current generator can be obtained by imagining a circuit in which there is a complex source of electromotive force giving a voltage V across its terminals when the current in the external load R_L is I. Because the circuit equations that result from the application of Kirchhoff's laws are linear, V must be related to I by a linear equation $V = A + BI$, in which A and B are constant. If the load R_L is removed, V is the open-circuit voltage (V_{oc}) of the system; thus A is equal to V_{oc}. If R_L is replaced by a short circuit ($V = 0$) and I is the short-circuit current (I_{sc}), then the equation becomes $0 = V_{oc} + BI_{sc}$ and $B = -\frac{V_{oc}}{I_{sc}}$. The whole circuit can therefore be replaced by a generator of open-circuit voltage V_{oc} with an internal resistance R equal to V_{oc}/I_{sc}, called a voltage generator, and in which the resistance (R) is obtained experimentally by measuring V_{oc} and I_{sc}.

The notion of a voltage generator arises naturally from a consideration of the nature of physical sources of emf. An electrode in a vacuum tube collecting a beam of energetic electrons, however, is a different case. The current I in the external lead will be more or less independent of the electrode voltage. This leads to the concept of a current generator, which forces a current I through the external circuit across its terminals whatever the voltage across them. If the terminals are connected by a resistance R, the voltage developed is RI. An ideal current generator cannot actually exist and any current source must have some parallel leakage conductance (G) from one terminal to the other. Such a circuit is equivalent in every respect to having a voltage generator with open-circuit voltage $V = I/G$ and internal resistance $R = 1/G$. The available maximum power (P_{max}) from each network is

The current generator

$$P_{max} = \frac{1}{4}\frac{V^2}{R} = \frac{1}{4}\frac{I^2}{G}.$$

The direct application of Kirchhoff's laws to complex networks generally leads to a large number of simultaneous equations. Over the years a number of techniques for organizing the work and minimizing the number of equations have been developed. These techniques are not considered here, but it should be clear that the numerical solution of a large set of linear simultaneous equations is an ideal task for a digital computer. Most computer facilities have ready-made programs for just this task.

The potentiometer. If V is the potential difference between terminals E and G in Figure 12 (top), the potential difference between F and G is $V\frac{R_2}{R_1 + R_2}$; thus if V is known, voltages that are a known variable fraction of V can be produced by altering the values of R_1 and R_2. The circuit of Figure 12 (top) is the basic potentiometer circuit. The potentiometers used in radio as volume controls usually consist of a fixed resistor R with variable tapping point; *i.e.*, $R = R_1 + R_2$. Laboratory potentiometers can be constructed either with switched resistors or with a continuous variable tap on a slide wire. If the potentiometer is constructed with enough care, it is possible to measure the ratio of two voltages within an accuracy of a few parts per million. A typical potentiometer circuit used in the laboratory is shown in Figure 12 (bottom). A standard cell and a galvanometer (G_1) are connected across a fixed length (SS') of the slide wire, and the current (I) through the slide wire is varied using R until the galvanometer reads zero. The potential difference (V_S) between SS' is now equal to the standard cell electromotive force on no load. The unknown voltage (V) is then found by moving the slider (Y) until the galvanometer (G_2) reads zero. The source of V is now under no load, and V is obtained in terms of V_S by the ratio of the lengths XY and SS'; thus, $V = (XY/SS')V_S$.

To calibrate a voltmeter, the meter is placed across terminals PQ and, with an auxiliary circuit, the meter reading is set at, say, one volt and this is compared with the potentiometer reading. The potentiometer is a basic instrument in any electrical laboratory, and the foregoing method is useful for checking a meter whenever one is required to an accuracy of better than a few percent.

If the null reading galvanometer in a potentiometer cir-

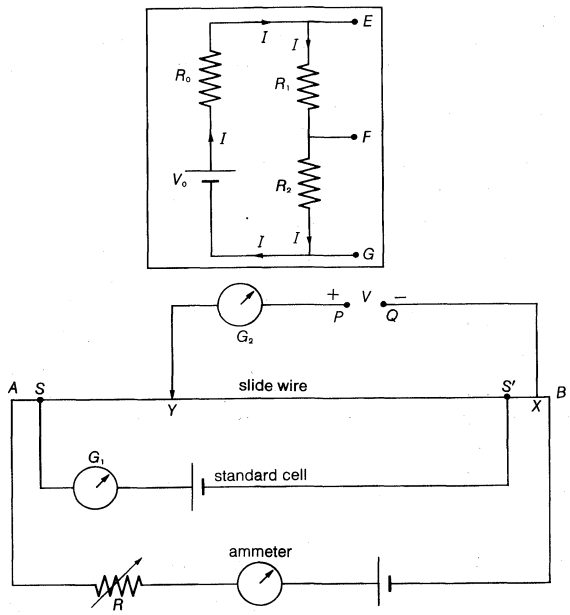


Figure 12: *Potentiometer.*
(Top) Schematic diagram; (bottom) potentiometer with a calibrated slide wire (see text).

cuit is replaced by an electronic amplifier, the output of this amplifier can be used to drive a servomotor that will automatically set the potentiometer tap at the null point. Self-nulling, direct-reading potentiometers are widely used in the laboratory. They are particularly convenient when a number of routine measurements have to be made.

The Wheatstone bridge. Figure 13 shows two potentiometers connected to form a Wheatstone bridge. The galvanometer reads zero when the ratio of resistances is such that

$$\frac{R_1}{R_2} = \frac{R_1'}{R_2'}.$$

If R_2 and the ratio R_1'/R_2' are known, this can be used to find the value of R_1. In one form of the instrument R_2 is a fixed resistance chosen to have a value near to R_1, and R_1' and R_2' are selected by switches or by plugging in fixed resistors. The Wheatstone bridge is a valuable instrument because its accuracy does not depend on the use of standard cells or even on having a steady source of electromotive force.

During the latter part of the 19th century a number of bridge circuits were designed for special applications. A typical example is the Kelvin double bridge used to compare two very low resistances. These special bridges are usually complicated to analyze and to operate, and their use has generally been replaced by simpler potentiometric methods. The simple Wheatstone bridge remains useful largely because it is cheap, easy to construct, and accurate enough for most laboratory purposes. In a section below a modification of this circuit is considered for use with alternating currents.

Kelvin double bridge

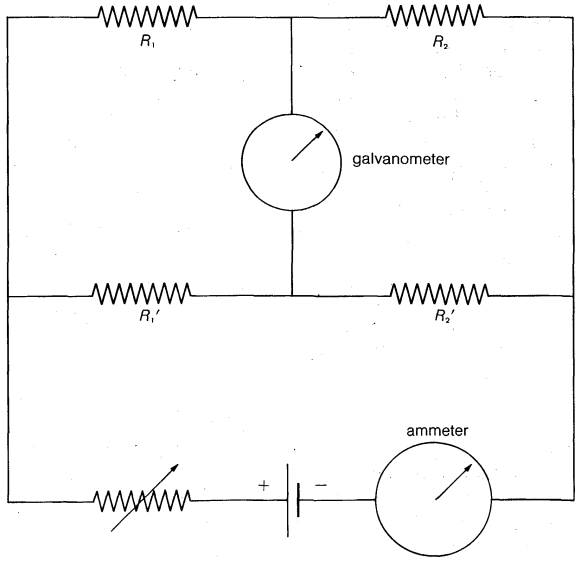


Figure 13: Wheatstone bridge (see text).

Nonlinear direct-current circuits. Only circuits in which Ohm's law is obeyed (*i.e.*, the potential difference across an element of a circuit is proportional to the current) have been dealt with so far. Many practical devices exhibit a more complicated behaviour. Most nonlinear devices are primarily of interest in connection with alternating currents and electronic circuits, but one familiar direct-current device, the electric lamp, displays such exceedingly nonlinear behaviour that it seems appropriate to end this section on direct-current circuits by considering the lamp's behaviour.

The common electric lamp bulb consists of a fine tungsten filament in an evacuated glass envelope. At the rated voltage, this filament will be at a temperature of some 2,800° C, and at this temperature the specific resistivity of tungsten is some 20 times its room temperature value. Thus if a standard 100-watt, 100-volt lamp, at 2,800° C carries one ampere, its resistance is 100 ohms (cold, its resistance is only five ohms). If, on the other hand, only 10 volts is applied to the lamp, the filament will barely heat up, the resistance will be slightly over five ohms, and the current will be greater than 0.1 ampere as it would be for a resistance of 100 ohms. Thus, if the current through the lamp is plotted against the applied voltage, the resulting curve will be nonlinear and bear little relation to Ohm's law.

A number of materials have specific resistivities that change rapidly with temperature, and these are frequently used to control or stabilize currents in, for example, electric motors. Although Ohm's law is so simple and so useful, it does not follow that it is obeyed by all media. Even if the temperature is held constant, the relation between I and V in some semiconductors, for example, is far from linear.

IV. The conduction of electricity

This section deals with electrical conduction in terms of the flow of charged particles and begins with a general treatment of the motion of charged particles in an electrical field. It then examines the principal types of electrical conductor, paying special attention to metals and semiconductors. In the two final sections of this part, other phenomena are considered (*e.g.*, thermoelectricity and photoelectricity) that are readily explicable in terms of the electron theory of conduction.

THE MOTION OF CHARGED PARTICLES IN AN ELECTRIC FIELD

If v is the velocity of a particle of mass m and charge q in an electrical field $\mathbf{E}$, the equation of motion is

$$m\frac{dv}{dt} = q\mathbf{E}.$$

The acceleration is parallel to the field, and, except in the presence of magnetic fields, there is little need to retain a three-dimensional vector description. For the time being, vector notation is omitted and it is assumed that E is parallel to the x-axis and all motion takes place in this direction. Thus the above equation becomes

$$m\frac{dv}{dt} = qE.$$

The three-dimensional equation $E = -\nabla\phi$ simplifies to $E = \frac{\partial\phi}{\partial x}$. Also, since $\frac{dv}{dt} = \frac{dv}{dx}\frac{dx}{dt} = v\frac{dv}{dx} = \frac{d}{dx}\left(\frac{1}{2}v^2\right)$,

the equation of motion can be written as

$$\frac{d}{dx}\left(\frac{1}{2}mv^2 + q\phi\right) = 0.$$

This is of course, a statement of the conservation of energy and gives

$$\frac{1}{2}mv^2 + q\phi = W_0,$$

in which W_0 is the total energy of the particle, a constant. These results are sufficient for an understanding of the behaviour of the motion of charged particles in vacuum. If, however, as in a metal, a gas discharge, or an electrolyte the particles of mass m move through a background of other particles they will be continually undergoing collisions, and this will obviously seriously affect the equations of motion. Some of these collisions will be violent and completely change the path and energy of the particle; others will be less violent and will result in only small changes; but both types of collision are likely to have sufficient effect to make it impossible to calculate the details of the motion of a simple particle. Instead, only the average motion of a large number of particles can be described. The general problem of setting up an approximate average equation of motion poses difficult and subtle questions in statistical mechanics which would be inappropriate to consider here. It is possible, however, to proceed in a more intuitive, and essentially correct, fashion as follows. The net effect of collisions is a change in the path of a particle and the average effect can be represented in terms of an average force F acting on the average particle. Thus the average equation of motion for the average velocity v becomes

Collision effects in the equation of motion

$$m\frac{dv}{dt} = qE + F.$$

It is reasonable to suppose that, as the average particle velocity increases, collisions will damp the motion and tend to limit the velocity. In other words, the effect of collisions is similar to friction. This effect could be represented by writing F as $-kv$, but it will be more convenient to write the relation as

$$F = -\frac{m}{\tau}v$$

so that the quantity τ has the dimensions of time. The equation of motion now becomes

$$\frac{dv}{dt} + \frac{v}{\tau} = \frac{q}{m}E.$$

If at time zero, the velocity equals zero, then the solution of this equation is

$$v = \frac{q\tau}{m}E(1 - e^{-t/\tau});$$

and as t increases v tends to a limiting value $(q/m)\tau E$. If τ is short this will happen almost instantaneously. If there are N charged particles per unit volume the current density is $J = Nqv$; and so

$$J = \frac{Nq^2\tau}{m}E.$$

Thus the current density J is proportional to E and the conductivity (σ) of the medium is

$$\sigma = \frac{Nq^2\tau}{m}. \qquad (37)$$

If, instead, $E = 0$, and at time $t = 0$ the velocity is v_0, then

$$v = v_0e^{-t/\tau};$$

so that τ can be interpreted as a relaxation time—*i.e.*, as the time it takes currents to decay from any specified value once the field has been turned off.

Instead of considering how the force F depends on velocity v it is possible to proceed in a different way. If the particles of mass m make collisions at a mean rate $1/\tau'$ per second so that τ' is the mean time between collisions and after each collision the velocity of the particle is reduced to zero, then between one collision at time $t = 0$ and the next, the velocity of the particle is qEt/m; and by the time it makes its collision at $t = \tau'$ it has travelled a distance $qE\tau'^2/2m$. Its mean velocity is

$$v = \frac{1}{2}\frac{q}{m}E\tau'; \qquad (38)$$

and, with N particles in unit volume, this gives a current density

$$J = Nqv = \frac{1}{2}\frac{Nq^2\tau'}{m}E$$

and a conductivity

$$\sigma = \frac{1}{2}\frac{Nq^2\tau'}{m}.$$

This equation will coincide with (37) if the mean collision time τ' is twice the relaxation time τ.

In most cases the particles have in addition to their mean drifting velocity v a random spread of velocities caused by thermal energy, etc.; and, in the case of electrons in metals, ions in an electrolyte, or electrons and ions in a high-pressure gas discharge, the random velocities are usually much bigger than the drift velocity. Thus, if the particles with which the drifting charged particles collide are on an average a distance l apart, it is plausible that τ', the collision time, should be of the order of l/u, in which u is the mean random velocity. This interval of time is not quite correct because the drifting charged particles do not necessarily collide violently with every particle that they encounter. This discrepancy can be accounted for by attributing to the particles a collision cross section s, such that, if particles coming within a distance r of each other undergo a violent collision, $s = \pi r^2$. If there are ν target particles per unit volume, a drifting particle makes, on the average, νs collisions per unit length of path. The mean free path (λ) between collisions is $\lambda = 1/\nu s$, and is inversely proportional to the density of target particles or collision centres.

Collision cross sections

The average charged particle drifts with a mean velocity given by equation (38), but now $\tau' = \lambda/u$, and so

$$\sigma = \frac{Nq^2\lambda}{2mu}.$$

Therefore, if the mean free path is relatively constant, σ decreases with increasing u, and, because the random velocity is expected to increase with temperature, σ will decrease with temperature. This expectation is, on the whole, fulfilled in metals, but in other cases the variation of N or λ with temperature is more important.

The treatment given above is, in all essentials, the standard approach used in understanding electrical conduction. It relates the bulk motion of the charge carriers to parameters such as N, q, m^* (effective mass), u, τ, τ', and λ, which might reasonably be expected to be characteristic of the medium and the charge carriers. A quantum mechanical approach makes little difference to the overall picture, but it does relate the parameters listed above to other, perhaps more fundamental, features of the structure of the medium.

In some cases (*e.g.*, conduction in an ionized gas) the simple picture thus developed is inadequate. The processes that go on are now too complicated to be described solely in terms of drift velocities, mean free paths, and the like. Almost equally complex phenomena can also occur in semiconductors, although fortunately most of the important practical applications of semiconductors involve relatively simple conduction processes.

The theory given above expresses the electrical conductivity in a form that may be written as

$$\sigma = Nq\mu,$$

in which N is the number of carriers of charge q per unit volume and μ would be the average velocity acquired by a carrier in unit field. The quantity μ is usually referred to as the mobility of the carriers and is much used in dealing with all types of conducting media, especially with semiconductors.

CONDUCTION IN VACUUM

A current can flow between two electrodes in vacuum only when at least one of the electrodes emits charged particles. Although radioactive materials may emit posi-

trons (positive electrons) or α-particles (positively charged helium nuclei) the emission expressed as a current is negligible (one curie yields about 5×10^{-9} ampere) and the only important cases are photoelectric, thermionic, or field emission of electrons. Because electrons have a negative charge, current will flow only when the emitter is the cathode (*i.e.*, negative electrode). If the rate of emission of electrons is small and the anode, or positive electrode, is at a high positive potential and near the cathode, all the emitted electrons will be collected by the anode and the current will be controlled by the rate of emission. If the emission is caused by the photoelectric effect the current will depend on the light intensity, whereas if the emission is thermionic it will be controlled by the cathode temperature. A diode operating because of thermal emission is said to be temperature limited when the current is controlled by the cathode temperature. Field emission can occur only when there is a strong field, usually in excess of 10^6 volts per metre at the cathode surface. In this case the emission depends critically on the field and is controlled by the anode potential. Neither photoelectric emission nor field emission usually leads to appreciable charge densities or space charge in the vacuum space between the electrodes. In the photoelectric case, this is because the emission current is usually small; in the field emission case, it is because the field required to cause emission sweeps the electrons rapidly away from the cathode toward the anode. The case is quite different with a thermionic cathode. If the cathode temperature is high enough, the emission can easily exceed 10^4 ampere per square metre (one ampere per square centimetre), and, if the anode to cathode potential difference is not too large, an appreciable negative electronic space charge can accumulate between the cathode and the anode. Because the space charge is negative it reduces the field at the cathode surface, and eventually the field drops to zero.

Dependence of field emission on field

Any additional electrons emitted by the cathode are then returned to the cathode, and the total current is controlled by the space charge. The diode is said to be space-charge limited. This is the usual case in practical devices such as the thermionic diode or triode. It is possible to calculate the maximum current that can flow between a plane cathode of area A with a plane anode at a positive potential V in front of the cathode and a distance d away from it. If d is much less than the transverse dimensions of the system, edge effects can be neglected and the problem is essentially one-dimensional. At a point at a distance x from the cathode, where the potential is ϕ, the velocity of the electrons is obtained from $\frac{1}{2}mv^2 = e\phi$ as $v = (2e\phi/m)^{1/2}$; and so, if the current density is J, the space charge is

$$\rho = -\frac{J}{v} = -\frac{J}{\left(\frac{2e\phi}{m}\right)^{1/2}}.$$

The potential (ϕ) satisfies Poisson's equation $\nabla^2\phi = -\rho/\epsilon_0$, and in one dimension this becomes

$$\frac{\partial^2\phi}{\partial x^2} = -\rho/\epsilon_0 = \frac{J}{\epsilon_0\left(\frac{2e\phi}{m}\right)^{1/2}}.$$

The desired solution of this equation gives $\phi = 0$ at the cathode surface, where $x = 0$, and also gives

$$\frac{\partial\phi}{\partial x} = -E = 0 \text{ at } x = 0$$

because the limiting current flows when the field E at the cathode surface is zero. It may be verified that the correct solution is

$$\phi = \left(\frac{9J}{4\epsilon_0\left[\frac{2e}{m}\right]^{1/2}}\right)^{2/3} x^{4/3}.$$

The potential, instead of increasing linearly with x, as it would do in the absence of space charge, increases as $x^{4/3}$. At x equal to d, however, the potential $\phi = V$. The resulting equation determines the limiting value of the current density as

$$J = \frac{4\epsilon_0\left(\frac{2e}{m}\right)^{1/2}}{9}\frac{V^{3/2}}{d^2}, \tag{39}$$

and so the total current is

$$I = AJ = 2.33 \times 10^{-6} V^{3/2}\frac{A}{d^2}$$

ampere, in which the numerical values of ϵ_0, e, and m have been used. This result, known as Child's law, states that the current I is proportional to the 3/2 power of the anode-to-cathode potential difference V. For example, if $A = 1$ square centimetre and $d = 1$ millimetre so that $A/d^2 = 10^2$, and if $V = 100$ volts, then $I = 0.233$ ampere (233 milliamperes). The 3/2 power law is not restricted to plane diodes but is a general result.

Child's law

This treatment has been idealized in one respect because it ignores the thermal energy of the electrons emitted from a hot cathode. If the cathode temperature is 1,000° K, the electrons will be emitted with a mean kinetic energy directed away from the cathode of the order of 1/10 electron volt. This additional electron energy does not make an appreciable difference to the result if the anode voltage exceeds about one volt, but at lower voltages the current is somewhat greater than predicted by equation (39). The effects of the initial thermal velocities are of significance mainly in determining the fluctuations or noise in the electron current. The transverse components of the thermal velocities also limit the maximum current density of an electron beam that can be focussed on the screen of a cathode-ray tube.

The current that flows between two electrodes in vacuum is transported by electrons, and it is of interest to consider the current that flows in the external circuit connected to the electrodes while the electrons are in transit in the vacuum space. Two electrodes, one at zero potential and the other maintained by a battery at a potential V, will produce a field $\boldsymbol{E}$ in the space between the electrodes; and, if there is an electron of charge $q = -e$ in the space with a velocity v, the field will do work on it at a rate $q\boldsymbol{E} \cdot \boldsymbol{v}$. This work must be done by the battery, and so the current I in the battery lead, while the electron is in transit, satisfies the equation $VI = q\boldsymbol{v} \cdot \boldsymbol{E}$, or $I = (q/V)\boldsymbol{v} \cdot \boldsymbol{E}$. If an electron starts at the zero potential electrode and proceeds to the positive electrode, charge will flow in the external circuit only while the electrons are in transit and not at the instant when they leave the cathode or the instant when they arrive at the anode. If, for example, the cathode and anode are one centimetre apart and the potential difference is 25 volts, the time of flight is approximately 10^{-8} second and the transit of each electron produces a pulse lasting 10^{-8} second (10 nanoseconds) in the external circuit.

This result is of fundamental importance in any consideration of high-frequency effects in thermionic devices, but it also has applications to semiconductor devices. Exactly the same argument applies to the currents in the external leads attached to a p–n junction diode and leads to the conclusion that the external current flows only while electrons (or holes) are in transit across the depletion layer at the junction.

In some systems (*e.g.*, a cathode-ray tube or an electron microscope) there is less interest in the current than there is in the trajectories of the electrons. Electron optics, or the study of electron trajectories, is a highly developed subject with an extensive literature and is dealt with in the articles ELECTRON; ELECTRON TUBE; MAGNETISM; PHOTOELECTRIC EFFECT; TELEVISION; and THERMIONIC DEVICES.

CONDUCTION IN GASES

Two essentially different types of electrical discharge in a gas can be distinguished: self-sustaining discharges in which, once the discharge has been initiated, enough electrons and ions are generated in the discharge to keep it going; and non-self-sustaining discharges in which the

electrons or ions must be continuously introduced from outside. The space-charge limited diode mentioned in the section above is a typical non-self-sustaining discharge. A somewhat more complicated type of non-self-sustaining discharge occurs in an electron tube having a hot emitting cathode and containing gas at low pressure.

At low anode-to-cathode voltages the current is reduced by collisions in the gas, which slow down the electrons and increase the space charge for a given current. As the anode voltage is increased, however, electrons eventually acquire enough kinetic energy to ionize gas atoms on impact. This ionization increases the number of electrons available and also forms positive ions that tend to neutralize the electronic space charge. Ions do this rather effectively because they have much larger mass and therefore move more slowly than electrons. At this stage, the current begins to increase rapidly with voltage. A condition is eventually reached in which the electronic space charge is almost entirely neutralized throughout the space between the cathode and anode, except near to the cathode where (1) the electrons are moving rather slowly and (2) the ions are being swept toward the cathode. If the voltage is increased still further, the current rises enormously and the number of collisions is so high that atoms can be ionized by multiple electronic collisions. At this time it is no longer necessary for the electrons to acquire enough kinetic energy to produce ions in a single collision. Thus, once this low-voltage arc phase has been initiated, the applied voltage can be reduced without putting out the discharge. The effect is used in the mercury-vapour thermionic rectifier, because it allows high currents of many amperes to flow even when the anode-to-cathode voltage is only 10 volts or so. Somewhat similar effects also occur in the mercury and tungsten arc lamps, the carbon arc, and the mercury pool rectifier. In these cases the electrons are supplied, either by thermionic emission from the electrodes heated by the energy dissipated in the arc or by field emission (see above) caused by the high potential gradients that can occur at the cathode surface.

The Townsend discharge

The simplest self-sustaining discharge is probably the Townsend discharge in a gas between two parallel metal plates. This is a low-current (a few microamperes) discharge at moderately high voltage in a gas at reduced pressure. Before this discharge is treated, a general remark about the stability of gas discharges is needed. In almost every case the voltage required to initiate a discharge is higher than the voltage needed to sustain it, and, in addition, the sustaining voltage decreases with increasing current. Thus to observe the Townsend discharge the current must be limited by a resistance (R) in series with the discharge tube.

In the Townsend discharge the primary ionizing event is accidental and attributable to stray X-rays, cosmic rays, and so on. The electron-ion pair formed in this event is pulled apart by the field, and the electron, colliding with gas molecules, then produces further ions and electrons. The number of pairs formed in this avalanche depends on the voltage and the pressure of the gas. The higher the voltage, the more energy the electrons pick up between collisions and the greater the ionization; on the other hand, the higher the pressure, the greater the number of atoms per unit volume, and as a result collisions are more frequent and the electron picks up less energy between collisions. If this were the only process then, once the avalanche was over, the discharge would go out; however, the ions formed in the discharge will hit the cathode and release further electrons. If the ions formed in a single avalanche release on the average more than one electron the discharge is self-sustaining.

In elastic collisions between electrons of mass m and atoms or ions of mass M approximately equal to $10^4\ m$, the average energy loss of the electrons per collision is the small fraction $2m/M$ approximately equal to 2×10^{-4}. Electrons formed at the cathode therefore drift toward the anode, acquiring energy from the field and losing little energy in collisions until they have enough energy, 10–20 electron volts, to excite or ionize atoms. The excited atoms emit light as they lose energy, and a luminous layer thus forms at a short distance from the cathode. The electrons lose most of their kinetic energy in this layer by inelastic collisions and then drift on toward the anode until again they have enough energy to excite atoms and another layer forms. These so-called Holst and Oosterhuis layers extend from the cathode to the anode.

At low currents, space charge is negligible; but if the external resistance (R) is reduced, allowing the current to increase from a few microamperes to a few milliamperes, space charge begins to be significant and, as in the thermionic arc, the potential drop in the discharge begins to concentrate in the region near the cathode. This phenomenon is called a glow discharge (see Figure 14). Electrons in a region near the cathode (the Crookes dark space) pick up energy from the field until they can excite and ionize atoms and form a luminous layer, the negative glow. By contrast with the Townsend discharge, there is sufficient ionization in this region to supply a neutral plasma (of electrons and ions of low velocity) to fill the tube out to the anode. The Holst and Oosterhuis layers disappear and, beyond the negative glow, there is a dark space extending out to the anode.

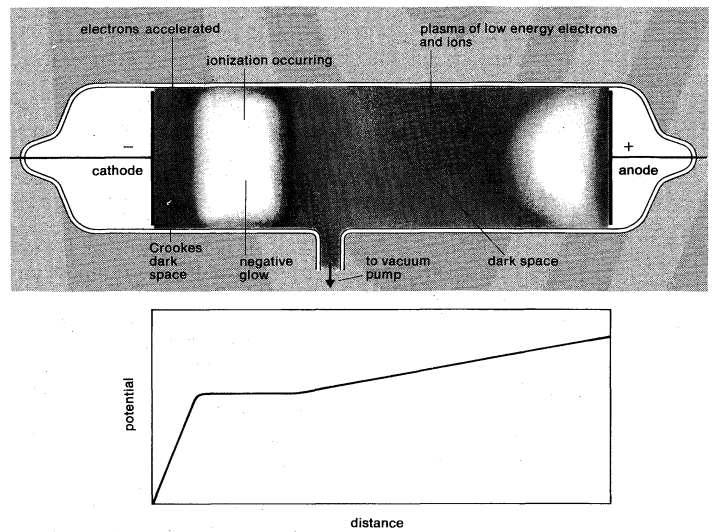


Figure 14: Conditions in a moderately intense discharge (see text).

The positive column

If the anode is moved farther away from the cathode, a new luminous region, the positive column, forms in front of the anode. In this region the electric field is small, but, if the field acts over a long enough distance, it can give electrons sufficient kinetic energy to ionize and excite atoms and produce a luminous region. Additional electrons and ions formed in the positive column are lost from the surface of the discharge to the walls of the vessel where they recombine.

The discharge configuration described above is stable only for a particular current density, determined by the nature of the gas, its pressure, and the configuration and material of the electrodes. Thus, variation of the current leads to a change in the area of the electrodes covered by the discharge and not to a change in current density and potential distribution. The voltage across the tube is almost independent of current as long as the whole cathode is not covered. This property of the glow discharge is important in the voltage stabilizing tubes used in electronics.

Once the current has reached a point at which the entire cathode is covered by the discharge, any further increase in current must result in a change in the potential distribution in the discharge. The field near the cathode increases, and, eventually, positive ions in this region acquire enough energy to excite atoms near the cathode surface. A luminous region then appears at the cathode surface. If a hole is made in the cathode the ions passing through the hole (canal rays) cause the cathode glow to extend behind the cathode.

As the current is further increased and the cathode fall in potential increases, ionization caused by positive ion collisions becomes significant; also the impact of energetic ions on the cathode may eject not only electrons but even atoms of the cathode material. This latter behaviour is known as sputtering. The ejected atoms themselves become ionized, and the cathode glow contains spectral lines characteristic of the cathode material rather than just the gas. Finally, as the current increases still further, the heating of the cathode from ion bombardment leads to thermionic emission, and the discharge takes, more and more, the form of an arc. The voltage across the electrodes collapses to a low value and eventually gives a typical low-voltage arc.

The form taken by the discharge in any particular practical case depends on the nature of the gas and the electrodes, the applied voltage, and the current density. It also depends on the pressure of the gas. Temperature does not of itself have much direct effect, except insofar as it alters the gas pressure and possibly facilitates thermionic emission. The effects of changes in pressure are rather more dramatic. At low currents, and therefore in the initiation of a discharge, the primary controlling parameter is the energy picked up by electrons between collisions, compared to the ionization potential ϕ_i of the gas. The ratio $E\lambda/\phi_i$, in which E is the electric field and λ the free path of electrons, is an important dimensionless parameter (a quantity without units). Because λ is inversely proportional to pressure (p), the ratio E/p is roughly constant. If, for a given electrode spacing, the gas pressure is doubled, the breakdown voltage more or less doubles. At high pressures in high fields, however, once the discharge has started, the energy density is much higher and the discharge tends to be altogether more violent. Simple and easily controlled discharges are, on the whole, confined to gases at pressures well below atmospheric.

The study of gas discharges is an important and flourishing branch of modern physics and generally considered under the heading of plasma physics. The reader will find relevant material in the articles on LIGHTING AND LIGHTING DEVICES; LIGHTNING; ELECTRON TUBE; THERMIONIC DEVICES; LASER AND MASER; and PLASMA STATE.

ELECTROLYTIC CONDUCTION

Apart from liquid metals, a few other pure liquids, notably molten salts, also conduct electricity. The flow is usually accompanied by chemical change. The most important case of this type of conduction, however, occurs in aqueous solutions. The purest water has a rather low electrical conductivity, about 10^{-4} mho per metre (reciprocal ohms per metre). Thus conductivity may be compared with metals for which a typical value is 10^7 to 10^8 mho per metre. The flow of electricity in water is usually accompanied by the evolution of hydrogen at the negative terminal and oxygen at the positive terminal. The addition of a salt such as potassium chloride increases the conductivity; thus, a normal solution (74.6 grams per litre) has a conductivity of 10 mho per metre.

Range of conductivity in materials

Electrolytic conduction is due to the motion of charged ions; *e.g.*, a hydrogen ion (H^+) or bare hydrogen nucleus, a positively charged potassium ion (K^+), a neutral potassium atom that has lost its single valence electron, a negatively charged chlorine ion (Cl^-), an atom with an extra valence electron, or a complex ion such as the phosphate ion (PO_4^{3-}). The conductivity is small for a number of reasons. First of all, the ionic concentration is small, usually less than 10^{20} per cubic centimetre, compared with the electron concentration in metals, about 10^{23} per cubic centimetre. Second, the ions move through a disordered background, with which they interact strongly, thus slowing down the drift of ionic charge. Third, the ions are massive compared with electrons and so do not acquire much velocity from the field between collisions.

The chemical aspects of electrolysis are treated in the articles on ELECTROCHEMICAL REACTIONS and ELECTROPLATING. The articles on LIQUID STATE and SOLUTIONS AND SOLUBILITY are relevant. This article concentrates on the basic features of the conduction process.

The atoms in an isolated water (H_2O) molecule are held together by electrostatic forces. The energy of the molecule is less than that of two hydrogen (H) atoms and one oxygen (O) atom because in the molecule the two valence electrons of the H atoms and the six valence electrons of the O atom can share the attractive potential of three positive nuclei. Repulsion between the nuclei and between the electrons, and quantum mechanical dynamic effects, somewhat reduces the binding energy and also determines the (triangular) shape of the molecule. The binding energy is so large that, in water vapour, hardly any molecules are thermally dissociated at normal temperatures.

The work required to remove a positive ion from a water molecule to produce H^+ and OH^- ions is of the order of $e^2/4\pi\epsilon_0 r$, in which r is a typical linear dimension of the water molecule. If r is taken to be 2Å (2×10^{-10} metre), the energy (W) is about five electron volts (5 eV) and the probability, $e^{-W/kT}$ ($k \simeq 10^{-4}$eV per degree is Boltzmann's constant), that a molecule will be dissociated at a temperature T is negligible. At 127° C, or 400° K, the probability is about e^{-120}, or 10^{-50}.

The work required in liquid water is entirely different. Once the hydrogen ion has been removed a small distance, it is in a medium with a high dielectric constant, $\epsilon = 80$, and the work required to remove it completely is reduced by a factor related to $1/\epsilon$. In fact, in pure water at room temperature, roughly one molecule in 10^9 dissociates; thus the work W must have been reduced to about 0.5 electron volt by dielectric effects.

In a potassium chloride crystal the potassium atoms are already present as K^+ ions and the chlorine atoms as Cl^- ions. The energy required to extract an electron from a potassium atom and attach it to a chlorine atom is positive, but it is more than counterbalanced by the strong attraction between the resulting ions in the crystal lattice. When potassium chloride (KCl) is dissolved in water, the attractive force between the ions is reduced, because of the high dielectric constant of water, and the salt goes into solution as free K^+ and Cl^- ions. In general, ionic crystals dissolve appreciably only in solvents of high dielectric constant, which means that usually the solvent molecules themselves have intrinsic, permanent, electric dipole moments and for this reason are classified as polar solvents.

Conduction with free ions in solution

With free ions in solution, the mechanism of conduction is obvious. In an electric field, positive ions drift toward the negative electrode and negative ions toward the positive electrode. Both processes conduct electricity from the positive to the negative electrode. It may happen that one ion has a higher mobility than the other, and in this case charge transport can be ascribed predominantly to the more mobile ion. This is especially the case when H^+ ions, which have some six times the mobility of the other common ions, are present.

As a conduction process, electrolytic conduction is rather uninteresting. There are no spectacular or even useful associated phenomena. Its importance is almost wholly in the realms of chemistry and biology. It does, however, have a profound historical significance for the physicist. J.J. Thomson's discovery of the electron really began with Faraday's discovery that, in electrolysis, the charge required to release one gram atom of a substance is always a small integral multiple of a basic unit. As atomic theory developed, this led to the obvious conclusion that electric charge itself might have a discrete structure.

CONDUCTION IN METALS

This section gives the classical theory of metallic conduction; it is followed by a look at how the theory is modified by quantum mechanics (see METALS, THEORY OF).

The classical theory of metallic conduction. There are two fundamental characteristics of metals. First, all metals have similar, and high, values of electrical conductivity, σ, in the range 10^6 to 10^8 ohm^{-1} (or mho) per metre; second, metals tend to occur in groups in the periodic table, toward the left of the table, where there

are relatively easily detached valence electrons. The first characteristic suggests that the mechanism of conduction is the same in all metals; the second suggests that the charge carriers are detached valence electrons. The classical theory of conduction is based on this hypothesis, and, because it leads to reasonable agreement with experiment, it is believed to be essentially correct.

The periodic potential

The classical model of a metal assumes that a cloud, or atmosphere, of mobile electrons moves against a background provided by the ionic cores of the atoms. These are arranged on a more or less rigid and regular crystal lattice. As the electrons move through the metal, they move through a periodic electric potential produced by the ions. In the classical theory the effects of this periodic potential are neglected and the electrons are regarded as moving in a region of constant positive (with respect to the outside world) potential. This potential, denoted by ϕ, binds the electrons in the metal. The work (W) required to move an electron from the influences of the lattice is the product of the electron charge (e) and the potential, or $e\phi$. Thus W is the work function of the metal expressed in electron volts. In quantum mechanics the periodic potential is dealt with more carefully. It is pointless to do so in classical theory, because if a more accurate theory is required it must include quantum effects as well. Despite this drawback, classical theory nonetheless gives a useful understanding of the basic processes of conduction.

It is reasonable to suppose that the number of mobile electrons contributed by each metal is, on average, about one. Each electron removed from a metal atom leaves it with an additional positive ionic charge e. Thus, even though aluminum (Al) has three valence electrons, it would be unlikely, on the classical model, that all three electrons would be released from each ion. For most metals the gram molar volume is between six and 25 cubic centimetres. Thus the number (N) of electrons, on the basis of one electron per atom and Avogadro's number (6×10^{23} molecules per gram mole), is of the order of 10^{22} to 10^{23} per cubic centimetre, or $N \simeq 10^{28}$ to 10^{29} per cubic metre. In a section above (see equation [37]) the conductivity was expressed as (setting $q = e$)

$$\sigma = N\frac{e^2\tau}{m}. \tag{40}$$

If $\sigma \simeq 10^7$ mho per metre and the numerical values of e and m are inserted, the result is $\tau \simeq 10^{-14}$ second. Because this time is so short, it is easy to understand why transient effects are not observed in metals as the electrons get moving. They reach their terminal drift velocity almost immediately.

If v is the drift velocity of the electrons, the current density is $J = Nqv = -Nev$. If J has the rather high value of 1,000 amperes per square centimetre, or 10^7 amperes per square metre, the numerical values of N and e give a velocity of one centimetre per second. Thus, at all usual current densities the drift velocity of electrons in a metal is small, at most a few centimetres per second.

The electrons in a metal are in thermal equilibrium, and classical kinetic theory predicts that their mean kinetic energy (in one dimension) would be $\frac{1}{2}mu^2 = \frac{1}{2}kT$, in which u is the average speed. If the temperature equals 300° K (27° C), then with the numerical values of m and k (Boltzmann's constant) the result is $u \simeq 10^5$ metres per second, which is much greater than any possible drift velocity. Because the relaxation time (τ) obtained from equation (40) is, to within a factor 2, the mean time between collisions, the mean free path (λ) between collisions is proportional to ut and approximately equal to 10^{-9} metre. In addition, if there are 2.7×10^{28} atoms per cubic metre, the separation of the atoms is

$$(2.7 \times 10^{28})^{-1/3} = \tfrac{1}{3} \times 10^{-9} \text{ metre}.$$

Thus the electrons on an average get past about three atoms before they make a collision. This figure is not unreasonable, although it should not be taken too seriously, for if the temperature of copper is lowered from room temperature (300° K) to liquid helium temperature (4° K), its conductivity increases roughly 1,000 times. At 4° K electrons make a collision only after passing several hundred atoms.

The Hall effect

If a magnetic field (B) is applied to a metal carrying a current density J at right angles to B, so that (by convention) the electrons are moving with a velocity v in the direction of $-J$, there will be a force at right angles to both v and B given by $-ev \times B$ acting on the electrons. This will deflect the electrons until enough charge has accumulated to set up an electric field $E = -v \times B$, which cancels the effect of B. Thus, if a metal bar carries current normal to B, there will be an electromotive force between points on opposite sides of the bar. This phenomenon, known as the Hall effect, is named for a U.S. physicist, Edwin Herbert Hall, who discovered it. For conduction by electrons with a charge $-e$ the Hall coefficient (R_H) in the expression

$$E = -R_H(J \times B) \tag{41}$$

is seen to be

$$R_H = -\frac{1}{Ne}. \tag{42}$$

For the alkali metals and copper and silver the value of R_H obtained experimentally is negative and the value of N deduced from (42) corresponds to one mobile electron per atom. In other cases (*e.g.*, zinc and cadmium) the numerical agreement is reasonable but R_H has the wrong sign. This anomalous Hall effect can only be explained quantum mechanically. Metals are noticeably better conductors of heat than most nonmetals are, and it is reasonable to attribute this characteristic to the transport of heat by the mobile electrons. If there are mobile particles with an average random velocity u and a mean free path λ that contribute a term C to the thermal capacity per unit volume, the general expression for the thermal conductivity (K) is

$$K = \tfrac{1}{3}u\lambda C.$$

If this expression is compared with the one for electrical conductivity,

$$\sigma = \frac{1}{2}N\frac{e^2}{mu}\lambda,$$

it is seen that their ratio is

$$\frac{K}{\sigma} = \frac{2}{3}\frac{mu^2C}{Ne^2}.$$

If the electrons behave as a classical gas, this ratio gives

$$\frac{1}{2}mu^2 = \frac{3}{2}kT$$

and

$$C = \frac{3}{2}Nk,$$

in which k is Boltzmann's constant and T is the absolute temperature. Thus, the ratio becomes

$$\frac{K}{\sigma} = 3\frac{k^2T}{e^2}. \tag{43}$$

The ratio of the thermal and electrical conductivities should therefore be the same for all metals and proportional to the temperature. This is indeed the case, and equation (43) expresses a well-known empirical rule discovered by two German physicists, Gustav Heinrich Wiedemann and Rudolph Franz, in the 19th century. There are appreciable deviations from this rule at low temperatures, but it is reasonably well obeyed near room temperature.

If, as has been assumed, the electrons have a molar specific heat $(3/2)R$, in which R is the gas constant, this should be observable directly, as an addition to the molar specific heat $3R$ expected as a consequence of thermal energy stored in lattice vibrations. This is not observed experimentally, and most metals have molar specific heats at room temperature that are quite close to $3R$. This result, like the anomalous Hall effect in some metals, has no classical explanation and is essentially a quantum mechanical effect. Therefore, although classical theory

gives a reasonable interpretation of much of the behaviour of metals, there are some phenomena to which it gives completely wrong answers.

Quantum theory of metallic conduction. In classical mechanics, electrons are regarded as point particles of mass m and charge q equal to $-e$ pursuing definite trajectories calculable from the equations of motion and the initial position and momentum of the particle. In quantum mechanics the uncertainty principle makes it impossible to assign precise values simultaneously to the position and momentum of a particle, and only approximate trajectories can be considered. These are described in terms of wave functions ψ (the Greek letter psi), which have the property that the product $\psi^2 dV$ is the probability of finding the particle described by ψ in a volume element dV. The classical equations of motion are replaced by a differential equation for ψ, called Schrödinger's equation (after Erwin Schrödinger, an Austrian physicist, who first proposed it), which is

Schrödinger's equation

$$\frac{\hbar^2}{2m}\nabla^2\psi + (U - q\phi)\psi = 0.$$

In this equation, U is the total energy of the particle, $q\phi = -e\phi$ is its potential energy in the electrostatic potential ϕ, m its mass, and $\hbar = h/2\pi$. When this equation is solved, the solutions must fit the prescribed physical conditions of the problem; thus in a metal ψ^2, and therefore ψ, must vanish at the surface of the metal. Prescribed boundary conditions such as this lead to the result that Schrödinger's equation has solutions for only particular values of the total energy U. The possible values of U are known as the eigenvalues of the problem and the corresponding wave function as the eigenstates of the problem. The eigenvalues and eigenstates are uniquely determined by the boundary conditions and the form of the potential ϕ. These in turn are determined by the atomic structure of the metal. Each electron in the metal must be assigned to a definite eigenstate, and Pauli's exclusion principle states that not more than two electrons, with opposite spins, can be assigned to each eigenstate.

In a metal there are huge numbers of electrons (10^{20} or so in a cubic millimetre), and the trajectories of individual electrons would be of no interest even in the classical case. Only the average behaviour of all the electrons is significant. In quantum mechanics the individual states of the electrons are not of interest but only the average distribution of the electrons among the possible states. Because the number of electrons is so large, the average properties derived from this average distribution are an excellent approximation to the actual behaviour. The process of finding this average distribution and using it to derive average properties is the subject matter of statistical mechanics; and so a treatment of electronic conduction in metals is necessarily a problem in quantum statistical mechanics. Fortunately, although the details of any calculation are exceedingly complicated, the results are simple, and virtually all the information of any use is available once the average distribution of the electrons among the various states has been found.

The first step in the calculation is to find the way the states themselves are distributed in energy—that is, to find a function (G) of the energy (U)—*i.e.*, $G(U)$—such that the product $G(U)\,dU$ is the number of available states in an energy range dU. The next step is to find the probability $F(U)$ that states in the energy range dU at U are actually occupied by electrons. Once both these steps have been performed, the number of electrons dN in states of energy between U and $U + dU$ can be written as

$$dn = F(U)\,G(U)\,dU.$$

Each state that is an allowed solution of Schrödinger's equation can accommodate two electrons according to Pauli, and in describing the states obtained in this way of calculating $G(U)$ it is convenient to incorporate this factor of two in the definition of $G(U)$, so that $F(U)$ becomes the probability of a state of energy U being occupied subject to the restriction that no state can contain more than *one* electron. The derivation of the probability $F(U)$ is purely a problem in statistics and the answer is that

$$F(U) = \frac{1}{e^{(U-W)/kT} + 1}, \qquad (44)$$

in which T is the temperature, k Boltzmann's constant, and W a constant energy. Equation (44) is the Fermi–Dirac formula (named for the physicists Enrico Fermi and P.A.M. Dirac), W is the Fermi energy, and particles described by this equation are said to obey Fermi–Dirac statistics.

Figure 15A shows the probability function F against U for three values of temperature T; one curve corresponds to absolute zero ($T = 0$), the second to a low temperature $kT << W$, (in which $<<$ symbolizes much less than), and the third to a temperature $kT > W$. At absolute zero, F equals one when U is less than W, and F equals zero when U is greater than W. All states of en-

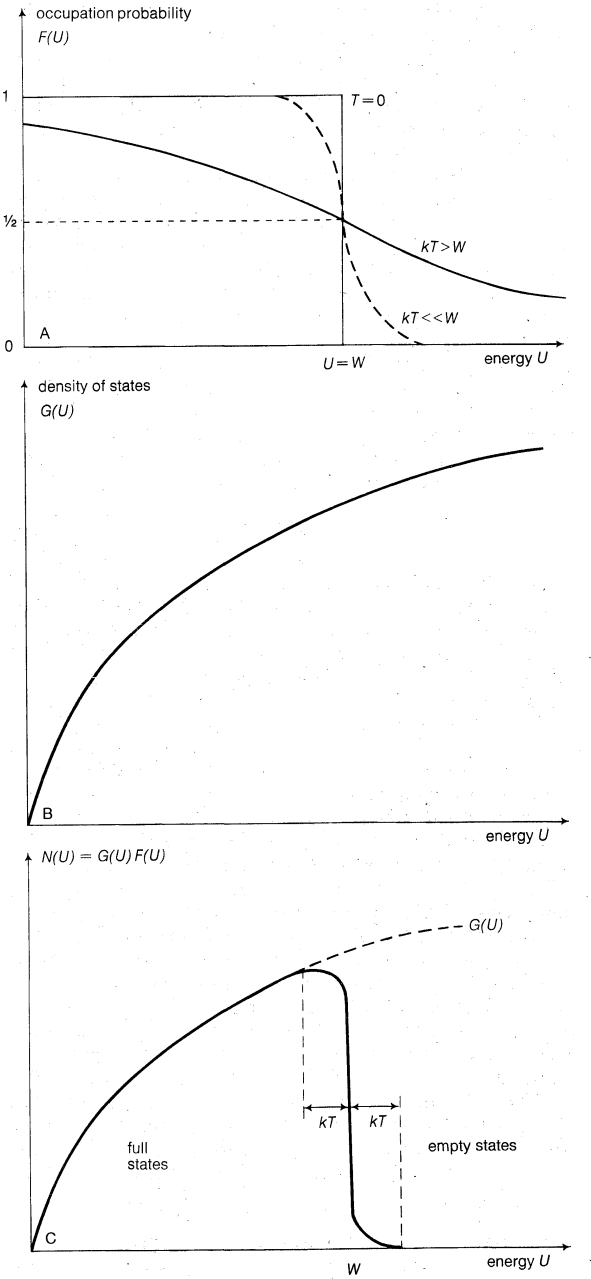


Figure 15: *Fermi–Dirac statistics.*
(A) Fermi–Dirac probability function at three different temperatures; (B) density of states available to electrons in an idealized metal; (C) electron distribution as a function of energy at a finite temperature (see text).

ergy less than W are occupied and all states of higher energy are empty. When kT is small a few states of energy just less than W will be empty and a few states of energy just greater than W will be occupied. When T is large the distribution is much more spread out, but in every case the probability that a state of energy U equals W is occupied is ½.

In metals and semiconductors, only the case in which kT is much smaller than W is important; then the Fermi level effectively divides the full low-energy states from the empty high-energy states.

The probability function F is determined by the general laws of statistical mechanics and is the same for every material; the density of states function $G(U)$, however, depends on the form of the allowed solutions of Schrödinger's equation. These in turn depend on the form of the potential ϕ, and this is a feature of the structure of the material. Before $G(U)$ can be investigated some sort of model for the material has to be adopted.

The classical theory simply assumed that $q\phi$ is constant inside a metal and increases suddenly at the surface. If this model is used in quantum theory, the resulting equations can be solved directly. The allowed wave functions are standing, or stationary, waves (*e.g.*, a vibrating rope attached at one end gives standing waves), with nodes at the surface of the metal, and the function $G(U)$ is given by

$$G(U)dU = \frac{\pi}{2}V\left(\frac{8m}{\hbar^2}\right)^{3/2}U^{1/2}dU,$$

in which V is the volume of the metal. In the treatment below, since the primary concern is with the number of electrons per unit volume, V may be taken as unity and omitted from the equations. The density of states function G for this model is plotted in Figure 15B. At absolute zero, all states up to W are filled, and the actual distribution of electrons in energy is given by the function $G(U)\,F(U)$, in which $F(U)$ is one for $U < W$ and zero for $U > W$. The total number of electrons (per unit volume) is

Distribution of electron energies

$$N = \int_{U=0}^{U=\infty} dN(U) = \int_0^\infty F(U)G(U)dU$$

$$= \int_0^W G(U)dU = \frac{\pi}{3}\left(\frac{8m}{\hbar^2}\right)^{3/2}W^{3/2}.$$

Because the value of N is fixed (there is one electron per atom), the result determines the value of the Fermi energy W. This gives

$$W = \frac{\hbar^2}{8m}\left(\frac{3N}{\pi}\right)^{2/3},$$

and using the numerical values of Planck's constant h and mass m and taking N equal to 27×10^{27} electrons per cubic metre, it leads to $W \simeq 5 \times 10^{-19}$ joule $\simeq 3$ electron volts $\simeq k \times 3 \times 10^4$ degrees. Thus, at temperatures up to the melting points of most metals, the equation yields a kT much less than W. In this case the distribution of the electrons in energy is as shown in Figure 15C. The value of W is little changed from the value appropriate to T equal to zero. Only the occupation of those states with an energy range kT about W is appreciably changed. It is now possible to see why the electrons do not contribute to the specific heat of the metal. As the temperature of the metal is increased, only a fraction, kT/W, of all the electrons change their energy states. Thus the electronic specific heat per mol, instead of being $\frac{3}{2}R$, is nearer to $\frac{3}{2}R\frac{kT}{W}$. At room temperature $kT/W \simeq 0.01$ and so this contribution is small and unimportant.

So far the wave functions, which are solutions of Schrödinger's equation, have been regarded as corresponding to standing waves in the interior of the metal, and in order to understand conduction (*i.e.*, the motion of electrons) this concept must be modified. A standing wave in quantum mechanics can never correspond to a net motion of a particle. Moving particles are described only by running waves. A typical standing wave describing an electron in a definite state in the metal is of the form $\cos \beta x \cos \omega t$, in which $\hbar^2\beta^2$ is the mean square momentum of the electron in its oscillatory motion and $\hbar\omega$ is its energy. It can, however, always be written as a combination of running waves in opposite directions; thus, for example,

$$\cos \beta x \cos \omega t = \frac{1}{2}(\cos [\omega t - \beta x] + \cos [\omega t + \beta x]),$$

which is the sum of two waves travelling in the positive and negative directions with velocities ω/β and $-\omega/\beta$. The allowed electronic states could equally well have been described in terms of pairs of running waves rather than standing waves.

In the presence of an electric field ($\boldsymbol{E}$), electrons in running-wave states that correspond to motion "downhill" in the field pick up energy from the field and, as a result, electrons occupying these states need to acquire less thermal energy from collisions with other electrons and the lattice of ionic cores. Thus these states tend to be preferentially occupied by electrons at the expense of states corresponding to "uphill" motion. If an electron spends a time τ between collisions in a state corresponding to a velocity v, the thermal energy required to get it into that state is reduced by $q\boldsymbol{E} \cdot \boldsymbol{v}\tau$. Thus, if U is the initial energy of the state in zero field, the probability that the state is occupied in a field $\boldsymbol{E}$ is no longer given by equation (44) but by the same equation with U replaced by $U - q\boldsymbol{E} \cdot \boldsymbol{v}\tau$. This makes no difference to the states with U much less than W, because these are occupied in any case; nor does it make any difference to the states with U much greater than W, which remain empty. It does, however, alter the distribution in states within the region where U approaches W and tends to increase the number of electrons in states that correspond to motion with the field. The effect can be estimated as follows: near the Fermi level, the exponential factor $(U - W - q\boldsymbol{E} \cdot \boldsymbol{v}\tau)/kT$ in the Fermi–Dirac formula, equation (44) is near unity and is, for small values of $\boldsymbol{E}$, approximately $1 - q\boldsymbol{E} \cdot \boldsymbol{v}\tau/kT$. Thus the Fermi–Dirac factor can be approximated, and a brief calculation leads to the expression

The Fermi–Dirac factor

$$J = \frac{1}{3}N\frac{q^2\tau}{m}\boldsymbol{E}.$$

A more accurate evaluation gives a different numerical factor. If this equation is compared with the classical result,

$$J = \frac{1}{2}N\frac{q^2\tau}{m}\boldsymbol{E},$$

it is seen to be virtually the same. Thus this particular quantum mechanical model leads to the same result as the classical model for the conductivity and can also be shown to yield the Wiedemann–Franz law. It has the advantage, however, that it explains the absence of any appreciable electronic specific heat. It does not, on the other hand, explain the anomalous Hall effect. To deal with this it is necessary to drop the assumption that the electrons move in a constant potential inside the metal.

In a real metal (or semiconductor) electrons move in a periodic potential caused by the ionic cores of the metal atoms. This appreciably modifies the solutions of Schrödinger's equation, and the density of states, instead of increasing steadily with energy, has the form shown in Figure 16A. There are regions of allowed energy (energy bands) and forbidden zones in which there are no solutions of Schrödinger's equation. Furthermore, a general theorem about the solutions of Schrödinger's equation ensures that total number of states per unit volume in each allowed band is always a small, even-integral multiple of the number of atoms per unit volume. If, as in diamond, the number of electrons is just sufficient to fill all the states in a band, then the Fermi level, which

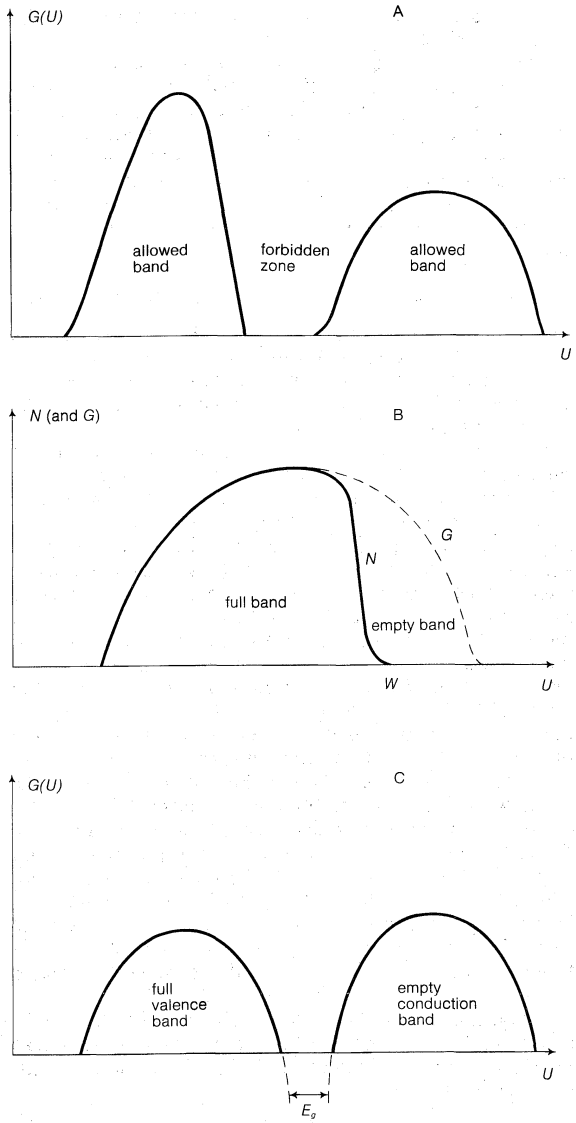


Figure 16: *Electron states in a solid.*
(A) General form of the density of states as a function of energy; (B) occupation of the available states in a partially filled (conduction) band; (C) density of states as a function of energy in a semiconductor (see text).

divides the full states from the empty states, lies in a forbidden zone. The electrical conduction results only from electrons with energies near the Fermi level, and in this case there are no electrons in this region of energy. Thus a material with a completely filled band and no electrons left over is an insulator. In the section that follows, it is seen how this mechanism is modified in semiconductors.

A metal must be a material in which one band of allowed states is only partially filled. This band is known as the conduction band, and the Fermi level lies within it. Figure 16B shows both the conduction band and the distribution of electrons in a typical metal.

In considering the motion of electrons in a constant potential, quantum mechanics does not lead to results essentially different from those of classical mechanics. The case, however, is quite different with a periodic potential. Then, according to classical mechanics, an electron situated where energy is less than the maximum potential energy cannot travel through the medium but will be trapped in one of the potential wells. In quantum mechanics, an electron with less energy than the maximum can still travel from one potential well to the next one.

Thus, the motion of electrons in real metals and semiconductors leads to mechanical behaviour not explicable in purely classical terms. On the other hand, in order to be able to visualize the response of electrons to periodic fields, it is useful when possible to employ equations that show some similarity to the equations of classical mechanics. In particular, if p is the momentum of an electron and a force F acts on the electron, it would be convenient to retain Newton's law,

Motion of electrons in periodic potentials

$$\frac{dp}{dt} = F, \tag{45}$$

and at the same time, if the electron has an equivalent classical velocity v, to be able to express the rate of change of the electron energy (U) when it responds to a force (F) as

$$\frac{dU}{dt} = Fv.$$

These equations are, of course, satisfied by free electrons, for which the electron energy is the same as its kinetic energy, $U = \frac{1}{2}mv^2$. The equations, however, are also satisfied in a more general way if the velocity is expressed as

$$v = \frac{\partial U}{\partial p}. \tag{46}$$

This gives, according to the laws of classical mechanics, the following:

$$\frac{dU}{dt} = \frac{\partial U}{\partial p}\frac{dp}{dt} = \frac{\partial U}{\partial p}F = vF.$$

For a classical particle the energy is the sum of a potential energy term (zero in this case) and a kinetic energy term $p^2/2m$, and equation (46) leads immediately to $p = mv$, which could be used to define the momentum of a classical particle. This equivalence is also true for a free quantum particle moving in a constant potential; but for a quantum particle moving in a periodic potential, if p is to obey equation (45), the energy can no longer depend on p simply through a term $p^2/2m$. Indeed, one of the important properties of solutions of the Schrödinger equation for a particle in a periodic potential is that the energy U is a periodic or oscillatory function of p. This means that for some values of p the energy actually decreases with increasing p. When this happens the velocity v is negative and the equivalent classical velocity is in the opposite direction to the momentum p. Clearly the behaviour of a particle with these properties would be strange. It is necessary to investigate this behaviour because, in many metals and semiconductors, the Fermi level falls in a region where it occurs, and, since it is just those electrons with an energy near the Fermi level that determine the electrical properties, the anomalous behaviour has important consequences.

No matter how energy depends on p, it cannot depend on the sign of p, for then the energy would be lower for one sign of p and there would be more electrons with this sign of p even in zero field. The electrons in the crystal would then have a net total momentum, which is clearly unacceptable. Thus U must be a function of p^2 to make it sign-independent. The actual dependence of U on p^2 is often quite complicated in real metals and semiconductors, but for many purposes a good approximation to U at the bottom of an energy band of allowed states is

$$U \simeq \frac{1}{2}Kp^2;$$

and near the top of a band of allowed states it is

$$U = U_0 - \frac{1}{2}K'p^2,$$

in which U_0, K, and K' are positive constants.

With v defined by equation (46) the rate of change of v is

$$\frac{dv}{dt} = \frac{d}{dt}\left(\frac{\partial U}{\partial p}\right) = \frac{\partial^2 U}{\partial p^2}\frac{dp}{dt} = F\frac{\partial^2 U}{\partial p^2},$$

and so, to present the motion in completely classical

terms using Newton's law in the form $\frac{dv}{dt} = \frac{1}{m}F$, the factor $\frac{1}{m}$ must be replaced by

$$\frac{1}{m^*} = \frac{\partial^2 U}{\partial p^2}.$$

The effective mass

This quantity m^* is known as the effective mass. For electrons near the bottom of a band, it is equal to $1/K$ and is positive, but for electrons near the top of a band it equals $-1/K'$ and is negative. If there is an electron near the top of a band with p positive and a force F (electric field) is applied that increases p, then according to equation (45) the energy of the electron decreases and work is done against the source of the force F. This is not what happens in a real metal, for it would imply that the current caused by an applied field is in the wrong direction. The reason this does not occur is that in real metals, if there are any electrons near the top of an energy band, then all the lower energy states in the band are already filled. Since there is never an occasion in which conduction is due solely to a few negative effective mass electrons near the top of a band, the only cases to be considered are that of either a few electrons in states near the bottom of a band, in which case they have positive effective masses and behave normally, or of an almost full band with a few vacant states, or holes, near the top of the band. The problem is to find a reasonable way of describing the latter situation.

In a completely filled band, for every electron with a velocity v there is another electron with a velocity $-v$, and so the total current resulting from a full band of electrons is zero. If one electron of charge $-e$ and velocity v is missing, a contribution $-ev$ to the current density is missing (considering unit volume of material). The net current density is then $+ev$. The total current density can be expressed either by summing $-ev$ over all the full states or by summing $+ev$ over all the vacant states or holes. In an electric field (E), electrons with momentum p (written here as a vector because momentum is a directed quantity) in the direction of the force $-eE$ will increase in momentum but decrease in energy. In the process, they do work against the source of the field. More thermal energy would have to be supplied by collisions to maintain electrons in states in which p is in the direction of the force. This will not happen, and at the first collision these electrons are likely to drop into states with smaller values of p or with p opposed to the force. Because v is opposed to p, this gives a net flow of electrons in the direction of the force; and a current of the normal sign (*i.e.*, J) could, however, equally be described in terms of holes. In a field E the number of holes in states corresponding to electrons with momenta parallel to the force $-eE$ increases. The electron velocity in these states obeys the law

$$\frac{dv}{dt} = \frac{F}{m^*} = \frac{1}{m^*}(-eE) = +e\left(-\frac{E}{m^*}\right),$$

and this is therefore the equation of motion for the hole. It is the equation of motion of a particle of charge $+e$ and effective mass $-m^*$, which is also positive, because effective mass m^* is negative. Because it is a hole that also contributes $+e$ of charge to the current density, the dynamics of the system can be treated entirely in terms of holes of positive charge and positive effective mass. This is so convenient, compared with treating negatively charged electrons with a negative mass and an esoteric relation between energy and momentum, that whenever electrons near the top of a band are involved, their behaviour is always considered in terms of holes.

The concept of hole

Because the energy levels that influence the electrical properties of a metal are those lying near the Fermi level, the factor that decides whether electrons or holes are appropriate is the position of the Fermi level in the band. If there are few electrons in the band and the Fermi level is near the bottom of the band, an analysis in terms of electrons is appropriate. If, on the other hand, the band is almost full and the Fermi level is near the top of the band, then the useful concept is hole conduction.

It is now possible to consider the anomalous sign of the Hall coefficient. According to equation (41), when a magnetic field B is applied perpendicular to current flow J in a conductor, an electric field E, perpendicular to both J and B, is produced and is given by

$$E = -R_H(J \times B) = \frac{1}{Ne}(J \times B).$$

In some metals (*e.g.*, sodium) this agrees with experiment, but in other cases (*e.g.*, zinc) the observed field has the opposite sign. The reason for this is that in zinc the conductors must be holes and the effective charge of the carriers is $+e$, not $-e$. This interpretation is strengthened by the fact that Hall effects of anomalous sign occur only in metals in which there are other grounds (chemical, crystallographic, etc.) for believing that the conduction band is almost full.

In a perfectly regular lattice, which produces a perfectly regular periodic potential, quantum mechanics would predict that there should be no collisions and no transitions between different electronic states or energy levels, which would lead to an infinite conductivity. If an electric field were applied to such a perfect medium, the electrons would acquire so much energy from the field that they might either ionize the atoms still further or perhaps emerge from the positive end of the medium. A real medium does not behave like this. One reason is that even single crystals of metals are not perfect. The periodicity of the lattice is disturbed by dislocations and impurity atoms. An impurity of as little as one part per million gives an impurity atom on an average once every hundred atoms in each direction. Impurities and lattice imperfections make an important contribution to electron collisions and, at low temperatures in many metals, are the dominant cause of a residual resistivity. Experimentally, however, the resistivity of most metals rises appreciably as the temperature T increases. Because the number of conduction electrons (if it changes at all) increases with T, this means that collisions become more frequent, the reason being that, as T increases, the vibration of the ionic cores about their mean lattice positions becomes more violent. Also, as T is raised, the lattice becomes less and less a perfect periodic structure until, at the melting point, it disintegrates entirely.

The atoms in a crystal are strongly bound together; as a result, the vibrations of the atoms are tightly coupled and one atom does not vibrate alone without influencing its neighbours. Thus the fundamental modes of vibration are much more like waves propagating through the lattice. These waves have the same nature as sound waves, although they are usually of much higher frequency, and are known as phonons, by analogy with photons or light waves. Just as a light wave or photon can be scattered by electrons, electrons in a metal can be scattered by phonons. In a normal metal, individual electrons are scattered by phonons and impurities, and this scattering is what gives them a finite mean free path. Under some circumstances, however, the interaction between electrons and the lattice waves, or phonons, has a quite different effect and leads to an interaction between the electrons themselves. This action couples together electrons with energies near the Fermi level; at temperatures near absolute zero, its effects are sufficient to cause these electrons to move as a coherent group through the medium. Thus, once a current has been set up, the lattice vibrations, or phonons, instead of being able to scatter individual electrons, must produce a change in the motion of the whole of this coherent group of coupled electrons. Neither the relatively weak, thermally excited phonons nor impurities are able to disturb the motion of such a large number of coupled electrons and so, once this type of interaction has taken hold of the electrons, they can move through the lattice without scattering. A current once set up in the material persists indefinitely, and the metal becomes what is called a superconductor of zero resistance. The effect normally sets in abruptly as the special metal is lowered through a critical temperature

Phonons

T_c, the superconducting transition temperature. Below T_c the metal has zero resistance, unless the current is made too large, in which case the coupling between electrons is strained to a breaking point and the coherent superconducting state is destroyed. As far as is known experimentally and as far as the detailed theory can be interpreted, the resistance in the superconducting state is not only very small but is identically zero.

A magnetic field in a metal appreciably raises the energy of electrons in the superconducting state and, as a result, if it is strong enough, can make the state unstable and cause the metal to revert to a normal conductivity. If, however, the field is not that strong a much more interesting effect—the Meissner effect (proposed in 1933 by the German physicists, W. Meissner and R. Ochsenfeld)—occurs. The electrons now set up coherent circulating currents, the net effect of which is just sufficient to cancel the field in the metal. Thus, when a metal is cooled through its superconducting transition temperature in a magnetic field, the field is expelled from the metal as it passes through the transition temperature. For further details see SUPERCONDUCTIVITY.

CONDUCTION IN SEMICONDUCTORS

Semiconductors are materials that have a low concentration of mobile charge carriers (electrons or holes) and a correspondingly low conductivity. They display many properties that can be put to practical use in electronic devices such as the transistor. The elements carbon (C), silicon (Si), germanium (Ge), and the gray allotrope of tin (Sn), together with a large number of compounds, have two things in common: they all contain an average of four valence electrons per atom, and they all crystallize in the diamond structure in which each atom is at the centre of a regular tetrahedron of neighbouring atoms. All these materials have a band of electronic levels, the valence band, which can just accommodate four electrons per atom and a second band of levels, the conduction band, the lowest level of which is separated from the highest valence-band level by a forbidden zone with no allowed levels. The width of this zone (E') represents the energy gap. It varies in width from a fraction of an electron volt in tin to several volts in diamond and cuprous chloride. The density of states available to electrons in these materials is shown in Figure 16C. According to the theory explained in the last section, all these materials, which contain one full band and one empty band, ought to be insulators. They have, however, one peculiarity in that the energy gap (E') is not large, which has a profound influence on their properties.

The conduction band

According to the general principles of statistical mechanics, the probability that an electron will pick up enough thermal energy to be excited from a state at the top of the valence band is very nearly $e^{-E'/kT}$. The energy represented by kT at room temperature is 1/40 electron volt and so this probability, with E' expressed in electron volts, is $e^{-40E'}$. There are about 10^{28} electrons per cubic metre in the valence band; as a result the probable number of electrons in the conduction band is of the general order of $10^{28}\ e^{-40E'}$ per cubic metre. This factor can also be expressed as $10^{(28-18E')}$. If, as in silicon, E' equals 1.1 electron volts, this gives about 10^8 conduction electrons per cubic metre of silicon (a better estimate is given below); but if, as in diamond, E' equals 3.5 eV, the number is virtually zero. Thus quite small changes in the energy gap (E') lead to a large difference in the number of electrons. Diamond may thus be considered to be an insulator, but silicon, with a few electrons in the otherwise empty conduction band and a few holes in the otherwise full valence band, is a poor conductor or semiconductor. Gray tin with a smaller energy gap is almost metallic.

The various elements mentioned above, as well as others of the same general type, are all regarded as semiconductors even though, as in the case of diamond, the semiconducting properties can occur only at high temperatures. A few elements of different structure—*e.g.*, tellurium—are also semiconductors; but, of them, only silicon is considered, partly because it is the most familiar and technologically important material and partly because it illustrates all the main features of semiconductor behaviour.

In silicon there are a few electrons in the conduction band and a few electrons missing from the valence band. The probability of a state in the conduction band being occupied is small and the probability of one being vacant in the valence band is also small. It follows that the Fermi level (W) must lie somewhere between the two bands—*i.e.*, in the energy gap. In fact, the Fermi level lies more or less in the middle of the forbidden energy gap, and the intrinsic hole and electron concentrations N_c and H_v are equal and given by

$$N_c = H_v = Ke^{-E'/2kT},$$

in which K is a constant. Because both holes and electrons contribute to the conductivity, the conductivity can be expressed as $\sigma = eN_c\mu_e + eH_v\mu_H$, in which μ_e and μ_H are the mobilities of electrons at the bottom of the conduction band and holes at the top of the valence band. (Here μ_e is taken as positive, although many textbooks take $-e\mu_e$ to be positive instead.) It follows then that the conductivity is

$$\sigma = Ae^{-E'/2kT}.$$

With E' equal to 1.1 electron volts, the conductivity changes by a factor of 10 as T changes from 300° K (27° C) to 340° K. This rapid exponential rise of conductivity with temperature is characteristic of semiconductors. Silicon, in common with most semiconductors, has two other important features. First, the mean free path and mobility of the charge carriers are high compared with those of metals, and, second, by introducing impurities, all the holes can be eliminated to produce many more electrons in the conduction band, or vice versa. The significance of the long mean free paths and high mobility is that silicon is a better conductor than might be expected with so few carriers and also that carriers in silicon can reach quite high velocities at relatively small current densities.

Silicon belongs to Group IV of the periodic table and has four valence electrons. Elements from neighbouring groups—*e.g.*, boron with three valence electrons and phosphorus with five valence electrons—can be introduced in small quantities as impurities and fit into the crystal lattice without appreciably disturbing the structure of the electronic energy levels. If boron is introduced (called doping), however, there will be one electron too few per boron atom to fill the levels of the valence band, and if phosphorus is introduced, there will be one too many. Thus, in boron-doped silicon, there are additional holes in the valence band, and in phosphorus-doped silicon there are additional electrons in the conduction band.

The number of electrons and holes in pure silicon at room temperature is about 10^{16} per cubic metre, and because there are some 10^{28} silicon atoms per cubic metre, the addition of one part per million of phosphorus—*i.e.*, 10^{22} atoms per cubic metre—increases the number of electrons available for the conduction band by a factor of $10^{(22-16)} = 10^6$. A few of these electrons are used to fill the holes in the valence band, but the remainder are available for conduction. Boron has the opposite effect. It "soaks" up all the conduction-band electrons and greatly increases the number of holes. Thus, whereas in pure, or intrinsic, silicon the conduction process involves equal numbers of holes and electrons, in phosphorus-doped silicon the process involves only electrons and in boron-doped silicon only holes. Silicon doped with a donor element from Group V, such as phosphorus, is known as *n*-type silicon because the conductivity is caused by negative carriers. Silicon doped with an acceptor element from Group III, such as boron, is known as *p*-type silicon because the conductivity results from positive carriers (holes). Adding electrons to a system always raises the Fermi level, and removing them lowers it. As a result, in *n*-type silicon the Fermi level lies near the bottom of the conduction band, and in *p*-type silicon it lies near the top of the valence band. If a sample of

n-type and *p*-type silicon

n-type silicon is placed in contact with a sample of p-type silicon some of the excess electrons in the n-type material flow into the p-type material and fill a few of the holes. This charges the p-material negatively and raises the potential energy of all the electrons in the material. According to statistical mechanics the process continues until the Fermi levels in the two materials coincide. This can be illustrated as shown in Figure 17. There is now a

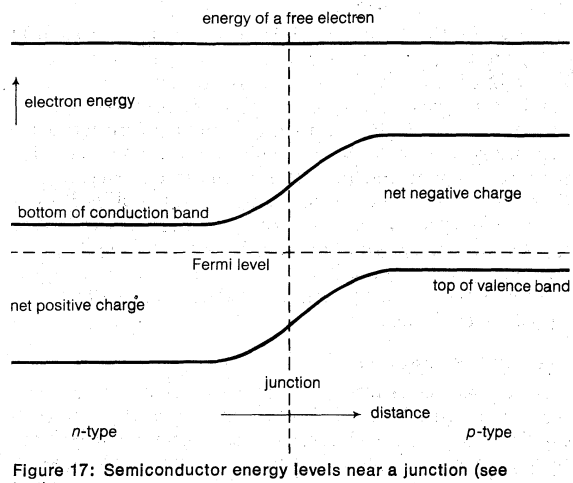


Figure 17: Semiconductor energy levels near a junction (see text).

contact potential difference between the two regions. If the Fermi levels actually coincided with the band edges, the contact potential ϕ_c would be given by $e\phi_c = E'$; *i.e.*, $\phi_c \simeq 1.1$ volts, with the p-material negative. This process is the origin of all forms of contact potential.

There are many interesting phenomena that occur at a p-junction, details of which may be found in SEMICONDUCTOR DEVICES. Electrical conduction in semiconductors differs in many ways from conduction in metals. For example, it will suffice to consider n-type silicon with 10^{22} conduction electrons per unit volume compared with 10^{29} in a metal. It follows immediately that, for a given current density, the velocity of the carriers must be much greater. Because the mobility of carriers in silicon (about 0.1 metre per second per volt metre^{-1} in SI units) is some 100 times larger than in metals, quite high current densities can be achieved in silicon. If the current density ($J = nev$) is 10^6 amperes per square metre, there are n equal to 10^{22} electrons per cubic metre, and because e is equal to 1.6×10^{-19} coulomb, the velocity will be 10^3 metres per second. This is comparable with the velocity of the lattice waves, or phonons, that scatter the electrons. Detailed analysis shows that the scattering becomes more effective as v approaches this velocity. In high fields, J is no longer proportional to E and there is a breakdown in Ohm's law. The limiting drift velocity is of the order of 5×10^4 metres per second. At still higher fields the drift velocity increases again and eventually electrons acquire enough energy from the field to create new pairs of charge carriers—*i.e.*, electrons and holes—in collisions. The fields in many semiconductor devices are strong enough to make all these nonlinear phenomena significant.

CONTACT POTENTIAL AND THERMOELECTRICITY

The atomic structure of a metal or semiconductor determines the position of its allowed energy levels relative to the energy of an electron at rest outside the metal. The number of electrons in the metal then determines, at any particular temperature (T), the energy of the Fermi level (W). The position of the Fermi level, relative to the standard energy zero, is a characteristic property of each metal and varies appreciably from metal to metal or metal to semiconductor. When two metals, a metal and a semiconductor, or two semiconductors are placed in electrical contact, electrons flow out of the material with the higher value of W. This continues until the resulting change in electrostatic potential brings the Fermi levels into coincidence (see Figure 17). If W_1 and W_2 are the two Fermi-level energies relative to an external electron at rest, the contact potential developed (ϕ_{12}) is given by

$$e\phi_{12} = W_1 - W_2.$$

In a chain of different metals (1, 2, 3 $\cdots$ n), the potential difference between (1) and (n) is $\phi_{12} + \phi_{23} + \cdots \phi_{n-1,n}$ or is equal to ϕ_{1n} because $e\phi_{1n} = (W_1 - W_2) + (W_2 - W_3) + \cdots (W_{n-1} - W_n)$. In a closed loop, therefore, there will be no net electromotive force and no current. If, however, in a closed loop of two metals, A and B, the two contacts are maintained at different temperatures T_1 and T_2, then, if the positions of the Fermi levels of A and B vary in a different way with temperature, the contact potential $\phi_{AB}^{(1)}$ at T_1 will no longer be equal to $\phi_{AB}^{(2)}$ at T_2. There is now a net emf acting around the loop and a current will flow. The current could be used to do mechanical work, and the laws of thermodynamics then require that heat should be taken in at the hot junction and given out at the cold junction. The production of a thermal emf is known as the Seebeck effect (after the Estonian-born German physicist Thomas Johann Seebeck), and the absorption and rejection of heat at the junctions when a current flows is known as the Peltier effect (after French physicist Jean-Charles Peltier).

The Seebeck effect

There is a third thermoelectric effect, the Thomson effect (discovered by William Thomson, later known as Lord Kelvin), which occurs in the interior of a single, homogeneous piece of metal in which there is a temperature gradient. This effect has two aspects. In an isolated piece of metal the presence of a temperature gradient means that the distribution of the electrons among the available energy levels will vary slightly from point to point in the metal. As a result there will be a redistribution of the negatively charged electrons in the metal, and this will cause a potential gradient to appear. If, on the other hand, a current flows in the metal from the cold end to the hot end so that electrons travel from the hot to the cold end, electrons arriving at any point in the metal will have thermal energy to spare. There will, therefore, be a local generation of heat proportional to both the temperature gradient and the electron flux. It is usual to express the heat generated per unit volume in terms of the conventional current density J and the temperature gradient as $-\sigma J dT/dx$. When J and dT/dx are positive, current flows from cold to hot and electrons from hot to cold. The Thomson coefficient σ is therefore expected to be negative. In those metals in which the conduction is caused by the positive holes (and also in p-type semiconductors) the coefficient will be positive. These expectations are not fulfilled in practice; thus, although sodium has a negative coefficient, as expected, lithium, a similar metal, has a positive coefficient. The reason for this is described next.

The only electrons that take part in the conduction process are those with energies within a few multiples of kT (the thermal energy) of the Fermi level, and further, if the mean free path of the electrons varies appreciably with energy, the electrons within this group that have the largest free paths will preponderate in the current. In general, the mean free paths near the Fermi level vary with energy in a way that depends critically on the atomic structure of the metal and it is quite possible for the mean free path to decrease rapidly with energy, in which case most of the electrons constituting the current will have thermal energies below, rather than above, the Fermi level. If the temperature is raised, electrons farther from the Fermi level in both directions can take part in conduction, but now the main increase will come from electrons considerably below the Fermi level. Thus, as the temperature is increased, the main thermal energy of those electrons that actually constitute the current will decrease, rather than increase, and this will reverse the sign of the Thomson coefficient σ.

Figure 18 shows two conductors labelled (1) and (2) and the various thermoelectric effects that occur. The Peltier coefficient $\Pi_{12}(T)$ is defined so that the heat ab-

The Peltier coefficient

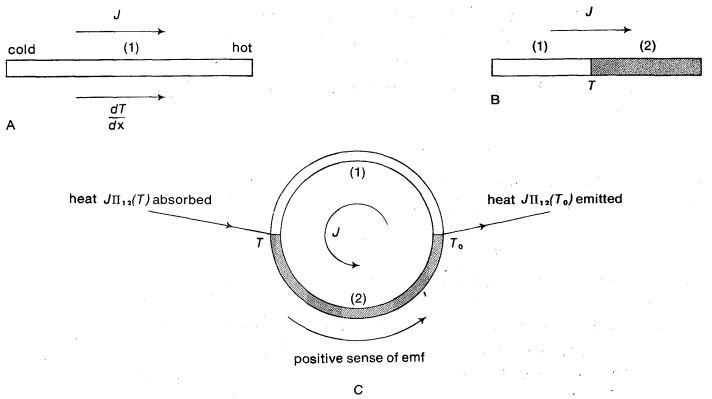


Figure 18: *The thermoelectric effects.*
(A) Heat generated at a rate $-\sigma_1 J\, dT/dx$ per unit volume; (B) heat absorbed at the junction at a rate $J\Pi_{12}(T)$ per unit area of cross section. (C) Seebeck electromotive force in a complete circuit (see text).

sorbed per unit area of cross section at the junction between (1) and (2)—when current I of density J flows from (1) to (2)—is $J\Pi_{12}(T)$. The Seebeck electromotive force acting round the complete circuit is V with the sense shown in the figure. The Seebeck coefficient S_{12}, or thermoelectric power of the junction, is defined with the result that

$$S_{12} = \underset{T \to T_0}{limit} \frac{V}{T - T_0} = \frac{dV}{dT}.$$

Thermodynamics yields two relations,

$$S_{12} = \Pi_{12}/T$$

and

$$\sigma_1 - \sigma_2 = -T\, dS_{12}/dT = -T\, d^2V/dT^2,$$

between the thermoelectric parameters.

The Thomson coefficients σ_1 and σ_2 for the two metals occur only as the difference $\sigma_1 - \sigma_2$, and the Peltier coefficient can be expressed as a difference between coefficients for each metal $\Pi_{12}(T) = \Pi_1(T) - \Pi_2(T)$ and the thermoelectric power as

$$S_{12} = S_1 - S_2.$$

Thus each metal can be characterized by three parameters S, σ, and Π, which are related by the Thomson equations, $\sigma = -T\, dS/dT$ and $\Pi = TS$. The third law of thermodynamics requires all thermoelectric effects to vanish at absolute zero, and so integration of the expression for dS/dT yields

$$\frac{\pi}{T} = S = -\int_0^T \frac{\sigma dT}{T}.$$

Experimentally, the Thomson coefficient σ is found to be zero in a superconductor; therefore, the absolute thermoelectric power of any metal can be measured by incorporating it in a loop with a superconductor. This is rather inconvenient, and, in practice, thermoelectric powers are usually expressed relative to the element lead which is known to have a low value.

From an experimental point of view the easiest thermoelectric parameter to measure is the electromotive force in a circuit containing two metals at different temperatures—*i.e.*, the thermoelectric power of the couple S_{12}. If S_{12} is known over a range of temperatures for a range of different pairs of metals, the other thermoelectric parameters can be obtained using the Thomson relations.

In practice, thermoelectric effects are significant in a number of ways. First of all, the possible presence of unwanted thermal electromotive forces in direct current circuits is a source of error in dc measurements. It is a prime cause of drift in dc bridges, for example. Second, thermal emf's can be used to measure temperature differences. Third, the Peltier effect may be used to produce refrigeration.

A typical pair of metals used in thermometry is the copper–constantan (alloy of 60 percent copper, 40 percent nickel) couple. Near room temperature the thermal emf is about 40 microvolts per degree temperature difference. This is too small to be easily measured and so a thermopile is usually used—*i.e.*, a circuit consisting of a set of alternating copper and constantan wires with alternate junctions at the reference temperature T_0 and the unknown temperature T. With 25 junctions at each temperature, the emf's add in series to give one millivolt per degree.

In metals, the Peltier cooling at a junction is usually completely masked by the irreversible joule heating in the bulk of the metal and is of no practical use. Semiconductors, however, have Fermi levels the exact energy of which is much more critically dependent on temperature than in metals; as a result, metal–semiconductor junctions exhibit much larger thermoelectric effects. The thermoelectric power is typically a few millivolts per degree compared with microvolts for metals. The Peltier coefficient is correspondingly larger and as a result practical refrigerating junctions can be constructed. The low thermal conductivity of semiconductors also helps to insulate the cold junctions.

ELECTRON EMISSIONS

Thermionic emission. A metal contains mobile electrons in a partially filled band of energy levels—the conduction band. These electrons, though mobile within the metal, are rather tightly bound to the metal. The energy required to release a mobile electron from the metal varies from about 1.5 electron volts to about six electron volts, depending on the metal. The electrons can acquire enough energy to escape in a number of ways. A few electrons, for example, can pick up enough thermal energy in collisions. This leads to thermionic emission. The number of emitted electrons, and therefore the thermionic emission current, depends critically on the temperature. If energetic electrons are injected into the metal from outside, some of their excess energy can be transferred directly to other electrons which may then escape from the metal. This process is known as secondary emission. In some cases one extra energetic electron incident on the metal may lead to three or four secondary electrons. Finally, if light of high enough frequency falls on the metal so that the photon energy $h\nu$ (Planck's constant h times frequency ν) exceeds the binding energy of the electrons, electrons are released from the surface, resulting in a photoelectric emission current. These are the three primary effects that will be of concern below, in addition to photoelectric effects in semiconductors. Light incident on a semiconductor leads normally not to the release of electrons but rather to the production of additional charge carriers within the semiconductor. This internal photoelectric effect has a number of interesting consequences. In a homogeneous semiconductor it leads to an increase in electrical conductivity, the photoconductive effect; but if the semiconductor contains a rectifying p–n-junction (permitting current to flow in only one direction), it can also generate an electromotive force. This is known as the photovoltaic effect. In a metal the conduction band levels are filled up to the Fermi level, which lies at an energy $-e\phi_w$ relative to a free electron outside the metal. The typical energy required to release an electron is $e\phi_w$ and this is known as the work function. At a temperature T, the number of electrons with enough energy to escape is proportional to $\exp(-e\phi_w/kT)$, in which k is Boltzmann's constant. If, for example, ϕ is two volts and T equals 1,000° K, this proportionality factor is about 10^{-10}, and thus only this small fraction of the mobile electrons will have sufficient energy to escape. Those electrons that can escape are moving so fast in the metal and have such high kinetic energies that they are little affected by the periodic potential caused by atoms of the metallic lattice. In other words, they behave like electrons trapped in a region of constant potential. Because of this, when the rate at which electrons escape from the metal is calculated, the detailed structure of the metal has little influence on the final result, and so a universal formula is valid for all

Secondary emission

metals. This result is known as Richardson's law (first proposed by English physicist Owen Williams Richardson). It is usually expressed in terms of the emission current (J) for unit area as

$$J = AT^2\exp(-e\phi_w/kT),$$

in which the universal constant A is 120 ampere degree^{-2} centimetre^{-2}, or 1.2×10^6 ampere degree^{-2} metre^{-2}, and T is in degrees Kelvin. If, for example, T equals 1,000° K and ϕ_w equals 2 volts the result is a J equal to 1.2×10^{-2} ampere per square centimetre. The rapid rise with temperature should, however, be noted. If T is increased to 1,200° K, J rises to ½ ampere per square centimetre. In most practical vacuum tubes, the cathode is coated with a mixture of barium and strontium oxides, and at 1,000° K this not only has a work function somewhat less than two electron volts but also acts as a reasonably good conductor. Typically, currents of several amperes per square centimetre can be drawn from oxide cathodes but, in practice, the current density is usually less than 0.2 ampere per square centimetre because higher densities lead to rapid degradation of the oxide layer. Tungsten has the potential ϕ_w equal to about 4.5 volts and does not emit appreciably below about 2,000° K. According to Richardson's law, at 2,800° K the emission should be about 10 amperes per square centimetre, which is roughly in agreement with experiment.

Secondary emission. If energetic electrons—*i.e.*, electrons of 10 to 1,000 electron-volt energies—strike a metal surface in vacuum, their energy is lost in collisions in a region near the surface and mainly transferred to other electrons in the metal. Because this occurs near the surface, some of these electrons may be ejected from the metal and form a secondary emission current. The ratio of the number of secondary electrons to the number of incident electrons is known as the secondary-emission coefficient. For low-incident energies (below about one electron volt) the primary electrons tend to be reflected, and the secondary emission coefficient is near unity. With increasing primary energy, the coefficient at first falls and then at about 10 electron volts begins to rise again, usually reaching a peak of from two to four at energies of a few hundred electron volts. At higher energies the primary electrons penetrate so far below the surface before losing energy that the excited electrons have little chance of reaching the surface and escaping. Thus at higher energies the secondary emission coefficients fall, and, when the electrons have a few thousand volts primary energy, are usually well below unity. If the incident primary particles are massive ions rather than electrons, they are stopped much nearer the surface. The increase in the secondary emission coefficient then persists to rather higher incident energies, and the peak value is somewhat larger. Secondary emission is not confined to metals but can also occur in insulators. Because many insulators have rather high secondary emission coefficients, it is often useful to coat a metal electrode with a thin layer of an insulator a few atoms thick when high secondary emission yields are required.

Secondary-emission coefficient

The principal use of secondary emission is in the multiplier tube, generally in association with a photo-emitting cathode. Light falling on the cathode generates photoelectrons, which are then accelerated to energies of a few hundred electron volts. The electrons strike a secondary emitting cathode and produce perhaps three secondaries per electron. These secondaries are then accelerated and strike a second electrode producing nine electrons and so on. With 10 stages of secondary emission the current gain is about 10^5, and a single photoelectron produces a current pulse large enough to operate electronic equipment. The main advantage of the secondary emission process as a gain mechanism is that it is largely free from noise.

Photoelectric effect. If light, the frequency of which is such that the photon energy $h\nu$ exceeds the work function $e\phi_w$, falls on a metal surface, some of the incident photons will transfer their energy to electrons, which will then be ejected from the metal. This is the primary photoelectric effect. If $h\nu$ is greater than $e\phi_w$, the excess energy $h\nu - e\phi_w$ transferred to the electrons will be observed as their kinetic energy outside the metal. The relation between electron energy E and ν ($E = h\nu - e\phi_w$) is known as the Einstein relation (after Albert Einstein, who first proposed it), and its experimental verification helped to establish the validity of quantum theory. It should be noted that the energy of the electrons depends on the frequency of the light, not its intensity. The intensity determines the rate of photoelectric emission. With quantum theory these relations seem almost obvious but are completely inexplicable in classical terms.

In practical devices based on photoelectric emission that are used to detect light and produce electrical signals from light signals, the photocathode is usually combined with a secondary-emission multiplier. This gives a large amount of low-noise amplification before the signal is processed by the normal techniques of electronic amplification. The internal photoelectric effect in semiconductors can now be explained. In a semiconductor the lower valence band of energy levels is almost completely full and the upper conduction band almost empty. The conductivity of the material derives from the few holes present in the valence band and the few electrons in the conduction band. Electrons can be excited from the valence to the conduction band by light for which $h\nu$ exceeds the energy gap E_g between the bands. The value of E_g depends on the particular semiconductor; for lead sulfide the threshold frequency occurs in the infrared, whereas for zinc oxide it is in the ultraviolet. For silicon, E_g equals 1.1 electron volts, and the threshold wavelength is about 9000 angstroms—*i.e.*, just in the infrared. Visible radiation produces electron transitions with almost unity quantum efficiency in silicon. Each transition gives a hole–electron pair—*i.e.*, two carriers—which contributes to the electrical conductivity. If photons arrive at a rate n per second in a volume V, and the excess carriers have a lifetime τ before they recombine, the net increase in the carrier density is $2n\tau/V$. A sample of pure silicon may be considered, in which there are some 3×10^{10} carriers per cubic centimetre at room temperature in the form of a thin plate one square centimetre and 0.03 centimetre thick, which is thick enough to absorb all incident light. If the incident radiation is 5×10^{15} quanta per second (about one milliwatt) and τ is approximately 2×10^{-4} second, the resultant density is 6×10^{13} carriers per cubic centimetre, with the result that the resistance of the plate will be decreased by a factor of about 2,000. In practice, photoconductive effects are not usually as noteworthy as this, but this example indicates that appreciable changes in conductivity can occur even with much less intense illumination. Photoconductive devices are simple to construct and are used to detect visible, infrared, and ultraviolet radiation. They are not, however, as sensitive either as vacuum photomultiplier tubes or as the devices that are described next.

The internal photoelectric effect

If a junction is formed between n-silicon containing donor impurities giving an excess of electrons and p-silicon containing acceptor impurities and an excess of holes, a contact potential is generated with the p-material negative. This is shown in Figure 17. The system is in dynamic equilibrium. The potential barrier determined by the contact potential inhibits the flow of electrons (the majority carriers in the n-material) from the n- to the p-material. Any electrons generated as minority carriers in the p-material are, however, free to cross the junction. In equilibrium the majority and minority electron currents just balance. Similarly, the majority carrier-hole current from p to n is just balanced by a minority carrier-hole current from n to p. If reverse bias is applied to the junction, making the p-material still more negative, the majority carrier currents are completely cut off, leaving only the minority carrier currents. These, in a typical diode, may amount to only 10^{-9} ampere or so. When light falls on the semiconductor, electron–hole pairs are generated in both the p- and n-material, and any holes generated in the n-material or electrons generated in the p-material are swept across the junction and yield an additional current. Each photon absorbed transfers one electronic charge across the junction. The device therefore acts as a photo-

detector, yielding one electronic charge in the external circuit for each photon absorbed. As a device for measuring weak photon fluxes the reverse-biassed photodiode is rather impractical largely because, although it has nearly unity quantum efficiency, it has no internal current gain such as occurs in a photomultiplier tube. As a result, fluctuations in the succeeding electronic amplifier (and also fluctuations in the dark current) mask weak input signals. On the other hand, when it is used to measure small changes in a large photon flux, it can be shown to achieve almost the ultimate theoretical sensitivity.

If no bias is applied to the diode, the holes generated in the *n*-material and the electrons generated in the *p*-material flow across the junction, driven by the contact potential difference. The resulting charge transfer reduces the contact potential, and the process goes on until the increased majority carrier flow, allowed by the reduced contact potential, just balances the additional photoelectric minority carrier current. Just as in thermoelectricity, a localized change in contact potential is available as an electromotive force in an external circuit connected to the diode. The diode now behaves as a photovoltaic detector of radiation. At low light intensities the output voltage is proportional to light intensity, but at high intensities the voltage saturates at a value not far off the initial value of the contact potential (but of the opposite sign). The light intensity then determines the current available from the diode. Although the photovoltaic diode can be used to detect and measure incident radiation, it is not as sensitive as the reverse-biassed diode. Its most important application is in the solar cell used (especially in artificial satellites) to generate electric power from solar radiation. In favourable circumstances a solar cell can give a 50 percent conversion efficiency, though a more usual figure is somewhat less than 10 percent. For solar radiation a typical power output is from five to 10 milliwatts per square centimetre of junction area. It is perhaps worth noting that this corresponds to several hundred kilowatts per acre.

The photovoltaic diode

Although the main use of all types of photoelectric effect is in devices that either detect light or convert radiant energy to useful electrical energy, the photoelectric effect is also useful as a research tool. The study of the energy distribution of photoelectrons released from either solids or single molecules in a gas is also a powerful technique for investigating the energy levels of the system (see SEMICONDUCTOR DEVICES).

V. Alternating electric currents

Attention has been paid so far only to steady electric currents, steady voltages, and fields and static distributions of charge. Most of the important practical applications of electricity are, however, concerned with currents, voltages, and fields that vary rapidly in time. For example, the domestic electricity supply in the United States reverses the sign of the voltage 120 times per second (called 60 cycle ac); but in electronic apparatus, voltages and currents are dealt with that change appreciably in times sometimes as short as 10^{-9} second; and in optics, the electromagnetic fields vary periodically in time with frequencies of around 10^{14} hertz (cycles per second). This section includes consideration of the behaviour of time-dependent currents in circuits, and especially one particular class of time-dependent current, the periodic current having a wave form that can be expressed as either the cosine or the sine of a phase angle that increases linearly with time. In other words, these are currents the time dependence of which can be expressed as either $I(t) = I_0 \cos(\omega t + \theta)$ or $I_0 \sin(\omega t + \chi)$, in which t is time, I_0 is the maximum current attained during a cycle, and ω, θ and χ are constants. Currents that can be expressed in this form are referred to as sinusoidal. Since $\cos(\omega t + \theta + 2\pi) = \cos(\omega t + \theta)$ and $\sin(\omega t + \chi + 2\pi) = \sin(\omega t + \chi)$, sinusoidal currents repeat at intervals given by $t = 2\pi/\omega$. The time length of one cycle is $2\pi/\omega$, and therefore the frequency (f) of repetition is $f = \omega/2\pi$, or $\omega = 2\pi f$, in which ω is called the angular frequency. Here, f and ω will be indiscriminately referred to as the frequency, the factor 2π being ignored.

If a coil (of area A and N turns) rotates about an axis normal to a steady magnetic field $\boldsymbol{B}$ at an (angular) frequency ω and the plane of the coil is perpendicular to the direction of $\boldsymbol{B}$ at zero time, the flux linking the coil is $ANB \cos \omega t$ and the induced electromotive force in the coil, written $V(t)$ for the emf at time t, is $V(t) = -\omega ANB \sin \omega t$. Thus, a system of this kind acts as a source of an alternating emf. Other sources of alternating emf can be constructed using electronic devices, gas discharges, microphones, or mechanical interruption in a direct current (dc) circuit. Some of the properties of alternating current (ac) sources will be treated in a later section but here it is necessary to know only that they exist. Alternating current circuits containing only resistive components—*i.e.*, those that obey the law that the voltage $V(t)$ at time t is equal to the resistance R times the current $I(t)$ at the same instant, $V(t) = RI(t)$—are no different from dc circuits. If, however, a circuit contains a capacitance (C), the voltage across the capacitor is related to the charge (q) by the constant C, or $q(t) = CV(t)$, and the charging current in the leads to the capacitor is $I(t) = dq(t)/dt = CdV(t)/dt$. Thus, with a time-dependent emf the behaviour of circuits containing capacitors begins to show new features. The peculiar simplicity of sinusoidal (also said of a variation with cosine) currents and voltages arises from the fact that the variation of a sine or cosine function with time is another sine or cosine function; *i.e.*,

$$\frac{d}{dt}(\sin \omega t) = \omega \cos \omega t = \omega \sin(\omega t + \pi/2)$$

$$\frac{d}{dt}(\cos \omega t) = -\omega \sin \omega t = \omega \cos(\omega t + \pi/2).$$

Thus, because of the relation $I = CdV(t)/dt$,

$$I = C\omega V_0 \sin(\omega t + \pi/2),$$

that is, if the voltage $V(t)$ applied to a capacitor is sinusoidal, the current across it is also sinusoidal with the same frequency.

A resistance is an element that dissipates electrical energy as heat (although resistance can also be used to represent the circuit properties of a device such as a motor, which converts electrical energy to mechanical work). A capacitor is an element that stores electrical energy and can then release the energy at a later time. The energy in a capacitor can be regarded as stored in the electric field between the plates. Magnetic fields can also store energy, and the question then arises as to how this will influence the flow of currents, because a current is always accompanied by a magnetic field. This is dealt with in the next section, and, although it is dealt with at more length in the article MAGNETISM, it is interesting to see what can be deduced from energy considerations alone. The magnetic field produced by a current (I) in a conductor is proportional to I and, by analogy with the electric case, the magnetic stored energy could be expected to be proportional to the square of the magnetic field and thus to the square of I. If the energy (U) is written $U = \frac{1}{2}LI^2$, L being a proportionality constant, and I increases slowly at a rate $\dot{I}$ (a superdot means the rate of change of a quantity with time), then $dU/dt = LI\dot{I}$; but the rate at which the source of the emf does work (during the increase of I) is VI (*i.e.*, $dU/dt = VI$), and thus $V = L\dot{I}$; and so the circuit produces a back emf $-V = -L\dot{I}$. The coefficient L is known as the inductance (or self-inductance) of the circuit. From the equation $V = L\dot{I}$, it is seen that if a voltage $V(t)$ is applied to an inductance, the current also will be sinusoidal with the same frequency. The three passive elements, resistance (R), capacitance (C), and self-inductance (L), together with a new concept, mutual inductance (M), which will be introduced in the following section, form the background of any treatment of ac circuits.

Self-inductance

FARADAY'S LAW OF ELECTROMAGNETIC INDUCTION

When a conductor moves in a magnetic field $\boldsymbol{B}$ with a velocity $\boldsymbol{v}$, the magnetic force $q\boldsymbol{v} \times \boldsymbol{B}$ acting on the charge

carriers q leads to a separation of positive and negative charge in the conductor and thus to an electromotive force. Faraday discovered that an emf also existed in a closed circuit when, instead of moving the circuit, he either moved or changed the sources of the magnetic field. The law of electromagnetic induction can be expressed as follows:

If a magnetic flux (ψ)—*i.e.*, the area integral of $\boldsymbol{B}$ threading a closed loop (which may coincide with an actual physical conductor or be a mathematically defined closed line)—changes, the loop remaining fixed, then the electromotive force induced around the loop is $-d\psi/dt$. The sense of the emf is given by Lenz's law: If the electromotive force acts in a real conducting loop, the resulting current produces a field that tends to oppose the change in $\boldsymbol{B}$ and ψ.

Faraday's law can be expressed in mathematical form by considering a plane area S bounded by a closed loop with a line element $\boldsymbol{dl}$. If the positive normal to S—that is, $\boldsymbol{dS}$—is taken to be in the direction that a right-handed screw would advance as the path $\boldsymbol{dl}$ is traced out, then the emf acting around the loop is the integral expression

$$\oint \boldsymbol{E} \cdot \boldsymbol{dl} = -\frac{d}{dt}\int_S \boldsymbol{B} \cdot \boldsymbol{dS}, \tag{47}$$

and, by considering a loop of infinitesimal area, this leads to the equation

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}. \tag{48}$$

Equation (48) will not be needed in this section but is required below in the section on *The primary effects and properties of electric fields and currents.* It is one of the most important formulas in the whole of electromagnetism. For this section the original statement of Faraday's law is adequate.

If a changing current I_1 flows in a circuit, the flux linking a nearby circuit will also change. According to Faraday's law there will be an induced emf proportional to an $\dot{I}$ equal to dI/dt in the second circuit. If this is written as $V_2(t) = M_{21}\dot{I}_1(t)$, M_{21} is called the mutual inductance between the two circuits (1) and (2). The units of M are volt ampere^{-1} second and are expressed as a henry (H).

Mutual inductance

The current in the first circuit, as it changes, will also change the flux in the first circuit, and there will be an induced emf in the first circuit. The sign will be such as to oppose the change in I. Treated as a back emf in the sense that the back emf in a resistance is RI and in a capacitance q/C, it gives $V_1(t) = L_1\dot{I}_1(t)$, and L_1 is the self-inductance of circuit (1). If two circuits have self-inductances L_1 and L_2 and mutual inductances M_{21} and M_{12} and if the current in (1) is raised to I, that in (2) remaining at zero, then the work done by sources of emf in (1) is

$$W_1 = \int I_1 V_1 dt = \int I_1 L_1 \dot{I}_1 dt = \frac{1}{2}L_1 I_1^2.$$

If the second current is raised to I_2, the source of emf in circuit (2) has to do work

$$W_2 = \frac{1}{2}L_2 I_2^2;$$

but, because of the back emf in (1) from $\dot{I}_2$, the first source also does work

$$W_{12} = \int I_1 M_{12} \dot{I}_2 dt = M_{12} I_1 I_2.$$

Thus the total work done to reach a final state, with I_1 in one circuit and I_2 in the other, is

$$\frac{1}{2}L_1 I_1^2 + \frac{1}{2}L_2 I_2^2 + M_{12} I_1 I_2.$$

This state could have been reached by increasing I_2 first and then I_1, which gives $\frac{1}{2}L_2 I_2^2 + \frac{1}{2}L_1 I_1^2 + M_{21} I_2 I_1$. These two expressions must be equal, and thus $M_{12} = M_{21}$.

THE MATHEMATICAL AND GRAPHICAL REPRESENTATION OF SINUSOIDAL PHENOMENA

As mentioned above, the basic circuit elements that constitute alternating current circuits are the voltage generator (source of alternating emf, symbolized ~), resistance (R), capacitance (C), self-inductance (L), and mutual inductance (M). Figure 19 shows a circuit incorporating all these elements. The equations that govern the operation of this circuit are

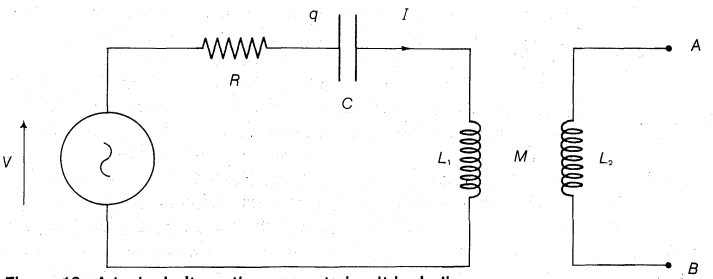


Figure 19: A typical alternating-current circuit including resistance R, capacitance C, inductance L, and mutual inductance M (see text).

$$V(t) = RI(t) + \frac{q(t)}{C} + L_1\dot{I}(t),$$

$$I(t) = \frac{dq(t)}{dt},$$

and

$$V_{AB}(t) = M\dot{I}(t),$$

in which V_{AB} is the voltage between terminals A and B. Finding the current in terms of voltage involves solving a set of linear differential equations with constant coefficients.

As shown above, the current in a circuit with resistance, capacitance, or inductance is sinusoidal if the impressed voltage is sinusoidal and is of the same frequency. The current and voltage, however, are not necessarily in phase, or step—*i.e.*, they do not rise and fall in unison so that they are maximum or minimum simultaneously. If there is only resistance in a circuit in which $V = V_0 \cos \omega t$, then because $I = RV$, or $I = RV_0 \cos \omega t$, it is seen that both V and I oscillate with $\cos \omega t$ and are in phase. If there is only an inductance L in the circuit, the current and voltage associated with L are related by $V = L\dot{I}$, and if $I = I_0 \cos \omega t$, then $V = -\omega L I_0 \sin \omega t = \omega L I_0 \cos(\omega t + \pi/2)$. Thus the amplitude of V is $\omega L I_0$, and it leads I by a phase angle $\pi/2$. If there is only a resistance in the circuit, the amplitude is RI_0 and the phase angle is zero. If there is a capacitance alone in the circuit, $I = C\dot{V}$, the amplitude of I is $\omega C V_0$, and I leads V by $\pi/2$; or, if I is given, the amplitude of V is $I_0/\omega C$, and V lags I by $\pi/2$. Figure 20 shows these relations, expressed in terms of V as the in-phase vector.

A circuit with inductance, resistance, and capacitance in series may now be considered. When a current $I(t) = I_0 \cos \omega t$ flows through the circuit the voltage is a sum of three terms of amplitudes I_0R in phase, $I_0/\omega C$ lagging by $\pi/2$, and $I_0\omega L$ leading by $\pi/2$. If these are represented on a vector diagram, as in Figure 20, the amplitude and phase can be found by simply adding the three vectors. Clearly, the amplitude is

$$V_0 = \left[R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2\right]^{1/2} I_0,$$

and the phase (leading is positive) is

$$\theta = \tan^{-1}\frac{\omega L - 1/\omega C}{R}$$

Although these relations are somewhat simple, they become hopelessly intricate when dealing with complicated circuits.

In order to simplify these relations, use is made of the notation of complex numbers. With real numbers, no meaning is attached to the square root of a negative num-

ber. If, however, an imaginary number j is defined to be the square root of -1 (*i.e.*, $j = \sqrt{-1}$) it is possible to define complex numbers; *e.g.*, $c = a + jb$, in which a and b are real, and therefore, $c^2 = a^2 + 2jab - b^2$. If b is equal to 0, then c is the square root of a positive number; if a is equal to 0, then c is the square root of a negative number.

Real numbers and imaginary numbers of the form jb are entities of different kinds and so if $c = a + jb$ and $\gamma = \alpha + j\beta$, the equation $c = \gamma$ implies the two equations $a = \alpha$ and $b = \beta$. If two complex numbers are equal, their real and imaginary parts are separately equal. This is reminiscent of vectors. If two vectors P and Q in two dimensions are equal, then their x components and y components are separately equal. Furthermore, the law of addition $c + \gamma = a + \alpha + j(b + \beta)$ is like the law of vector addition $(P + Q)_x = P_x + Q_x$, $(P + Q)_y = P_y + Q_y$. Thus, complex numbers can also be represented by a vector diagram, or Argand diagram, so-called because Jean Robert Argand, a Swiss mathematician, first used it (see Figure 21). The complex number c is regarded as a vector of length $(a^2 + b^2)^{1/2}$ at an angle $\theta = \tan^{-1}\frac{b}{a}$. This corresponds to writing c in the form

The Argand diagram

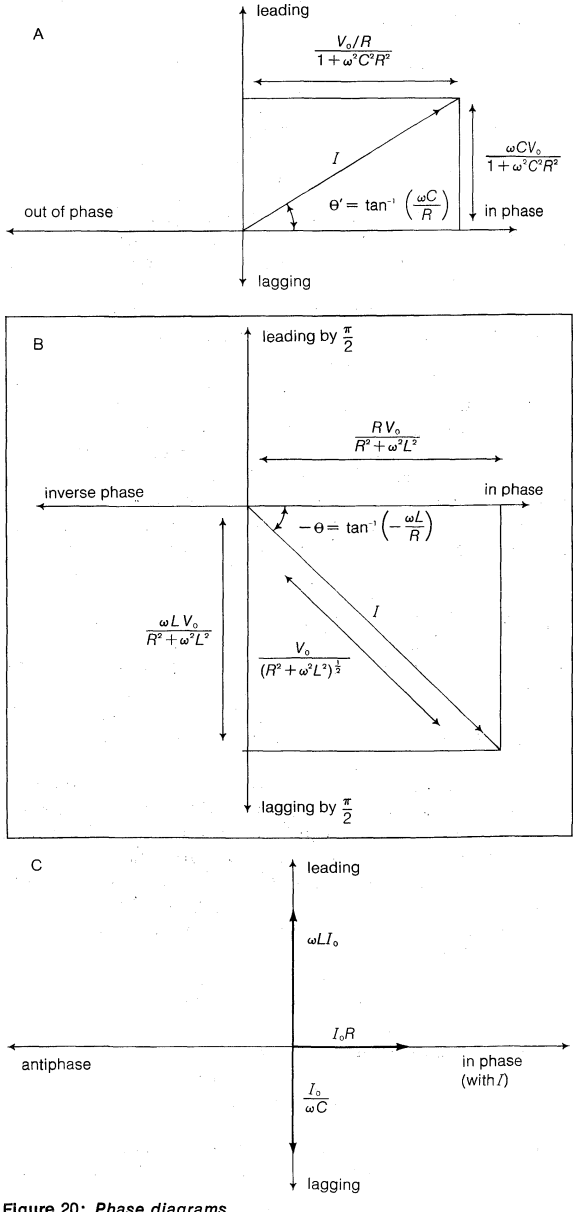


Figure 20: *Phase diagrams.*
(A) Resistance R and capacitance C in series; (B) resistance R and inductance L in series; (C) resistance, inductance, and capacitance in series (see text).

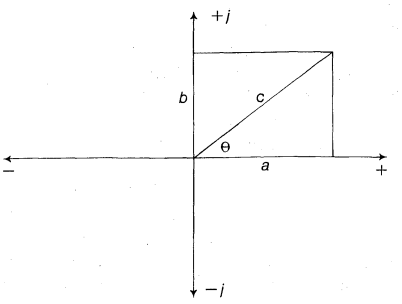


Figure 21: Argand diagram for a complex number $c = a + jb$ (see text).

$$c = (a^2 + b^2)^{1/2}\cos\theta + j(a^2 + b^2)^{1/2}\sin\theta$$
$$= (a^2 + b^2)^{1/2}(\cos\theta + j\sin\theta).$$

As an example in the application of complex numbers, a current $I_0 \cos \omega t$ flows through a circuit in which the voltage $V(t)$ across the circuit has been found to be $V(t) = I_0(R\cos\omega t - X\sin\omega t)$, so that the amplitude of $V(t) = I_0(R^2 + X^2)^{1/2}$ and the phase leads by $\tan^{-1}\frac{X}{R}$. This could be expressed in complex notation simply as $V = I(R + jX)$, and the Argand diagram for V with I defining the positive real axis would be the same as the vector diagram for $V(t)$ with $I(t)$ defining the in-phase axis.

The complex notation is a powerful technique for solving ac problems. With sinusoidal currents and voltages of frequency ω (strictly $\omega/2\pi$) each inductance L is replaced by an "effective resistance" $j\omega L$, and each capacitance C is replaced by an effective resistance $1/j\omega C$. A resistance remains a real number R. The total effective resistance is then calculated as though dealing with dc phenomena. At the end of the calculation the real part of the effective resistance describes the in-phase terms and the imaginary part describes the out-of-phase terms.

These considerations so far have been based on the analogy between the Argand diagram and the vector diagram. There is a much more fundamental and useful way of arriving at the same results, based on De Moivre's theorem (named after the French-born British mathematician Abraham De Moivre). De Moivre's theorem makes it easier to find solutions to the differential equations of ac theory. This theorem states that the logarithmic base (e) raised to an imaginary power is equal to the sum of a real and an imaginary number, $e^{j\theta} = \cos\theta + j\sin\theta$.

It follows from De Moivre's theorem that

$$e^{j\omega t} = \cos\omega t + j\sin\omega t$$

and that

$$\cos\omega t = Re(e^{j\omega t})$$
$$\sin\omega t = Im(e^{j\omega t})$$

in which *Re* means "real part of" and *Im* means "imaginary part of."

The simplification that results from complex notation is ultimately the result of the fact that

$$\frac{d}{dt}(e^{j\omega t}) = j\omega e^{j\omega t}.$$

As a result all time derivatives are simply replaced by $j\omega$. Because of this relation, the differential equations pertinent to the circuit in Figure 19 may be reduced to algebraic equations. This may be contrasted with the first two equations in this section, which involve both a multiplication by ω and a phase shift.

The rules for dealing with sinusoidal phenomena can now be summarized. If $I(t) = A\cos\omega t + B\sin\omega t$, then $I(t)$ is represented as either $Re(I_0e^{j\omega t})$, in which $I_0 = A - jB$, or, if preferred, by $I(t) = Re(I_{00}e^{j\omega t + j\phi})$, in which $I_{00} = (A^2 + B^2)^{1/2}$ and $\tan\phi = -B/A$. In the differential equations describing the circuit $I(t)$ is replaced by $Re(j\omega I_0e^{j\omega t})$. In the complete equations all the *Re* signs are dropped and the common factors $e^{j\omega t}$ are can-

celled. This leaves a set of algebraic equations relating complex numbers. When these equations have been solved and V_0 is given in terms of I_0, for example, the real time-dependent voltage is $V(t) = Re(V_0 e^{j\omega t})$. Further, if $V_0 = (X - jY)$, the amplitude of

$$V(t) = (X^2 + Y^2)^{1/2}$$

and its phase angle is $\tan^{-1} y/x$. In practice this reduces to leaving resistances as real quantities and replacing inductances L and capacitances C by reactances $j\omega L$ and $1/j\omega C$. The circuit equations are then solved using dc techniques.

Reactance and impedance

The term reactance is reserved for an imaginary quantity such as $j\omega L$, $1/j\omega C$, or $j(\omega L - 1/\omega C)$. A mixed term (*e.g.*, $R + j\omega L$) is called an impedance.

If $c = a + jb$, the number $c^* = a - jb$ is called its complex conjugate (cc), and, therefore, the product of a number and its complex conjugate is $cc^* = a^2 + b^2$. Thus $(cc^*)^{1/2}$ is the length of c on the Argand diagram. If γ and c are complex numbers and $\gamma = \alpha + j\beta$, then

$$\frac{\gamma}{c} = \frac{\gamma c^*}{cc^*} = \frac{\alpha a + \beta b + j(a\beta - \alpha b)}{a^2 + b^2}. \qquad (49)$$

The denominator is real and so equation (49) can be used as a general rule for picking out the real and imaginary parts of a complex fraction.

If $V(t) = Re(V_0 e^{j\omega t})$ and $V_0 = U + jW$, this means that $V(t) = U \cos \omega t - W \sin \omega t$. Similarly, if $I(t) = Re(I_0 e^{j\omega t})$ and

$$I_0 = H + jJ, I(t) = H \cos \omega t - J \sin \omega t,$$

then

$$I(t)V(t) = UH \cos^2 \omega t + WJ \sin^2 \omega t - (UJJ + WH) \sin \omega t \cos \omega t.$$

The average power (P_{av}) dissipated in a circuit in a time T is given by

$$P_{av} = \frac{1}{T}\int_0^T I(t)\, V(t) dt,$$

into which the preceding equation of three terms can be substituted. It is presumed that T contains many cycles or that T is exactly an integral number of cycles. In other words, either $\omega T \to \infty$ or $\omega T = 2n\pi$, with n an integer. The three resulting integrals may be integrated as follows:

$$\frac{1}{T}\int_0^T \cos^2 \omega t\, dt = \tfrac{1}{2}, \qquad (50)$$

$$\frac{1}{T}\int_0^T \sin^2 \omega t\, dt = \tfrac{1}{2},$$

$$\frac{1}{T}\int_0^T \sin \omega t \cos \omega t\, dt = 0,$$

resulting in an expression for the average power, $P_{av} = \tfrac{1}{2}(UH + JW)$.

From the relations

$I_0 V_0^* = (H + jJ)(U - jW) = UH + JW + j(JU - HW)$

and

$$I_0^* V_0 = UH + JW - j(JU - HW),$$

it follows that the average power can be written as either

$$P_{av} = \tfrac{1}{2}\, Re(I_0 V_0^*),$$

$$P_{av} = \tfrac{1}{2}(Re I_0^* V_0),$$

or

$$P_{av} = \tfrac{1}{4}(I_0 V_0^* + I_0^* V_0).$$

Assuming that

$$V_0 = ZI_0 = (R + jX)I_0,$$

then the average power becomes

$$P_{av} = \tfrac{1}{2}\, Re([R + jX]I_0 I_0^*) = \tfrac{1}{2} R I_0 I_0^*,$$

so that P_{av} depends only on the real part R of the impedance Z.

The results (50) are of course of more general interest. If a current $I_0 \cos \omega t$ or $I_0 \sin \omega t$ flows in a resistance R, the average power is $\tfrac{1}{2} R I_0^2$. In many cases, especially power engineering, the power (the basis of the monetary charge) is of more interest than anything else. Thus it is quite usual to find in texts a voltage $V_0 \cos \omega t$ or a current $I_0 \cos \omega t$ specified in terms of its rms (root mean square) amplitude $V_0/\sqrt{2}$ or $I_0/\sqrt{2}$. For example, the 120-volt main has a root-mean-square voltage of 120 volts; however, its amplitude or peak voltage is 170 volts. The peak to peak voltage is 340 volts. (The use of amplitudes in this article will not be abandoned for rms quantities unless explicitly stated.)

With a little practice the use of complex numbers to represent ac quantities becomes easy, and, when, for example, it is found that $V_0 = (R + jX)I_0$, it is unnecessary to remark that the amplitude of $V(t)$ is $(R^2 + X^2)^{1/2}$ times that of $I(t)$ and that $V(t)$ leads the phase of $I(t)$ by $\tan^{-1} X/R$; a second thought need not be given to the mathematical infrastructure.

THE BASIC ALTERNATING-CURRENT CIRCUIT LAWS

In dc circuits the basic laws are those defining emf (electromotive force) and Ohm's law together with Kirchhoff's laws. In ac circuits, Ohm's law is extended so that the back emf's caused by capacitors and inductors can be included; in addition, there is a completely different type of effect involving four terminals, the mutual inductance. The back emf's due to resistance R, capacitance C, and inductance L are IR, q/C and LI and in addition, for a capacitance $I = dq/dt$.

Kirchhoff's laws applied to ac circuits maintain the same form. The sum of all the currents leading to any node is zero and the sum of the emf's around any closed loop is zero.

Kirchhoff's laws in ac circuits

Next, the physical significance of the two Kirchhoff laws in ac circuits may be considered. If two resistances bring currents I_1 and I_2 to a node connected to a capacitance with a charge q, conservation of charge implies that $I_1 + I_2 = dq/dt$; but dq/dt is, in ac circuit terms, the current I_3, and so the law $I_1 + I_2 - I_3 = 0$ still depends on the law of charge conservation. In dc circuits the second Kirchhoff law is related to the property of the electrostatic field expressed by $\nabla \times \boldsymbol{E} = 0$—*i.e.*, the line integral of $\boldsymbol{E}$ round any closed loop is zero. In time-dependent situations, Faraday's law gives $\nabla \times \boldsymbol{E} = -\boldsymbol{B} \neq 0$. In the circuit shown in Figure 22, the line integral of

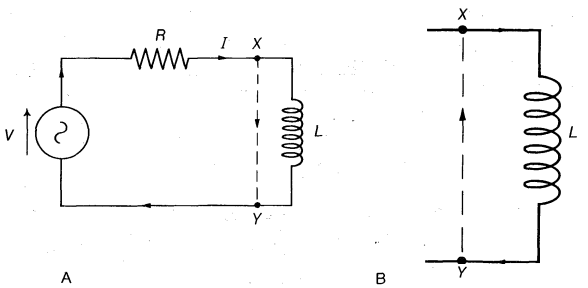


Figure 22: *Paths of integration (shown by arrows).* (A) Excluding changing magnetic flux; (B) including changing magnetic flux (see text).

$\boldsymbol{E}$ is taken round a closed loop that starts at the top of the voltage generator, passes through R, then leaves the conductor passing outside L from X to Y, and rejoining the conductor below L, finally returns through the generator, and is zero because it links no changing magnetic flux. The contribution to this line integral from the path XY outside L is LI, which is equal to $d\psi/dt$, the flux change linked by the windings of L. The electric field in these windings satisfies the equation

$$\oint \boldsymbol{E} \cdot \boldsymbol{dl} = -\frac{d\psi}{dt}, \qquad (51)$$

taking the closed path of integration as shown in Figure 22B. Because the section YX contributes $-d\psi/dt$, $\boldsymbol{E}$ must be zero in the actual conductor. Thus Kirchhoff's law is no longer related to the line integral of $\boldsymbol{E}$ round the actual circuit but to the equation (51). By including back emf's from self-inductance, or induced emf's from mutual inductance, however, the term $d\psi/dt$ has been explicitly included in the circuit equations.

In dc circuits any real source of emf, such as a battery, has a finite internal resistance. Real sources of alternating emf's such as motor alternators or electronic oscillators will not only have internal resistance but may also have a more complicated internal impedance. Whatever load is connected to the terminals, the voltage V_L and current I_L in this load must bear a linear relation to each other. In complex notation this gives

$$V_{LO} = -ZI_{LO} + V_{SO},$$

in which $Z = R + jX$ may be complex. The open circuit voltage is V_{SO}, and the short circuit current is V_{SO}/Z. Thus a source of emf can always be represented by an equivalent circuit incorporating an internal impedance Z. It is notable, however, that Z will almost certainly depend on the frequency of the alternating voltage. Again, as in dc circuitry, impedance can be measured as the ratio of the open circuit voltage to the short circuit current. Here a new problem may arise, for the short circuit current and the open circuit voltage need not be in phase. Thus the measurement gives only Z's magnitude *i.e.*, $|Z| = (ZZ^*)^{1/2}$. To obtain a complete specification requires a more elaborate measurement.

The power delivered to the load, averaged over a cycle, is $\frac{1}{2}R_L I_0 I_0^*$ and this is a maximum when the load reactance X_L is chosen to cancel the reactance X. It is then

$$P_{av} = \tfrac{1}{2}R_L V_0 V_0^*/(R + R_L)^2,$$

and this can be further maximized by choosing R_L to be equal to R. When this is done the maximum average power is $V_0 V_0^*/8R$, or, in terms of the rms open circuit voltage, $V^2_{rms}/4R$. The available power thus depends only on voltage V_0 and resistance R and not at all on reactance X. Whereas in dc circuits the optimum load is equal to the internal resistance of the generator, in ac circuits the optimum load is the complex conjugate of the internal impedance $Z_L = Z^* = R - jX$.

In some applications, particularly those in which there are a number of circuit elements in parallel across a source emf, it is convenient to represent the source as a current generator in parallel with an impedance Z. If the strength of the current generator is $I(t) = Re(I_0 e^{j\omega t})$, the open circuit voltage is $V(t) = Re(I_0 Z e^{j\omega t})$ and the short circuit current is $I(t) = Re(I_0 e^{j\omega t})$.

In dc circuits the conductance (G) of a circuit element is the inverse of its resistance. In ac circuits the quantity $Y = 1/Z$ is called the admittance and is complex. If $Z = R + jX$, the admittance is

$$Y = \frac{1}{R + jX} = \frac{R - jX}{R^2 + X^2};$$

and, when the admittance Y is separated into real and imaginary parts, $Y = G + jB$, the conductance G is

$$G = \frac{R}{R^2 + X^2},$$

and the susceptance B is defined as

$$B = -\frac{X}{R^2 + X^2}.$$

THE DETECTION AND MEASUREMENT OF ALTERNATING CURRENTS AND VOLTAGES

Heating effects of ac in a resistance

The simplest method of detecting the presence of either ac currents or voltages is by observing their heating effects in a resistance. A number of ac meters operate on this principle. In one form of the hot wire ammeter, the current flows in a fine wire attached to a pivoted needle, and thermal expansion of the wire moves the needle. A somewhat more sensitive instrument uses the heat produced to alter the temperature of a small bead of a semiconducting compound, and the heat produced is detected or measured by observing the change in the dc resistance of the bead. This technique is especially useful at high (microwave) frequencies. If the frequency lies in the audio range, quite weak ac signals can be detected using a pair of headphones. The force acting between the two plates of a capacitor is proportional to the square of the potential difference. If this is $V \cos \omega t$ the force $KV^2 \cos^2 \omega t = \frac{1}{2}KV^2(1 + \cos 2\omega t)$, in which K is a constant, contains a static component that can be measured. This is not a sensitive method of detecting weak signals, but it can be used to construct a simple instrument for measuring alternating voltages above about 100 volts. The usual electrostatic voltmeter found in the laboratory is not an accurate instrument, but it is possible to construct accurate meters and these are of practical importance.

The most sensitive method of detecting the presence of a weak ac signal undoubtedly involves the use of an electronic amplifier to bring it up to a level at which it can be used to operate a loudspeaker, a cathode ray oscilloscope, or some form of meter. The two most basic methods of measuring alternating currents and voltages are by the cathode ray oscilloscope and by the use of a rectifier to produce a direct current that can then be measured in the usual way.

In the cathode ray oscilloscope (CRO), or cathode ray tube (CRT), an electron beam in a high vacuum is focussed to a point on a fluorescent screen, where it produces a bright spot of light. The beam passes in vacuum between two pairs of electrostatic deflection plates, and potential differences applied across these plates deflect ("sweep") the beam, and the spot of light, horizontally (by convention, the X direction) or vertically (in the Y direction). The amplitude of an alternating voltage can be measured to about one-percent accuracy by the magnitude of the deflection it produces. The phase angle between two different voltages can be obtained by applying one voltage to the X plates and the other to the Y plates and observing the elliptical trace that is produced. The most usual mode of using a CRO is, however, to apply a linear sweep to the X plates and the unknown voltage to the Y plates. If the sweep duration occupies several cycles of the alternating voltage, the wave form is traced out on the screen. The cathode ray tube itself is a relatively insensitive device and requires a potential difference between the Y plates of 100 volts to produce a full scale deflection; an amplifier is thus usually required to detect and measure small signals. Most commercial oscilloscopes incorporate a calibrated linear amplifier and a time-base circuit for producing the linear X sweep. An oscilloscope in conjunction with a sensitive, high-gain amplifier is perhaps the most useful instrument for detecting the presence of small alternating signals.

Rectifying ac currents

A junction between n-type and p-type silicon passes current easily when the n-type material is made negative and is almost an insulator when the polarity is reversed. Many other systems have this property in varying degrees and can be used to rectify alternating currents. If a perfect rectifier, one which behaved as a short circuit in one direction and an open circuit in the other, existed, it would rectify all small alternating signals. Practical rectifiers are not perfect and usually, although little current flows with reverse voltage, an appreciable voltage, of from 0.1 volt to one volt is required before the forward current begins to rise rapidly. As a result, ac instruments based on the dc measurement of rectified ac cannot be used at extremely low voltages unless they are preceded by an amplifier. Numerous electronic circuits have been devised to overcome this problem, and in practice there is no difficulty in measuring ac voltages down to a few millivolts.

The simple rectifier circuit consisting of a rectifier element and a dc meter in series is not an efficient system and most ac instruments use the bridge rectifier shown in Figure 23. When the point X is positive with respect to Y, the rectifiers A and C are conducting, the + output lead is connected to X, and the — lead is connected to Y. When X is negative, D and B conduct and again the + lead is connected to the positive input, which is now Y, and the — lead is connected to X, which is negative. Both half cycles of the alternating voltage contribute to the output.

In general, with modern equipment, ac measurements at frequencies below about one megahertz (10^6 Hz) are no more difficult than dc measurements, although the accuracy attainable is appreciably less. Above about one

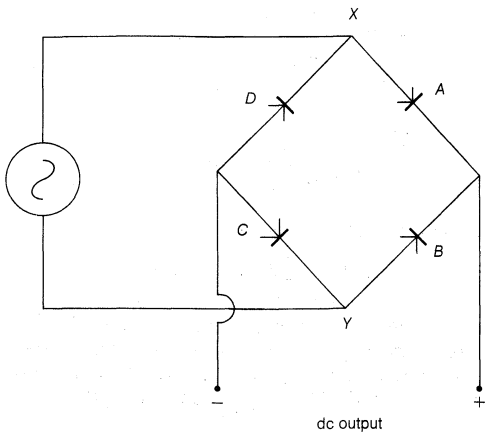


Figure 23: Bridge rectifier (see text).

megahertz, effects caused by stray lead inductance and capacitance become a serious nuisance, and, in addition, the sensitivity and calibration of measuring instruments tends to deteriorate. The only high-frequency electrical measurement of high accuracy that can be made with any ease is that of frequency itself.

In the subsections below it is assumed that no problems arise in the detection and measurement of ac quantities. It should be kept in mind, however, that elaborate calculations of the properties of ac networks at high frequencies are of doubtful utility when precise measuring instruments, as well as precisely specified circuit components, are not available. Computer applications to ac circuit theory are justified only because the calculations are faster and simpler, not because they are more accurate, than manual calculations.

RESONANT CIRCUITS

The parallel resonant circuit shown in Figure 24 is not only important in its own right but its behaviour also has far-reaching analogies throughout the rest of physics. The equations that relate the current (I), the charge (q), and the voltage (V) in terms of the capacitance (C) and inductance (L) are

$$I = \frac{dq}{dt},$$
$$q = CV,$$
and
$$L\dot{I} + V = 0.$$

If the quantities I and q are eliminated from the above three equations there remains

$$LC\frac{d^2V}{dt^2} + V = 0.$$

This is the classical harmonic oscillator equation and its general solution is

$$V(t) = A\cos\omega_0 t + B\sin\omega_0 t,$$

in which

$$\omega_0 = \left(\frac{1}{LC}\right)^{\frac{1}{2}}. \qquad (52)$$

This solution corresponds to an undamped oscillation of natural frequency $\omega_0/2\pi$. The current in the circuit is $I(t) = C(dV/dt) = \omega_0 C(B\cos\omega_0 t - A\sin\omega_0 t)$.

The general solution can also be written as

$$V(t) = V_0\cos(\omega_0 t + \phi)$$

and

$$I(t) = \omega_0 CV_0\cos(\omega_0 t + \phi + \pi/2).$$

The energy (U) stored in the system is

$$U = \tfrac{1}{2}LI^2(t) + \tfrac{1}{2}CV^2(t),$$

and, since $\omega_0^2 LC = 1$ from (52), this can be expressed as

$$U = \tfrac{1}{2}CV_0^2(\cos^2[\omega_0 t + \phi + \pi/2] + \cos^2[\omega_0 t + \phi]) = \tfrac{1}{2}CV_0^2.$$

In terms of the peak current, $I_0 = \omega_0 CV_0$, this is

$$U = \tfrac{1}{2}LI_0^2.$$

At the instant when V reaches its peak value, I is zero and all the energy is stored in C; a quarter of a cycle later, I reaches its peak value and V is zero; the energy is now all stored in L. Thus the energy oscillates back and forth between the capacitor and the inductor.

If the resonant circuit contains resistance in series with L as shown in Figure 24B, the basic equations will be $I = dq/dt$, $V = q/c$, and $L\dot{I} + RI + V = 0$; and these give the differential equation

$$LC\frac{d^2V}{dt^2} + RC\frac{dV}{dt} + V = 0.$$

The behaviour of the solutions of this equation depends critically on whether $(RC)^2 \geq 4LC$, or $(RC)^2 < 4LC$. If $(RC)^2 > 4LC$, the solution is of the form

$$V(t) = Ae^{-\alpha_1 t} + Be^{-\alpha_2 t},$$

in which the coefficients α_1 and α_2 are real. This behaviour corresponds to an exponentially decaying excitation. When $(RC)^2 = 4LC$ the solution is of the form

$$V(t) = (A' + B't)e^{-Rt/2L}$$

and is also exponentially damped. The most interesting case is that in which $(RC)^2 < 4LC$ and, in fact, in most cases $(RC)^2 << 4LC$. In this case the general solution is a damped oscillation

$$V(t) = (A\cos\omega t + B\sin\omega t)e^{-Rt/2L} \qquad (53)$$

in which

$$\omega = \omega_0\left(1 - \left[\frac{R}{2\omega_0 L}\right]^2\right)^{\frac{1}{2}},$$

and $\omega_0 = (LC)^{-\frac{1}{2}}$ is the natural (angular) frequency of the undamped oscillations.

The dimensionless parameter $\omega_0 L/R$ is of such importance that it is given a name, the quality factor Q. In terms of Q,

The quality factor Q

$$\omega = \omega_0\left(1 - \frac{1}{4Q^2}\right)^{\frac{1}{2}}.$$

In practice, the quality factor is usually quite large, values of from 10 to 10^5 occurring in resonant circuits, and so approximately

$$\omega \simeq \omega_0\left(1 - \frac{1}{8Q^2}\right).$$

Even if Q is as small in value as three, ω differs from ω_0 by only 1.3 percent. The difference between ω and ω_0, the damped and undamped frequencies, may almost always be neglected in calculations.

The exponential decay term in (53) can be expressed in terms of Q as $e^{-\omega_0 t/2Q}$. Thus the amplitude of the oscillations decays by a factor e^{-1} in a time $\frac{2Q}{\omega_0}$, and in this time the number of cycles of oscillation is Q/π. The decay (damping) in one cycle of oscillation is $e^{-\pi/Q}$, and the parameter π/Q is sometimes called the logarithmic decrement.

The stored energy in the circuit can be expressed in terms of the peak current (I_0) as $\frac{1}{2}LI_0^2$. This current flows through the resistance (R) and the average power (P) dissipated is $P = \frac{1}{2}RI_0^2$. The dimensionless ratio $\omega_0 U/P$ is equal to the quality factor since

$$\frac{\omega_0 U}{P} = \frac{\omega_0 \frac{1}{2}LI_0^2}{\frac{1}{2}RI_0^2} = \frac{\omega_0 L}{R} = Q.$$

Thus Q can also be regarded as a measure of how long, measured in radians of the oscillation, it takes the stored energy to decay. For many purposes this is the most useful definition of Q. It can be used in circumstances in which Q is not given by any simple expression in terms of circuit parameters, such as L, C, and R.

The parameter Q has two other important properties related to the response of a resonant circuit to an impressed electromotive force or current. Figure 24 includes in the circuit a voltage generator $v \cos \omega t$, in which ω is not necessarily equal to ω_0. The current in the circuit is

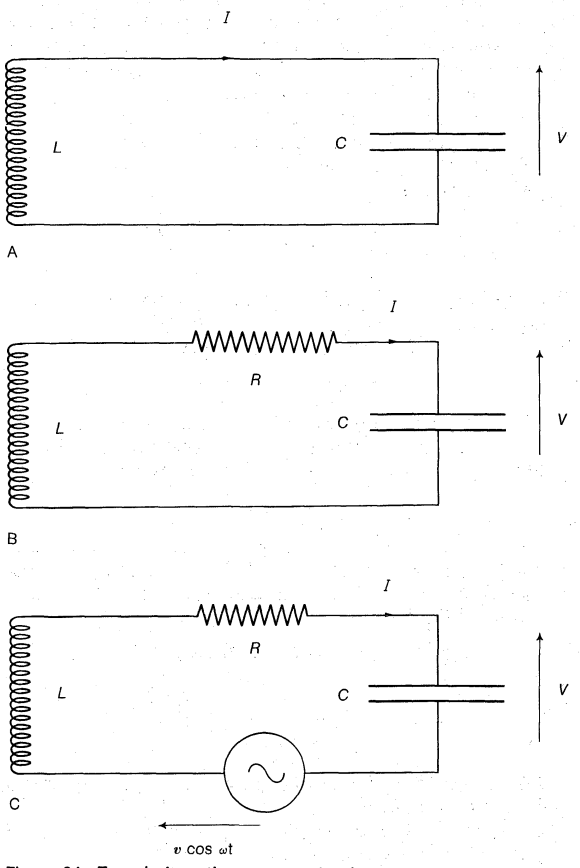


Figure 24: *Tuned alernating current circuits.*
(A) Lossless; (B) series loss; (C) forced (see text).

$$I = \frac{v}{j\omega L + R + \frac{1}{j\omega C}}$$

and the voltage (V) across the capacitor is

$$V = \frac{I}{j\omega C};$$

so that

$$V = \frac{v}{1 - \omega^2 LC + j\omega CR}$$

or

$$V = \frac{[(1 - \omega^2 LC) - j\omega CR]v}{(1 - \omega^2 LC)^2 + \omega^2 C^2 R^2}.$$

It can be seen from this equation that there are in-phase and out-of-phase components of V and that

$$|V|^2 = \frac{v^2}{(1 - \omega^2 LC)^2 + \omega^2 C^2 R^2}.$$

In many applications in electronics, a tuned circuit is used as a frequency selective element, and the interest is in the behaviour of the impedance or admittance of the circuit, especially near the resonant frequency. The impedance Z measured across the capacitance C of the circuit of Figure 24C when the voltage generator $v \cos \omega t$ is removed is

$$Z = \frac{j\omega L + R}{1 - \omega^2 LC + j\omega CR}.$$

If ω is near the resonance frequency, $\omega_0 = (LC)^{-\frac{1}{2}}$, Q is large, and it is permissible to neglect R in the numerator compared with ωL and express Z as

$$Z = \frac{Z_0}{1 + j\frac{\omega_0 L}{R}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)},$$

or

$$Z = \frac{Z_0}{1 + jQ\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)},$$

in which the constant Z_0 has the value

$$Z_0 = \frac{L}{CR} = \frac{Q}{\omega_0 C} = Q\omega_0 L$$

and is the impedance at resonance when $\omega = \omega_0$. If Q is large and only the region near resonance is of interest and ω is expressed as $\omega = \omega_0(1 + \Delta)$ in which Δ is small, the resonance term Z/Z_0 can be approximated by the function

$$F = \frac{1}{1 + jQ\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)} \simeq \frac{1}{1 + 2jQ\Delta}.$$

This can also be written as

$$F = \frac{1 - 2jQ\Delta}{1 + 4Q^2\Delta^2}.$$

The real part—*i.e.*, the part of Z (or Y) that leads to a component of voltage (current) in phase with the impressed current (voltage)—has a peak when $\Delta = 0$ (*i.e.*, when $\omega = \omega_0$) and is down to half its peak value at $\Delta = \pm \frac{1}{2}Q$. Thus the bandwidth, or frequency interval, between these points is ω_0/Q in terms of angular frequency, or ν_0/Q in terms of frequency $\nu = \omega/2\pi$. The imaginary, or out-of phase, component has a positive peak in which $F = +j/2$ at $\omega = \omega_0(1 - \frac{1}{2}Q)$ and a negative peak in which $F = -j/2$ at $\omega_0(1 + \frac{1}{2}Q)$.

The modulus, or absolute value, of F is

$$|F| = \left(\frac{1}{1 + 4Q^2\Delta^2}\right)^{\frac{1}{2}},$$

and this peaks at $\Delta = 0$. At $\Delta = \pm \frac{1}{2}Q$, $\omega = \omega_0(1 \pm \frac{1}{2}Q)$, $|F|$ is reduced by $1/\sqrt{2}$. The two frequencies $\omega_0(1 \pm \frac{1}{2}Q)$ are often referred to as the half-power points.

Half-power points

Many practical applications (*e.g.*, in radio receivers) require resonant circuits in which the quality factor Q and selectivity are higher than can easily be achieved with simple inductances or coils. In practice it is relatively easy to achieve values of Q of about 200 at frequencies up to about 100 megahertz, and above 100 megahertz, although ordinary tuned circuits of high Q are difficult to construct, it is possible to use resonant lines or hollow cavities.

Indeed, in the optical region it is practicable to design resonant systems with quality factors exceeding 10^6. There is, however, a need for circuits with Q of this order at low radio frequencies, and, for this purpose, quartz-crystal piezoelectric resonators are employed.

If a quartz crystal is mounted between two plates, the piezoelectric stress S induced in the crystal is proportional to the plate voltage, thus $S = \gamma V$, in which γ depends on the plate geometry and the piezoelectric coefficient of the crystal. The induced polarization (β) of the crystal, and thus the charge (q) on the plates, is proportional to the strain s resulting from the stress S—*i.e.*, $q = \beta s$. Thus the current in the lead to the plate is $I = \beta \dot{s}$. The crystal is a mechanical resonant system the natural frequency ω_0 of which is determined by its shape and elastic moduli. The resonance is damped by internal friction and also air damping and the presence of the electrodes. It is then possible to show that, if ρ is the frictional damping term, the frequency response of the admittance (Y) is

$$Y = \frac{Y_0}{1 + j\frac{\omega_0}{\rho}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)},$$

which is a typical resonant response with $Q = \omega_0/\rho$. The importance of the quartz-crystal resonator is that Q can easily exceed 10^5 throughout the frequency range 10^5 to 10^8 hertz—*i.e.*, the normal radio frequency range. It is also relatively easy to design the crystal so that Y_0 has a value compatible with conventional electronic circuit components.

When two tuned circuits are coupled together, new phenomena appear. If both circuits have the same natural resonance frequency ω_0 and quality factor Q and if a coupling term γ is defined so that, when V_1 and V_2 are the voltages across the two circuits, the stored energy can be expressed as $U = \frac{1}{2}\Gamma_1 V_1^2 + \frac{1}{2}\Gamma_2 V_2^2 + \gamma(\Gamma_1\Gamma_2)^{\frac{1}{2}} V_1 V_2$, then, when $Q\gamma$ exceeds unity, the single resonance at $\omega = \omega_0$ splits into two resonances (here Γ_1 and Γ_2 are energy coefficients or capacitances associated with the individual tuned circuits). If the coupling is through a capacitance, one resonance occurs at $\omega = \omega_0$ and the other at a lower frequency. If, as in the interstage transformers of a radio receiver, the coupling is through mutual inductance, the two frequencies lie above and below ω_0. The coupled, tuned circuit finds widespread applications in electronics as a band-pass element.

ALTERNATING-CURRENT BRIDGES

In dc circuits the Wheatstone bridge is a simple and accurate instrument for measuring resistance. The ac analogues of the Wheatstone bridge are equally important for determining the impedance of circuit elements. There are, however, a number of significant differences between ac and dc measurements. First of all, an ac impedance involves two numbers. These numbers may be regarded as either its real and imaginary parts, or as its in-phase and out-of-phase parts, or as its magnitude and phase angle, but, however they are regarded, a more elaborate measurement is required. A second difference is that both components of an impedance are likely to be frequency dependent so that measurements may be needed over a range of frequencies. A third difference is that, whereas in dc measurements all possible elements are of the same type—*i.e.*, resistances—in ac measurements there are four basic elements: the resistance, the capacitance, the inductance, and the mutual inductance. Furthermore, with the possible exception of capacitance, these elements are rarely separable, any real coil of wire has not only inductance but also resistance and some stray capacitance, and any real resistor also has inductance and stray capacitance. It is in fact difficult to construct standard elements of R, L, and C with which to compare an unknown element. Finally, the wiring of the bridge itself contributes significant components of R, L, and C, especially at high frequencies. For this reason ac bridge measurements are usually rather less accurate than dc measurements, except at low frequencies. In many cases it is easier to obtain accurate values by other methods. As an example, if an element is known to be an inductance (L) about one microhenry (10^{-6} henry) and a resistance (R) of about 0.1 ohm much the easiest way of measuring L is to place a standard capacitance C in parallel with the element and measure the resonant frequency ω_0; then $L = 1/\omega_0^2 C$. At the same time, if Q is obtained from the frequency response, then $R = \omega_0 L/Q = 1/Q\omega_0 C$. In the practical laboratory concerned with ac measurements, the Q meter, measuring both ω_0 and Q, is much more generally useful than a bridge.

Differences between ac and dc measurements

Obviously any bridge measurement is only as accurate as the reference elements used in the arms of the bridge, and it is also desirable that the elements should be as nearly as possible simple resistances, capacitances, or inductances. It is relatively easy to construct a capacitance with negligible series inductance and resistance in its leads and not too difficult to construct a resistance in such a way that it has only a slight series inductance or parallel capacitance. It is, however, almost impossible to wind an inductance coil of negligible resistance. Thus, as far as possible, in an ac bridge an attempt is made to use only resistances and capacitances in the reference arms.

The basic ac bridge circuit is similar to the Wheatstone bridge, Figure 13, but has the resistances replaced by impedances. The null detector may be an ac meter, although it is usually more convenient to use an amplifier and oscilloscope. The bridge is balanced and the null detector reads zero when

$$\frac{Z_1}{Z_2} = \frac{Z_1'}{Z_2'}.$$

Because the impedances are complex, this implies two conditions involving both the real and imaginary parts. If $Z_1 = R_1 + jX_1$ is the unknown impedance, then at least one of the other impedances must be complex. There is no need for all three to be complex. If X_1 is positive so that Z_1 is inductive, it would be inconvenient to make Z_1' and Z_2' real since then Z_2 would have to be inductive, which is to be avoided. The impedances Z_2 and Z_1' may be made real, however, and Z_2' capacitative —*i.e.*, $Z_2 = R_2$, $Z_1' = R_3$ and $Z_2' = R_4 - jX_4$. The balance condition then becomes $(R_1 + jX_1)(R_4 - jX_4) = R_2R_3$, or, $R_1R_4 + X_1X_4 = R_2R_3$ and $R_1X_4 = R_4X_1$. If $Z_1 = R_1 + j\omega L_1$ and $Z_2' = R_4 + 1/j\omega C_4$, as shown in Figure 25 (top), then the two balance conditions become

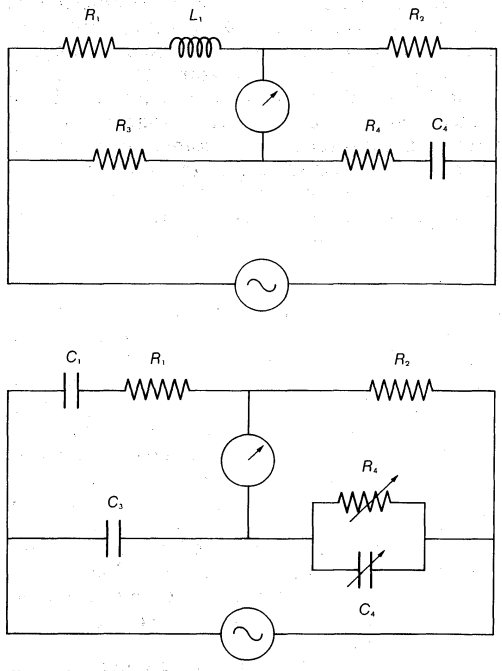


Figure 25: *Alternating-current bridges.*
(Top) Impractical circuit; (bottom) Schering capacitance bridge (see text).

$R_1R_4 + L_1/C_4 = R_2R_3$ and $\omega^2 L_1 C_4 R_4 = R_1$. This result is most unsatisfactory, first of all because the second condition depends on ω, which may not be known, and also because variation of R_4 or C_4 to satisfy this second condition upsets the balance implied by the first condition. For this reason, practical ac bridges do not usually take this form. It is impossible to discuss all possible forms of bridge but one special case: the Schering capacitance bridge, named for the German engineer H. Schering, serves to illustrate the principles involved.

The Schering capacitance bridge

The Schering bridge is shown in Figure 25 (bottom). The unknown capacitance is C_1, R_2 is a fixed known resistance, and C_3 a fixed known capacitance of the same order of magnitude as C_1. Balance is achieved by varying the calibrated components C_4 and R_4. The balance condition is

$$j\omega C_3(R_1 + 1/j\omega C_1) = R_2(1/R_4 + j\omega C_4),$$

and the in-phase and out-of-phase, or real and imaginary, equations are

$$C_3R_4 = C_1R_2 \quad (54)$$
$$C_3R_1 = C_4R_2 \quad (55)$$

Both equations are independent of ω, and (54) can be sat-

isfied by variation of R_4 alone, whereas (55) is satisfied by variation of C_4 alone. Thus, the procedure is first to minimize the detector reading by varying R_4 and then to obtain the final null using C_4.

It has been pointed out above that one of the most useful laboratory ac instruments is the Q meter, and this section on bridges ends with a brief résumé of its operation. The basic circuit is shown in Figure 26. Provided that the

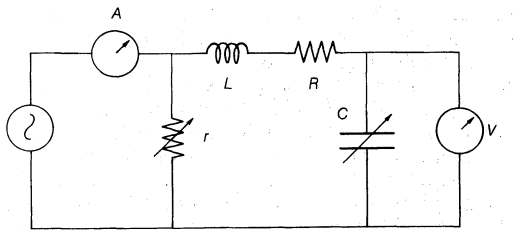


Figure 26: The Q meter (see text).

voltmeter V has a high enough impedance, its reading is a maximum when $\omega^2 LC$ equals one and has the value QIr. In practice neither the ammeter A nor the voltmeter V is accurate enough to make much use of this relation, and, in the usual instrument, the source of the ac voltage is adjusted to bring the ammeter reading to a fiduciary mark, say half full scale; C is then adjusted to maximize the voltmeter reading, and finally the variable resistor r is adjusted to bring this reading to another fiducial mark. The instrument is calibrated so that the value of Q can be read from the setting of r. A typical instrument has a variable frequency-calibrated oscillator covering a range of, say, 100 kilohertz to 100 megahertz.

NONLINEAR ELEMENTS

Metallic resistors, air-cored inductors, and capacitors employing normal dielectric insulators are nearly linear elements; *i.e.*, in the equations $V = IR$, $V = L\dot{I}$, and $V = q/C$, the quantities R, L, and C are nearly independent of V, I, or q. There are, however, many practical components the behaviour of which is appreciably nonlinear. An extreme case is the semiconductor diode in which the current-voltage relation is of the form $I = I_0(e^{V/V_0} - 1)$, in which I_0 and V_0 are constants, but in many other cases there are less drastic departures from linearity. For example, in an iron-cored inductor, the value of L decreases at high currents because of saturation, and, in a capacitor with a ferroelectric ceramic dielectric, the relation between charge and applied voltage is quite often extremely nonlinear.

In most cases the presence of nonlinear components in a system has undesirable effects, but in a few cases the nonlinear response can be put to practical use. The basic effects associated with nonlinear components can easily be seen by considering a nonlinear resistor passing a current $I(t) = I \cos \omega t$. If the voltage current relation can be expressed as the sum of current components (1), (2), (3), etc.,

$$V(t) = R_1 I(t) + R_2 I^2(t) + R_3 I^3(t) + \text{etc.},$$

then, since

$$\cos^2 \omega t = \frac{1}{2}(1 + \cos 2\omega t),$$

$$\cos^3 \omega t = \frac{1}{4}(\cos 3\omega t + 3 \cos \omega t), \text{ etc.},$$

the voltage across the resistor will, in addition to terms at the fundamental frequency, contain terms at the second and third harmonic frequencies 2ω and 3ω. Furthermore, if the current $I(t)$ contains terms at two or more frequencies ω_1, ω_2, the voltage will contain terms at the various beat frequencies $\omega_1 + \omega_2$, $2\omega_1 + \omega_2$, etc. If, for example, the nonlinear element forms part of an audio amplifier used for the reproduction of music, the spurious tones generated in the output will produce a most disagreeable effect on the listener.

The parametric amplifier used in the sensitive microwave receiver required for satellite communication relies on the nonlinear voltage–charge relation $V = Aq^2$ in a special semiconductor diode, utilizing a reverse-biased junction. If the diode is fed with both a signal current at ω_1 and a large "pump" current at $\omega_1 + \omega_2$, it generates an idler voltage at ω_2. This idler voltage drives current at the same frequency ω_2 in a resistive load, and then this idler current recombines with the pump current at $\omega_1 + \omega_2$ to generate a component at ω_1. The new signal frequency current can, if conditions are right, deliver more power to a load than is extracted from the signal source. The device is, therefore, an amplifier. Because it generates much less internal noise than conventional amplifiers, it is particularly useful in dealing with weak signals.

ELECTRIC WAVES IN CABLES

Capacitance and inductance in cables

Any two-wire cable—*e.g.*, a coaxial cable, a parallel wire line, or a twisted pair—has some stray capacitance between the two wires of the cable. Furthermore, when a current flows in the cable, it produces a magnetic field near the conductors, and, when the current changes, the changing magnetic field produces a back electromotive force in the cable. Thus, in general, any cable has a certain capacitance and self-inductance. The capacitance and inductance are proportional to the length of the cable and so may be defined as a capacitance C per unit length and an inductance L per unit length. In some cases, in which the shape and arrangement of the conductors are particularly simple, L and C can be calculated. For example, in a coaxial cable having a circular inner conductor of radius a within a circular sheath of radius b and with the space between the two conductors being filled with an insulator of dielectric constant ϵ, the values for inductance and capacitance are

$$L = \frac{\mu_0}{2\pi} \log \frac{b}{a} \quad \text{and} \quad C = 2\pi\epsilon\epsilon_0 \left(\log \frac{b}{a}\right)^{-1}$$

In this case the product is $LC = \mu_0\epsilon_0\epsilon$, which not only is independent of the dimensions of the line but also can be written $LC = (1/c')^2$, in which c' is the velocity of light in the dielectric. This is in fact a particular example of a general result that follows from Maxwell's field equations. For any cable, the quantity $(LC)^{-1/2}$ is less than the velocity of light in vacuum; if the cable consists of uniform straight conductors $(LC)^{-1/2}$ equals c', in which c' is the velocity of light in the medium surrounding the conductors. It will not be unexpected to find, as explained below, that $(LC)^{-1/2}$ is the velocity with which electric signals propagate in the cable. The fact that $LC \geq (1/c')^2$ means that a cable cannot be designed to have a low stray capacitance C without at the same time increasing its self-inductance L, and vice versa. As an example, the low-capacitance cable used to connect the stylus pickup of a record player to the audio amplifier is not suitable for connecting the amplifier to the speaker, for which a low-inductance cable is required.

If, as this suggests, electric signals propagate with a finite velocity along cables, then, when a sinusoidal voltage $V \cos \omega t$ is applied to one end of a cable of length l with a phase velocity v, the phase of the voltage at the other end of the cable is changed by $\omega l/v$ and the voltage is $V \cos (\omega t - \omega l/v)$. If $\omega = 2\pi \times 60$ second^{-1}, which corresponds to a 60-hertz supply frequency, and, if v is about equal to the velocity of light, 3×10^8 metres per second, then $\phi \simeq 4\pi 10^{-7} l$ and will not be appreciable unless the length, l, is approximately 1,000,000 metres, or 1,000 kilometres, or about 600 miles. On the other hand, if $\omega = 2\pi \times 10^8$ second^{-1} (a typical high radio frequency), then, if $l = 0.75$ metre (about two feet), the phase change is $\pi/2$ (*i.e.* 90°). This phase change is clearly significant. At high radio frequencies the phase shifts and time delays that occur in normal laboratory lengths of wire cannot be disregarded.

Figure 27 illustrates in a schematic way a long two-wire cable with a current I in the upper conductor (and I in the lower conductor) and a voltage V between the conductors. It also shows the series inductance Ldx and capacitance Cdx associated with elements of the line of length dx. Both the voltage and the current vary with

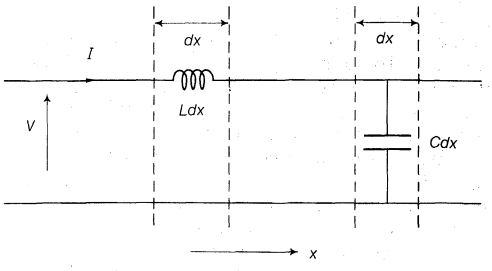


Figure 27: Series inductance and parallel capacitance in a transmission line (see text).

position and time. The instantaneous power flow to the right, past a point x, is $V(x,t)I(x,t)$.

The voltage drop in a length dx is $Ldx\ \partial I/\partial t$, and this yields the differential equation

$$\frac{\partial V}{\partial x} = -L\frac{\partial I}{\partial t}. \tag{56}$$

The charge stored in dx is $CVdx$, and, since a current I flows into dx from the left and $I + dI$ flows out from the right, this charge decreases at a rate dI. This, therefore, gives

$$\frac{\partial}{\partial t}(CV) = -\frac{\partial I}{\partial x},$$

or, because the capacitance C is independent of time,

$$\frac{\partial I}{\partial x} = -C\frac{\partial V}{\partial t}. \tag{57}$$

Equations (56) and (57) govern the behaviour of signals in the line. Before the general solution is written down it is useful to define two new quantities $v = (LC)^{-\frac{1}{2}}$ and $Z_0 = (L/C)^{\frac{1}{2}}$. In terms of these parameters v and Z_0, the current at time t is expressed as

$$I(x,t) = I_f(x - vt) + I_r(x + vt) \tag{58}$$

and the voltage as

$$V(x,t) = V_f(x - vt) + V_r(x + vt) \tag{59}$$

in which

$$V_f = +Z_0 I_f, \tag{60}$$

$$V_r = -Z_0 I_r. \tag{61}$$

The quantities I_f and V_f are functions of x and t in the combination $x - vt$, whereas I_r and V_r involve $x + vt$. If I_f is now defined as $dI_f/d(x - vt)$ and $I_r' = dI_r/d(x + vt)$ this gives $\partial I_f/\partial x = I_f'$, $\partial I_r/\partial x = I_r'$, $\partial I_f/\partial t = -vI_f'$, and $\partial I_r/\partial t = vI_r'$. Equation 56 then yields

$$Z_0(I_f' - I_r') = +Lv(I_f' - I_r'),$$

which is an identity, since $Lv = L(LC)^{-\frac{1}{2}} = (L/C)^{\frac{1}{2}} = Z_0$. The equation (57) is also satisfied identically. Thus the expressions (58) and (59) are the general solution of the cable equations if V and I are related by (60) and (61).

If $I(x,t) = I_f(x - vt)$, the current has the same value at $x = x_0 + vt$ at time t as it had at x_0 at $t = 0$. This solution represents a wave travelling to the right with velocity v. It is a forward wave. Similarly, $I_r(x + vt)$ represents a reverse wave travelling to the left with velocity v.

When only a forward wave is present,

$$V = V_f = Z_0 I_f = Z_0 I,$$

and the power flowing to the right is $VI = Z_0 I^2 = V^2/Z_0$, which is positive. When only a reverse wave is present, $V = -Z_0 I$, and the power flowing to the right is negative —*i.e.*, power flows to the left. In each case the power flows in the direction of travel of the wave.

Characteristic impedance

The quantity Z_0 is called the characteristic impedance of the line. It is the ratio of V to I in a pure forward wave. For a coaxial air-spaced cable,

$$Z_0 = \left(\frac{L}{C}\right)^{1/2} = \frac{1}{2\pi}\left(\frac{\mu_0}{\epsilon_0}\right)^{1/2} \log\frac{b}{a} \simeq 60 \log\frac{b}{a} \text{ ohms}.$$

For reasonable values of b/a, Z_0 lies between 20 and 200 ohms. Other types of cable—*e.g.*, parallel wire lines—have somewhat higher characteristic impedances, but Z_0 rarely exceeds 1,000 ohms. Practically, a two-wire cable will terminate in a resistance R. The voltage (V) at the end of the line has a value IR, but also $I = I_f + I_r$, and $V = V_f + V_r = Z_0(I_f - I_r)$,

so that

$$Z_0\frac{I_f - I_r}{I_f + I_r} = R,$$

and therefore

$$\frac{I_r}{I_f} = \frac{Z_0 - R}{Z_0 + R}.$$

Unless the resistance R and impedance Z_0 are equal, there must be a reflected wave on the line, and some of the power incident on R from a source connected to the left end of the line will be reflected back to the source. When $R = Z_0$, all the power delivered to the line by the source is absorbed in the load R and there is no reflected wave. The load R is then said to be matched to the line. This is a desirable state of affairs to achieve when cables are used to transmit radio-frequency power.

So far no assumptions have been made about the wave form of the electrical signal. There is no need for it to be sinusoidal. The results apply equally well to a train of pulses or a single pulse. If it is specifically assumed that the waves are sinusoidal so that a typical forward wave is

$$\cos\beta(x - vt) = \cos(\beta vt - \beta x),$$

then $\omega = \beta v$ is the (angular) frequency of the wave and $\lambda = 2\pi/\beta$ is the wavelength. Furthermore, $\beta = \omega(LC)^{\frac{1}{2}}$. The typical forward wave can also be expressed as an exponential function so that

$$I_f(x,t) = \mathrm{Re} I_f e^{j(\omega t - \beta x)},$$
$$V_f(x,t) = \mathrm{Re} V_f e^{j(\omega t - \beta x)},$$
$$V_f = Z_0 I_f.$$

The reverse wave is

$$I_r(x,t) = \mathrm{Re} I_r e^{j(\omega t + \beta x)},$$

with

$$V_r(x,t) = \mathrm{Re} V_r e^{j(\omega t + \beta x)},$$

in which

$$V_r = -Z_0 I_r.$$

If in this example the origin of x is taken to be at the load so that x is negative along the line toward the electromotive source, this gives

$$\frac{V_r}{V_f} = -\frac{I_r}{I_f} = \frac{R - Z_0}{R + Z_0},$$

and the voltage at x is

$$V = (V_f e^{-j\beta x} + V_r e^{j\beta x})e^{j\omega t}.$$

The square of the voltage amplitude is found by multiplying the voltage (V) and its complex conjugate (V^*) together,

$$VV^* = V_f^2 + V_r^2 + 2V_f V_r \cos 2\beta x,$$

and this has a maximum value $(V_f + V_r)^2$ when $2\beta x$ equals $n\pi$ ($n = 0, 1, 2, 3$, etc.), and a minimum value $(V_f - V_r)^2$ when $2\beta x$ equals $(n + \frac{1}{2})\pi$. The voltage standing wave ratio (VSWR) is the ratio of the maximum to minimum amplitudes—*i.e.*,

$$\text{VSWR} = \frac{V_f + V_r}{V_f - V_r} = \frac{R}{Z_0}. \tag{62}$$

(Definitions of VSWR differ; many books define the ratio as minimum to maximum rather than maximum to minimum.) A measurement of the VSWR relates R to Z_0. This is an important technique in radio engineering.

The power (P) transferred to the load is $Z_0(I_f^2 - I_r^2) = (1/Z_0)(V_f^2 - V_r^2)$ (using root-mean-square amplitudes). This can be expressed in terms of the maximum rms amplitude: $V_{max} = V_f + V_r$ as $P = \frac{1}{Z_0}\frac{V^2_{max}}{\text{VSWR}}$ or $V^2_{max} = Z_0 P(\text{VSWR})$. Thus, if the load is badly matched to the line and the VSWR is high (it is implicitly assumed in writing [62] that $R > Z_0$, and V_f and V_r have the same sign), then V_{max} is much greater than it need have been to achieve the same power transfer with a matched load. Because at high radio and microwave frequencies the

power-handling capacity of a line is often limited by voltage breakdown, this is a strong reason for attempting to match the load to the line.

Similar calculations can be used to obtain the ratio V/I at points on the line a distance l back from the load. This is, of course, the impedance that would be measured at this point. The calculation, though straightforward, is tedious, and the result is

$$Z = \frac{V}{I} = Z_0 \frac{R \cos \beta l + jZ_0 \sin \beta l}{Z_0 \cos \beta l + jR \sin \beta l}.$$

Here, $\beta l = 2\pi l/\lambda$ so that $\beta l/2\pi$ is the length of the line measured in wavelengths.

Whenever $\beta l = n\pi$ ($n = 0, 1, 2, 3$, etc.)—*i.e.*, the line is an integral number of half wavelengths long—this gives $\sin \beta l = 0$ and $Z = R$ again; but if $\beta l = (n + \frac{1}{2})\pi$—*i.e.*, the line is an odd number of quarter wavelengths long—this gives $\cos \beta l = 0$ and $Z = Z_0^2/R$. The larger R is made, the lower is the value of Z measured at the other end of the line. A quarter wavelength of line acts as an impedance transformer. A cable of characteristic impedance Z_0 and a load R can be matched by inserting a quarter wavelength of line of characteristic impedance Z_0' so that $Z_0 = Z = (Z_0')^2/R$. In other words, Z_0' is the geometric mean of Z_0 and R. This is a useful technique in microwave electronics. It may also be noted that it is essentially the principle underlying the use of thin surface layers (lens coating, or blooming) to reduce reflection of light waves from optical components, such as camera lenses. In the case when R is equal to zero—*i.e.*, the line is shorted at the end—then

Impedance matching

$$Z = jZ_0 \tan \beta l. \tag{63}$$

When βl is small, $\tan \beta l \simeq \beta l$ and this gives

$$Z \simeq jZ_0 \beta l = j(L/C)^{1/2} \omega (LC)^{1/2} l = j\omega Ll.$$

In other words, the line appears to have an inductance Ll, which is hardly unexpected because L is the inductance per unit length. If R is infinite—*i.e.*, the line is open circuited—the result is

$$Z = \frac{Z_0}{j \tan \beta l};$$

and, if βl is small,

$$Z \simeq \frac{Z_0}{j\beta l} = \frac{1}{j\omega Cl}.$$

The line behaves in this case like a capacitance Cl. The effects resulting from the use of cables having lengths that are a small fraction of a wavelength can be treated solely in terms of inductance Ll and capacitance Cl. This is the standard low-frequency approximation.

When $\beta l = \pi/2$, equation (63) shows that $Z \to \infty$. Thus a shorted line one-quarter wave long appears at the other end as an open circuit. As the frequency is changed so that $\beta l = \omega l(LC)^{1/2}$ increases through $\pi/2$, the phase angle of Z changes from $\pi/2$ to $-\pi/2$ and $|Z|$ goes through a sharp maximum. This is much like the behaviour of a tuned circuit; indeed, in microwave electronics, short-circuited lines are used as substitutes for tuned circuits. Although real lines have resistive loss and this broadens the resonance, the effective quality factor Q is still high. At 3,000 megahertz, when the wavelength is 10 centimetres, a line 2.5 centimetres (one inch) long behaves as a resonant circuit. This example demonstrates the way in which the transmission delay in cables makes microwave circuit analysis quite different from low-frequency ac circuit analysis.

Real transmission lines always have a finite resistance R per unit length. If finite resistance is included in the theory, the equation that replaces (56) is

$$\frac{\partial V}{\partial x} = -L\frac{\partial I}{\partial t} - RI,$$

and the wave, instead of propagating unattenuated as $e^{j(\omega t - \beta x)}$, propagates as $e^{(\omega t - \beta x) - \alpha x}$, in which $\alpha \simeq R/2Z$. The value of R is determined by the so-called skin effect (see below) and increases with frequency as $\omega^{1/2}$. At 100 megahertz—*i.e.*, a frequency used for radio and television transmissions—α is of the order of 3×10^{-3} metre^{-1}. The amplitude of the voltage and current on the line decrease exponentially along the line as $e^{-\alpha x}$ and the power as $e^{-2\alpha x}$. In the laboratory, with a few metres of line, the decrease is not serious, but if, for example $x = 10^4$ metres (six miles), the power is reduced by a ratio $e^{-60} \simeq 10^{-26}$. If 10^6 watts are fed into one end of the line, a power of only 10^{-20} watt emerges at the other end, and this would be barely detectable. Attenuation by the skin effect in cables at radio frequencies, although at first sight small, is a serious bar to their use for distributing radio signals over extended paths. Practical cable-distribution systems rely on amplifiers to boost the signal. These amplifiers have to be placed at frequent intervals in the cable.

It will be seen then that, at fairly low frequencies in which cables and connections are only a fraction of a wavelength long, the time delay in the cable can be more or less neglected and at most an allowance made for the total capacitance and inductance of the cable. At high frequencies, however, the time delay introduces quite new effects. The theory of the propagation of waves on cables is a highly developed branch of electrical engineering, but it has a more general significance. In many ways the behaviour of waves on cables is simpler and better understood than any other form of wave motion. Thus, notions drawn from transmission-line theory, especially the concepts of characteristic impedance and matching, are widely used in other branches of physics, particularly in acoustics and optics.

Importance of transmission line theory

FILTERS

A filter is a circuit network built up by the repetition of a single basic unit. This can be taken to be the T-section shown in Figure 28, which also indicates the currents

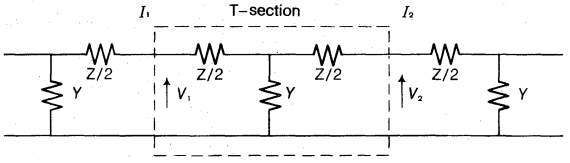


Figure 28: Filter partitioned as T-section (see text).

and voltages I_1, I_2 and V_1, V_2 at the ends of a typical section (Y represents some repeated combination of inductance and capacitance). If Kirchhoff's laws are applied to this T-section, they give the relations

$$V_2 = (1 + \tfrac{1}{2}ZY)V_1 - \tfrac{1}{2}Z(2 + \tfrac{1}{2}ZY)I_1$$

and

$$I_2 = -YV_1 + (1 + \tfrac{1}{2}ZY)I_1.$$

If, as the signal progresses through the chain of filter elements, it is to change regularly by the same amount as it traverses each element, the ratio V/I at the input to each element must be the same. In other words, $V_2 = \lambda V$, and $I_2 = \lambda I_1$ must be the input to the next section, in which λ is a (possibly complex) number. If λ is substituted in the equations for V_2 and I_2, there results, by eliminating V_1 and I_1, a quadratic equation for λ,

$$\lambda^2 - 2(1 + \tfrac{1}{2}ZY)\lambda + 1 = 0.$$

This equation has two roots, λ_1 and λ_2, which satisfy the relations $\lambda_1\lambda_2 = 1$ and

$$\lambda_1 = 1 + \tfrac{1}{2}ZY - \sqrt{\left(1 + \tfrac{1}{2}ZY\right)^2 - 1}\,.$$

One of the roots, say λ_1, corresponds to a signal propagating to the right (Figure 28) through the filter. The other root with λ_2 equals $1/\lambda_1$ gives V_1 equals $\lambda_1 V_2$ and I_1 equals $\lambda_1 I_2$. This corresponds to a signal propagating in the opposite direction.

The nature of the constant λ_1 depends, first, on whether ZY is real and, second, on whether $(1 + \frac{1}{2}ZY)^2$ is greater or less than one. If ZY is real and $(1 + \frac{1}{2}ZY)^2$ is less than one, it is possible to write $1 + \frac{1}{2}ZY = \cos \phi$, and then

$$\lambda_1 = \cos \phi - \sqrt{\cos^2 \phi - 1} = \cos \phi - j \sin \phi = e^{-j\phi}.$$

In this case the signal, as it passes each section, is merely shifted in phase by ϕ and corresponds to a signal in the pass band of the filter. If ZY is real and $(1 + \frac{1}{2}ZY)^2$ is greater than one, λ_1 is real and less than unity; the voltage V_2 and current I_2 will be less than V_1 and I_1 and the signal will be attenuated as it passes through the network. When ZY is complex the root λ_1 is also complex. If it is written as $\lambda_1 = \alpha e^{-j\phi}$, then $\alpha < 1$ and the signal is both attenuated and shifted in phase as it traverses each filter section. Most practical filters have attenuation and so further exact analysis becomes tedious and rather unprofitable. Fortunately, however, most of the properties of filter networks constructed of T-sections, as shown in Figure 29, can be understood by considering four basic

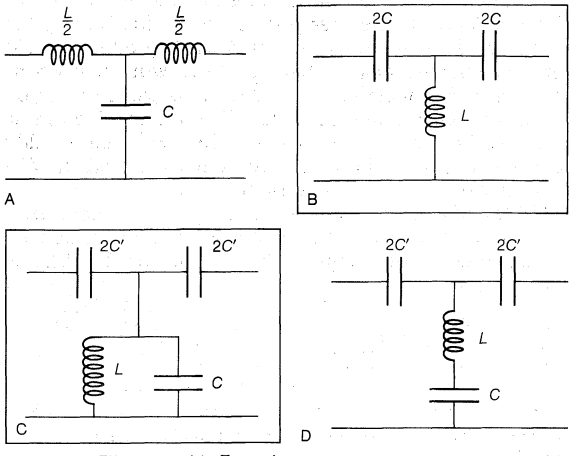


Figure 29: *Filters used in T-sections.*
(A) Low-pass; (B) high-pass; (C) band-pass; (D) band-stop (see text).

idealized types, the low-pass filter 29A, the high-pass filter 29B, the band-pass filter 29C, and the band-stop filter. If the filter is operating in a pass band with a phase shift ϕ per section, it must satisfy $\cos\phi = 1 + \frac{1}{2}ZY$. Because $-1 < \cos\phi < 1$, the condition defining a pass band becomes simply $-4 < ZY < 0$. In the low-pass filter of Figure 29A, in which Z equals $j\omega L$ and Y equals $j\omega C$, the pass band is defined by $-4 < -\omega^2 LC < 0$. Thus the pass band extends from dc up to the angular frequency $(4/LC)^{1/2}$. In the high-pass filter of Figure 29B,

$$Z = \frac{1}{j\omega C}, \quad Y = \frac{1}{j\omega L},$$

and then

$$-4 < -\frac{1}{\omega^2 LC} < 0.$$

This filter therefore passes all frequencies above $1/(4LC)^{1/2}$. For the band-pass filter of Figure 29C, in which

$$Z = \frac{1}{j\omega C'}$$

and $Y = j\omega C + \dfrac{1}{j\omega L}$, the condition becomes

$$\left(\frac{1}{L(C + 4C')}\right)^{1/2} < \omega < \left(\frac{1}{LC}\right)^{1/2}.$$

Thus in this case there is transmission only over a finite range of frequencies. If $C' << C$ and $\omega_0 = (1/LC)^{1/2}$, the pass band extends from $\omega_0(1 - 2C'/C)$ to ω_0, and the width of the pass band is $2\omega_0 C'/C$. The band-pass filter is an exceedingly useful type of network section for picking out wanted signals in a limited range of frequencies from a background of unwanted signals. At microwave frequencies it is quite feasible, using resonant cavities rather than tuned circuits, to design filters that transmit signals in a band width of only five megahertz about a centre of frequency of 10,000 megahertz. At lower radio frequencies resistive loss in the inductance limits the fractional frequency resolution to about 1 percent, but filters using resonant piezoelectric elements can have a much better resolution.

Usefulness of band-pass filter

For the band-stop filter shown in Figure 29D, $Z = \dfrac{1}{j\omega C'}$ and $Y = \dfrac{j\omega C}{1 - \omega^2 LC}$, and this gives transmission only when $\omega > (LC)^{-1/2}\left(\dfrac{C + 4C'}{4C'}\right)^{1/2}$. However, if C' greatly exceeds C, high attenuation only occurs near $\omega = (LC)^{-1/2}$. Ideal band-stop filters require extra circuit elements.

All these filters have one important feature in common, namely, $\cos\phi = \pm 1$ at the edges of the pass band. This means that the phase shift ϕ per section lies in the range $0 < \phi < \pi$. This is a characteristic feature of all periodic structures built up by repetition of identical units and is not confined to electrical filters. It applies equally well to the quantum mechanics of electrons in a crystal lattice, the propagation of sound waves (phonons) in a crystal lattice, and the propagation of electromagnetic waves through the periodic structures used in linear accelerators and microwave tubes.

TRANSIENT PHENOMENA

Usually in discussing ac circuits it is assumed that there are only sinusoidal currents of constant amplitude and frequency. There are, however, many cases in which currents vary in a rather more complicated way. For instance, it may be necessary to know what happens in a circuit, either immediately after the current is switched on or immediately after it is switched off. In radio and electronics there are modulated alternating currents in which amplitude and phase change rapidly, and in some applications the response of circuits to short pulses and sudden changes in an otherwise steady direct current is also required. There is an elegant mathematical technique based on the use of the Laplace transform for dealing with all these transient phenomena, but it will not be employed here; rather there will be a qualitative treatment of what is expected to happen in a number of simple cases. Only at the end of this section is there a brief outline of a more general approach.

Any circuit that contains inductive or capacitative elements is capable of storing electrical energy, and thus, when a circuit is first switched on, a finite time will be taken to acquire this energy. Equally, when the circuit is switched off, it takes time for the stored energy to disappear. If the stored energy is U and the time in which it is dissipated is τ, then the rate of dissipation U/τ can often be much greater than the power handled by the circuit in its steady-state condition. In this case, switching transients may lead to the destruction of the circuit. If the circuit contains both inductance and capacitance, it is also possible for it to resonate, and this resonance can be shock-excited by a switching transient. In high-power circuits this too can lead to destructive effects, and in low-power electronic equipment to spurious responses.

The problem of switching off an electromagnet illustrates most of the features that accompany switching transients. In a typical magnet used in high-field experiments, the windings have an inductance of about 0.02 henry, a resistance of 0.1 ohm, and operate at a current of 3,000 amperes. In addition, the windings have a distributed capacitance of the order of 10^{-7} farad. The magnet is energized by a motor-generator set having a low internal resistance. If the switch is in an open position, the magnet forms a tuned circuit. The behaviour of this circuit is determined by the parameter $(L/CR^2)^{1/2}$, essentially the quality factor Q of the circuit. If Q is greater than $\frac{1}{2}$, the circuit is underdamped and can resonate. In this case $(L/CR^2)^{1/2} \simeq 2{,}200$ and the circuit is highly resonant at a frequency $(LC)^{-1/2}/2\pi \simeq 3{,}000$ hertz. In practice the Q is appreciably less than this because of eddy currents in the metal casing and of skin effect losses, but, nevertheless, Q is of the order of 100.

If the switch is closed and the coil energized with $V_0 = 300$ volts and $I_0 = 3{,}000$ amperes, the energy stored in C is only $\frac{1}{2}CV_0^2 = 10^{-2}$ joule, but the energy $\frac{1}{2}LI_0^2$ stored in the coil is 10^5 joules. If the switch is now opened, it is as though a tuned circuit has been caught at the instant when all the energy is stored in the in-

ductance. Half a cycle later—*i.e.*, in just over one millisecond—all this energy is transferred to the capacitance C. The voltage across C rises to V, giving $\frac{1}{2}CV^2 = \frac{1}{2}LI_0^2 = 10^5$ joules. Because C equals 10^{-7} farad, V rises to about 10^6 volts or, rather, would rise to 10^6 volts if a destructive spark had not already occurred at the open switch or through weak insulation.

Importance of transient effects

The example just given was chosen as an extreme case to illustrate the magnitude of transient effects in high-power systems, but similar if less spectacular effects can also occur whenever current is suddenly interrupted in an inductive circuit. Transients are used in the ignition system of an automobile to generate a spark, but in most cases their effects are harmful rather than useful. In particular, electronic apparatus incorporating transistors (which are easily destroyed by transient over-voltage) has to be carefully designed to avoid high peak voltages being generated by transient effects when switches are closed or opened.

A rather different aspect of transient phenomena is associated with nonsinusoidal wave forms. In electronics, where short pulses, trains of pulses, and triangular or sawtooth wave forms occur, little is gained by attempting to interpret the response of circuits in terms of their response to sinusoidal signals. For example, the basic equations for a circuit in which a voltage V_{in} is applied to a capacitance C through a resistance R are

$$V_{in} = \frac{q}{C} + R\frac{dq}{dt} \text{ and } V_{out} = R\frac{dq}{dt}.$$

If q in this circuit changes relatively little in a time RC, the first equation can be replaced by $V_{in} \simeq q/C$, and so

$$V_{out} = RC\frac{dV_{in}}{dt}.$$

Thus, for slowly varying signals, this circuit gives an output proportional to the time derivative of the input. It can be verified that if the input is a regular train of square pulses in which q changes abruptly (dq/dt is large), the output is a sequence of alternately positive and negative spikes.

In radio, and especially in the electronic reproduction of music, short bursts of more or less sinusoidal signals occur. Thus, for example, a pizzicato double bass note may consist of only 10 complete oscillations of a wave of frequency 30 hertz (duration ⅓ second, a transient signal), whereas a triangle produces a few hundred oscillations at frequencies above five kilohertz. In general, provided that the response of the electronic circuit to sinusoidal signals varies smoothly with frequency in the frequency range of interest, transient signals such as these will be faithfully reproduced. Bad transient response in audio equipment is more likely to be associated with a poor overall response or mechanical resonances in the transducers (*e.g.*, stylus pickup) than with any peculiar features of the ac circuits themselves.

This section ends with a brief résumé of a more general approach (the Laplace transform) to transient phenomena using the theory of complex variables. In a linear system —*e.g.*, an ac bridge, a filter network, or an electronic amplifier driven by an external time-dependent electromotive force, $V(t)$—the application of Kirchhoff's laws to the system leads to a set of simultaneous linear differential equations for the voltages and currents at various points in the network. With a sinusoidal input voltage of the form $Ve^{j\omega t}$, these equations yield the voltage $Ve^{j\omega t}$ at one particular point in the network in the general form

$$v = \frac{f_0 + f_1(j\omega) + f_2(j\omega)^2 + \text{etc.}}{g_0 + g_1(j\omega) + g_2(j\omega)^2 + \ldots g_N(j\omega)^N} V,$$

in which the numerator and the denominator are finite polynomials in $j\omega$ with real coefficients. For a voltage input of a more general form, Ve^{pt}, in which p replaces $j\omega t$, p may be real, complex, or imaginary. When the input equals zero the output v will be zero unless the denominator vanishes. This will occur for N different values of p. Each of these so-called poles of the expression for v corresponds to a natural resonance of the system. If a typical pole is of the form $p = -\alpha + j\omega$, the system has a damped resonance of frequency ω with a decay constant α. The theory of the Laplace transform implies that the transient response of the system is entirely determined once the location of the poles of the response in the complex p plane is determined.

A modest familiarity with Laplace transform usage is necessary in order to follow current progress of work in electronics, feedback network design, and control system engineering.

EDDY CURRENTS AND THE SKIN EFFECT

In a shorted turn around a transformer core, large currents will circulate when the primary is energized by a source of ac. This current is caused by the electromotive force induced by the changing magnetic flux linking the turn, and it is in such a direction as to produce a field opposing the flux change. Similar effects will occur whenever there is a metallic or conducting object in a region of a changing magnetic field. The resulting currents are then known as eddy currents. The existence of eddy currents has many important consequences, and the discussion begins with the most familiar example.

Eddy currents

If a transformer is wound on a solid core, the core itself will act as a shorted turn. For a long, cylindrical core (see Figure 30) of length l and radius a, the flux linking

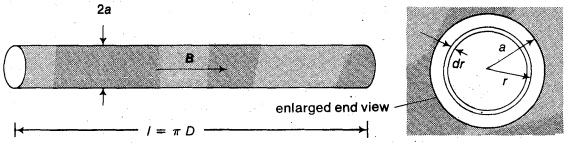


Figure 30: Equivalent cylindrical core; side and end views (see text).

an annular ring of radius r and thickness dr is $\pi r^2 B$, and, if the conductivity of the iron is σ, the current flowing in this ring is

$$dI = \sigma\pi r^2\left(\frac{dB}{dt}\right)\frac{l\,dr}{2\pi r}.$$

The rate at which heat is dissipated in the ring is

$$\frac{2\pi r}{\sigma l\,dr}(dI)^2, \text{ and this is } dQ = \frac{\pi}{2}\sigma l\left(\frac{dB}{dt}\right)^2 r^3 dr.$$

The total rate of dissipation in the core is the integral of dQ, or

$$Q = \frac{\pi}{2}\sigma l\left(\frac{dB}{dt}\right)\int_0^a r^3 dr = \frac{\pi}{8}\sigma l\left(\frac{dB}{dt}\right)^2 a^{24}.$$

The total stored magnetic energy in the core is $\frac{1}{2}\pi a^2 lHB$, in which $H = B/\mu\mu_0$ and μ is the relative permeability. The rate of change of this energy is $\dot{U} = \frac{\pi a^2 l}{\mu\mu_0} B\frac{dB}{dt}$ which is a measure of the rate at which energy flows in and out of the core because of the excitation of the primary winding. The ratio $Q/\dot{U}$ is a measure of the fraction of this energy lost as heat. If ω is the frequency of the ac, then $dB/dt \simeq \omega B$ and $Q/\dot{U} = \frac{1}{8}\omega\mu\mu_0\sigma a^2$.

For iron, $\mu \simeq 500$, $\sigma \simeq 10^7$ mho metre^{-1}, and $\mu_0 = 4\pi \times 10^{-7}$ so that, if $\omega = 2\pi \times 60$ hertz and $a = 10^{-2}$ metre (one centimetre), this gives $Q/\dot{U} \simeq 30$. Since this is greater than unity, all that can be deduced is that virtually all the energy fed into the primary will be dissipated as heat. Not much is left to drive whatever load may be connected to the secondary. This type of yoke or core would be of little use in a transformer. The ratio $Q/\dot{U}$, it should be noted, depends on a^2. Thus, if the core were subdivided into, say, 1,000 insulated iron filaments of much smaller radius, this would also reduce $Q/\dot{U}$ by a factor of 1,000. In fact, it is not only easier but also more effective to build up the core from thin, flat, insulated laminations, as indeed all transformers are. Further improvement is also obtained by using iron alloys of lower electrical conductivity than pure iron.

Next to be considered is a rather similar case using a

nonmagnetic core, a copper tube of radius a and wall thickness z with its axis parallel to a changing magnetic field dB/dt. The current flowing round the tube per unit length of the tube is $-\frac{\sigma z}{2\pi a}\cdot\pi a^2\frac{dB}{dt}$. This produces an additional field $-\frac{1}{2}\mu_0\sigma a z B_0$ within the tube, and this in turn induces a further current and produces yet an additional field. For a field B within the tube and a current I per unit length, $B = B_0 + \mu_0 I$ and $I = \frac{1}{2}\sigma a z\frac{dB}{dt}$.

If the fields vary with time as $e^{j\omega t}$, this leads to

$$B = B_0 - \frac{1}{2}j\omega\mu_0\sigma a z B,$$

so that

$$B = \frac{B_0}{1 + \frac{1}{2}j\omega\mu_0\sigma a z}.$$

The field within the tube lags behind B_0 in phase, and its magnitude is reduced by a factor $(1 + \frac{1}{4}\omega^2\mu_0^2\sigma^2a^2z^2)^{-\frac{1}{2}}$. If, for example, a equals 2.5×10^{-2} metre (one inch) and z equals 2.5×10^{-3} metre, then with $\omega = 2\pi \times 60$ hertz, $\mu_0 = 4\pi \times 10^{-7}$, and $\sigma = 10^8$ metre^{-1}, this gives a reduction factor of 0.8. Thus, even a nonmagnetic copper tube tends to exclude a changing magnetic field.

In this example the field in the tube had to be calculated allowing for the reaction of the field induced by the current. At high frequencies or with thick conductors, this effect eventually alters the field within the conductor itself. This leads immediately to the consideration of the skin effect.

In the section below, *Electromagnetic radiation*, it is shown that in a conductor a wave decreases exponentially in amplitude as it progresses through a conductor. It decreases by a factor $1/e$ (*i.e.*, 0.368) in a distance

$$\delta = \left(\frac{2}{\omega\mu_0\sigma}\right)^{1/2}.$$

This distance is known as the skin depth. All quantities associated with the wave decrease in the same way. Thus the electric and magnetic fields and the current density J all decrease in this way. For copper with $\sigma \simeq 6 \times 10^7$ mho metre^{-1}, at a frequency ν hertz, $\delta \simeq 1/15\nu^{\frac{1}{2}}$ metres. Thus at a radio frequency of one megahertz, the skin depth is approximately 1/15 millimetre, whereas at power frequencies (60 hertz) the skin depth is about 8.5 millimetres.

Skin effect

Alternating currents in conductors are driven by fields applied to their surfaces, and if, therefore, wave fields penetrate only a few skin depths, it is quite obvious that the alternating currents can only flow in a layer a few skin depths thick near the surface. This is generally known as the skin effect. It is a serious problem both in radio-frequency engineering and in power engineering. If an alternating electric field at the surface of a block of metal is E_0 and the current density parallel to E_0 at the surface is $J_0 = \sigma E_0$, then the current density (J) a distance x below the surface is $J = J_0e^{-x/\delta} = \sigma E_0e^{-x/\delta}$, and the total current, per unit length normal to E, is

$$I = \sigma E_0\int_0^\infty e^{-x/\delta}\,dx = \sigma\delta E_0.$$

It is as though there were a current density σE_0 in a layer of thickness δ.

In a thick metal bar of width w and length l, when δ is appreciably less than the thickness z, the resistance will be $R \simeq l/2\omega\delta\sigma$ rather than $l/\omega z\sigma$. Because of the skin effect there is little point in using conductors the thickness of which is much larger than δ. Thus, for example, the bus bars used in a power station are usually built up from strips about six millimetres (¼ inch) thick rather than from more massive copper bars. Equally, there is little point in using round section conductors that have diameters more than about $\frac{1}{2}\delta$. If a lower resistance conductor is required, it is more economical to use a tube of wall thickness $\frac{1}{2}\delta$ or less. At radio frequencies the concern is not so much with economizing on material but with achieving a low resistance at almost any price. In the frequency range up to about 10 megahertz, it is common to use litz wire made up of many fine strands of insulated wire. Above 10 megahertz even this artifice becomes less and less worthwhile, and rather high resistive losses must be tolerated. The resistance per unit length of a round wire of diameter d approaches $1/\pi d\sigma\delta$, compared with the dc resistance $4/\pi d^2\sigma$. The ratio $d/4\delta$ for wire of one millimetre diameter at 100 megahertz is about 40. At still higher frequencies there is a need to keep the wire circumference down to about ¼ wavelength or less. Thus ultimately at microwave frequencies the lowest resistance per unit length that can be achieved is about $4/\lambda\sigma\delta$. This expression can be simplified to $1.7 \times 10^{-2}\lambda^{-3/2}$ ohms per metre for copper. If λ equals 10^{-2} metre it gives 17 ohms per metre. In a coaxial cable of 50 ohms characteristic impedance, this leads to a reduction in the power in a wave by a factor of $1/e$ in only three metres.

In a metal such as iron, in which the relative permeability μ is different from unity, the skin depth δ is $(Z/\omega\mu\mu_0\sigma)^{\frac{1}{2}}$, and as a result the skin depth is further reduced and the resistivity increased. The skin depth in soft iron at 60 hertz is a fraction of a millimetre. This is one of the reasons why electric trains, which use the iron rails as a return conductor, have either to operate on dc power or to use a high-voltage, low-current ac distribution system.

In ac circuits perhaps the most important consequence of the skin effect is the influence it has on the resistance of inductors and thus on the quality factor Q attainable with tuned circuits. It also influences the minimum attenuation that can be achieved in filter networks. There are, however, some benefits that arise from the skin effect; for example, in copper δ is approximately equal to 10^{-2} millimetre at 50 megahertz, and so a thin-walled can of metal is adequate to exclude all unwanted signals of this frequency from the interior of electronic apparatus.

Because the skin depth in iron at power frequencies is only a fraction of a millimetre, at radio frequencies it is so small that it is impracticable to make thin enough laminations for use in transformer cores. In the 1930s, those transformers for use at radio frequencies that were not air-cored used cores made up of iron dust dispersed in an insulating resin or wax. This type of dust core is still in general use, but the postwar developments in ferrites, a new type of insulating ferromagnetic material, have now made available an alternative means of producing cores for radio-frequency transformers. Ferrites are free of both eddy current and skin effects and so can be used even at microwave frequencies.

POWER GENERATION AND DISTRIBUTION

The technology of this subject is covered in the article ELECTRIC POWER, and here only the general principles will be of concern. At its simplest, an electric alternator is a device that rotates a coil of wire in a magnetic field and picks up the induced electromotive force by a system of slip-rings and sliding contacts. It is instructive to consider a generator from this point of view. If a plane loop of wire of area A rotates about the axis, at an angular rate ω in a field B normal to the axis, and if at time (t) equal to zero the loop is normal to B, the flux linking the loop will be $AB\cos\omega t$, and the induced emf is therefore equal to $+\omega AB\sin\omega t$. This emf causes a current I to flow in the load. The current I also flows in the loop, and as a result the loop experiences a couple (Γ), which is easily seen to be equal to $ABI\sin\omega t$. The motor driving the coil must exert an equal and opposite couple to maintain the rotation. It therefore does work on the coil at a rate $\omega\Gamma = \omega ABI\sin\omega t$. But $\omega AB\sin\omega t = V$, and so the rate at which the motor does work is equal to the rate IV at which energy is delivered to the load. The machine is apparently a perfectly efficient device for converting mechanical to electrical power. It may, therefore, reasonably be asked why practical generators are so much more complicated than this and consist of com-

The electric alternator

plex systems of iron armatures with many windings and elaborate iron-cored stator coils. The reasons for this are varied, and only the most important will be mentioned. First of all, no notice has been taken of resistive losses in the rotating loop. As an example, a single turn of copper strip 10 centimetres by one centimetre in cross section is made into a loop one metre square, which then rotates at 60 hertz in a steady field (B) of one tesla (weber per metre^2). The peak induced emf is $2\pi \times 60 \simeq 400$ volts, and if, say, the mean power is 100 megawatts (10^8 watts), the peak current must be about 5×10^5 amperes. The resistance of the loop is, however, about 6×10^{-5} ohm, and so 30 volts are dropped in the loop and something approaching 7.5×10^6 watts dissipated in the loop as heat. Not only has the efficiency dropped to 92 percent but there is a serious cooling problem.

The field B was assumed to have a strength of one tesla. This is a large field, and to produce a field of this order over a region of a cubic metre or so using permanent magnets would be quite impracticable. Thus the field must be produced by an electromagnet. Almost the entire power output of the generator would be required to produce the field, and hence iron must be used to concentrate the field near the windings. This will also help to get more flux through a smaller and lower resistance winding.

The power delivered to the load $IV = V^2/R$ varies with time as $\sin^2 \omega t$. Since the rate of rotation of the generator is constant, the couple exerted by the motor driving the generator must also vary as $\sin^2 \omega t$. Such a reaction would lead to serious mechanical problems, comparable to trying to run a single-cylinder internal-combustion engine at constant speed with almost no flywheel. By mounting several coils at different angles on the generator axle, thus forming an armature for the generator, the load on the motor should be smoothed out. In fact, if there are three coils at 120° to each other ($2\pi/3$), the couple varies as

$$\sin^2\omega t + \sin^2\left(\omega t + \frac{2\pi}{3}\right) + \sin^2\left(\omega t + \frac{4\pi}{3}\right)$$

$$= \frac{1}{2}\left(3 - \cos 2\omega t - \cos\left[2\omega t + \frac{4\pi}{3}\right] - \cos\left[2\omega t + \frac{2\pi}{3}\right]\right),$$

which is seen to be constant and equal to 3/2. Thus, provided the coils are equally loaded, the load on the motor is steady. Placing more coils on the shaft has a further advantage because the field is now used more of the time. There is indeed no reason why there should not be six, nine, or 12 equally spaced coils on the same shaft. Thus practical considerations lead from the original idealized device to something like a conventional power-station generator. The final form of the generator is, of course, further influenced by the electrical, magnetic, and mechanical properties of available materials and mechanical-design considerations.

Power distribution

It will be useful to turn from the problem of generating ac power to the problem of distributing it, setting aside such important and difficult problems as switchgear. With ac, transformers can be used, and it is immaterial whether the power is distributed at high voltage and low current or low voltage and high current. When a power P equal to IV is distributed by a line having resistance R, however, power lost in the line is equal to PIR/V. Thus the higher the voltage and the lower the current, the less power is lost in the line. For this reason power transmission usually employs the highest practicable voltage. This brings other problems to bear—*e.g.*, insulation and breakdown and hazardous operation. In an overhead cable, a compromise has to be struck between the cost of power loss and the cost of higher towers and better insulation. A further factor enters into this calculation. The field at the surface of a circular conductor of radius r at a voltage V is approximately V/r, and, when this ratio exceeds the breakdown strength of air, a discharge is likely to be initiated. There is therefore a minimum radius for the conductors, which increases with V. In practice, conductors are grouped to produce a larger effective radius, but, even so, there comes a point at which further increase in V (although it reduces I) actually increases the cost of the conductors. In the case of long-distance cross-country power lines, carrying large power, the optimum (rms) voltage is in the range 100,000 to 500,000 volts. In metropolitan areas, at lower powers, and in underground installations, the optimum balance may be struck at a much lower figure—*e.g.*, 5,000 volts.

As explained above, if the armature or a generator has three windings arranged at equal angles to each other on the shaft and each supplies a similar load, the load on the motor or prime mover is constant. This type of generator will produce three separate output voltages: $V_0 \cos \omega t$, $V_0 \cos(\omega t + 2\pi/3)$, and $V_0 \cos(\omega t + 4\pi/3)$, and is known as a three-phase generator. The advantages of three-phase generation are by no means confined to the simplification of generator design. A typical three-phase distribution line will have three live conductors carrying currents $I_0 \cos \omega t$, $I_0 \cos(\omega t + 2\pi/3)$, and $I_0 \cos(\omega t + 4\pi/3)$, and a single neutral conductor that will carry a current (in the return direction) I_0 times a factor

$$\left(\cos \omega t + \cos\left[\omega t + \frac{2\pi}{3}\right] + \cos\left[\omega t + \frac{4\pi}{3}\right]\right).$$

As can be seen, this factor is zero. Thus if $R/2$ is the resistance of each conductor, the average power loss is $3(R/2)\ (I_0^2/2) = \frac{3}{4}RI_0^2$ for a generated power $\frac{3}{2}I_0V_0$. The ratio is $\frac{1}{2}RI_0/V_0$. If this is compared with a single-phase system with two conductors of resistance $R/2$, each carrying a current I_0 and dissipating energy $\frac{1}{2}RI_0^2$ for a generated power $\frac{1}{2}I_0V_0$, it will be seen that the fractional loss in a three-phase system is reduced by a factor of $\frac{1}{2}$. Thus the use of three-phase currents not only simplifies generator design but also reduces transmission losses. The ac distribution of electric power almost invariably employs three-phase working. A standard system of domestic power distribution is to provide consumers in one group of houses with a neutral-line and a single-live-phase, whereas houses in other groups are furnished with the neutral line and one or the other of the two remaining live phases. Because the balance of current in the final main neutral line depends on each phase being equally loaded, the groups have to be chosen so that there is a good statistical chance that each group will at any one time be drawing roughly the same current. Large installations are usually supplied with all three phases. It is then desirable that equal loads be presented to each phase. In particular, if one phase has a resistive load; *e.g.*, a heater, it is especially important that another phase should not have a reactive load; *e.g.*, a large transformer running on low load. If the currents in two phases are $I_0 \cos(\omega t + 2\pi/3)$ and $I_0 \cos(\omega t + 4\pi/3)$ and the current in the third phase $I_0 \cos(\omega t + \phi)$ has a large phase angle ϕ, the current in the neutral lead will be of magnitude $2I_0 \sin \frac{1}{2}\phi$. If $\phi = \pi/2$, this is $\sqrt{2}I_0$ and might seriously overload the neutral line.

A schematic drawing of a three-phase generating system is shown in Figure 31A and of a typical load in Figure 31B. This circuit, for obvious reasons, is known as a star-connected or Y-connected system. There is, however, available between the phases a set of voltages

$$V_{12} = V_0\left(\cos \omega t - \cos\left[\omega t + \frac{2\pi}{3}\right]\right)$$

$$= V_0\sqrt{3} \sin\left(\omega t + \frac{\pi}{3}\right);$$

$$V_{23} = V_0\sqrt{3} \sin\left(\omega t - \frac{\pi}{3}\right);$$

$$V_{31} = V_0\sqrt{3} \sin(\omega t + \pi).$$

Delta connections

Another way of connecting a load is shown in Figure 31C. This is known as a delta (Δ) connection. The current in the neutral line is now automatically zero, but, unless $Z_{12} = Z_{23} = Z_{31}$, the currents in the three phases

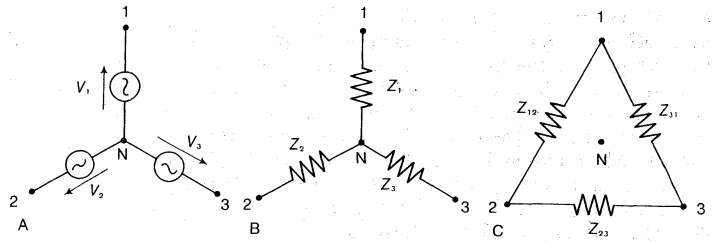


Figure 31: *Three-phase connections.*
(A) Three-phase voltage source; (B) star-connected three-phase load; (C) delta-connected three-phase load (see text).

will be different, although their algebraic sum is still zero. At the generator end one of the phases supplies more power than the others, and this puts an unbalanced load on the generator.

The usual electrical supply uses alternating current, rather than direct current, so that transformers can be used. Thus power transmission over a distance can be made at high voltage, minimizing resistance loss in the cables, and yet the final domestic supply can be reduced to a safe voltage, using step-down transformers. Furthermore, the domestic supply can be further transformed to give voltages appropriate for use in electronic equipment, furnace control gear, and so on. From the consumer's point of view, ac has few disadvantages. Perhaps the most important one is that it is difficult to design large ac motors that will start under load.

The primary disadvantages of ac power transmission are twofold. First of all, the skin effect limits the worthwhile thickness of conductors to about six millimetres (¼ inch), and secondly, the power delivered depends on the root-mean-square voltage, V_{rms}, whereas breakdown and insulation problems depend on the peak voltage $V_p = V_{rms}\sqrt{2} \simeq 1.4V_{rms}$. Thus, for a given power, an ac system must be able to stand some 40 percent higher voltage. Furthermore, in a three-phase system the peak voltage between phases is $\sqrt{3}$ times the peak voltage of each phase measured from neutral; *i.e.*, $\sqrt{6}V_{rms} \simeq 2.5V_{rms}$. There is, therefore, a good reason in considering high voltage dc for long-distance power transmission. High voltage dc transmission is used in Sweden in a main feeder cable that originates at a generating plant in the Arctic and ends in the south of the country. The voltage used is 400,000 volts, which presents no greater insulation problems than 160,000 volts three-phase. The power is generated as ac and then rectified for transmission. The power derived from the line is finally reconverted to ac using solid-state switching devices, so that ac is available for local distribution. The transmission of dc is also used in a number of underwater links—*e.g.*, that between England and France. Here the reason is somewhat different. A submarine cable has, necessarily, to be a closed, insulated structure and, as a result, has a rather high capacitance between conductors, which would produce intolerable switching transients in ac circuits. The transient response is easier to handle in a dc system.

In treating polyphase systems of generation and transmission, attention has been concentrated on three-phase systems because these are not only in common use but also illustrate most of the principles involved. There is, however, no difficulty in winding a generator armature to produce a greater number of phases, and, in some cases, especially if the output is to be rectified, this may be worthwhile.

It is worth considering briefly the reasons that determine the choice of frequency for an ac power supply system. Most systems operate at either 50 or 60 hertz, although there are a few much lower frequency systems —*e.g.*, 15 hertz and, in marine equipment, 600 hertz is common, whereas in some aircraft systems frequencies of 400 hertz and above are used.

VI. The primary effects and properties of electric fields and currents

This section deals with the magnetic effects of currents and electric fields, the way fields store and transport energy and stress, and electromagnetism. The material here mentioned forms the core of classical electromagnetic theory (see also the article on ELECTROMAGNETIC RADIATION).

MAGNETIC EFFECTS OF STEADY ELECTRIC CURRENTS

The magnetic field $\boldsymbol{B}$ caused by a current I flowing in a wire in vacuum (see Figure 32A) can be calculated by summing the contributions from line elements of the wire. The contribution $d\boldsymbol{B}(r_1)$ to the field at position $\boldsymbol{r}_1$ coming from a line element $d\boldsymbol{l}(r_2)$ at $\boldsymbol{r}_2$ is

$$d\boldsymbol{B}(r_1) = \mu_0 \frac{Id\boldsymbol{l}(r_2) \times \boldsymbol{r}_{12}}{4\pi r_{12}^3}. \qquad (64)$$

It obeys an inverse square law and is at right angles to $d\boldsymbol{l}$ and $\boldsymbol{r}_2$. Equation (64) is called the Biot-Savart law,

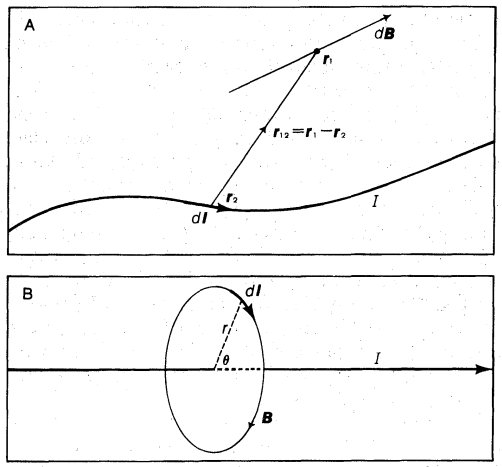


Figure 32: Magnetic fields $d\boldsymbol{B}$ and $\boldsymbol{B}$ arising from current I in a wire. (A) Elemental distance $d\boldsymbol{l}$ along a wire, (B) $d\boldsymbol{l}$ measured along field line around a long, straight wire (see text).

first enunciated by the French physicists Jean-Baptiste Biot and Félix Savart. The force experienced by a line element $d\boldsymbol{l}$ at r of a circuit carrying a current I when the magnetic field $\boldsymbol{B}$ at r is

$$d\boldsymbol{E} = Id\boldsymbol{l} \times \boldsymbol{B} \qquad (65)$$

and is at right angles to the wire element ($d\boldsymbol{l}$) and the magnetic field.

If equation (64) is applied to the calculation of the field that results from a current I flowing in a long straight wire, it shows that the lines of $\boldsymbol{B}$ form closed circular loops about the wire (as shown in Figure 32B) and that the value of the field at a distance r from the wire is $\mu_0 I/2\pi r$. The line integral of $\boldsymbol{B}$ around a closed field-line loop—that is, the integral of the scalar product of $\boldsymbol{B}$ and a loop element $d\boldsymbol{l}$—is

$$\oint \boldsymbol{B} \cdot d\boldsymbol{l} = \int_0^{2\pi} Brd\theta = \int_0^{2\pi} \frac{\mu_0 I}{2\pi r} rd\theta = \mu_0 I.$$

Ampère's law

This result can be generalized to give Ampère's law: the line integral of $(1/\mu_0)\boldsymbol{B}$ around any closed loop in vacuum is equal to the total current threading the loop. If in some region of space there is a current density $\boldsymbol{J}$ and a plane loop defining a plane area is taken the surface element of which is $d\boldsymbol{S}$ (see Figure 32B), this law can be expressed as

$$\oint \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 \int \boldsymbol{J} \cdot d\boldsymbol{S}$$

and, by taking an infinitesimal area, converted to a differential equation involving the curl of $\boldsymbol{B}$. This equation is

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{J}. \qquad (66)$$

All magnetism is due to currents. These may be either currents circulating in wires or currents circulating in atoms. Equation (64) can then be used to show that the

total number of lines of force of the field $\boldsymbol{B}$ emerging from any closed volume is zero. In mathematical terms, if S is a closed surface with dS as a positive outward normal (*i.e.*, a vector perpendicular to the surface and pointing away from the centre), then

$$\oint_S \boldsymbol{B} \cdot d\boldsymbol{S} = \oint_S B_n dS = 0,$$

and this can be expressed in terms of the divergence of $\boldsymbol{B}$ as

$$\nabla \cdot \boldsymbol{B} = 0. \tag{67}$$

Equations (66) and (67) are together equivalent to (64) in the same way that the electrostatic equations

$$\nabla \times \boldsymbol{E} = 0.$$
$$\nabla \cdot \boldsymbol{E} = \rho/\epsilon_0$$

are equivalent to the inverse square law of forces between charges. These electric equations express the following properties of $\boldsymbol{E}$. Lines of electrostatic field do not form closed loops ($\nabla \times \boldsymbol{E} = 0$). Lines of electrostatic field terminate on charges. Equations (66) and (67) express the properties of $\boldsymbol{B}$, which are very different from electric-field properties. Lines of $\boldsymbol{B}$ form closed loops ($\nabla \cdot \boldsymbol{B} = 0$) and do not terminate. The value of $\boldsymbol{B}$ around any closed loop depends on the current linking the loop.

Equations (66) and (67) are still valid for magnetic fields in the presence of matter, provided that the current density ($\boldsymbol{J}$) includes not only currents circulating in circuits or caused by streams of free electrons and ions but also currents circulating within atoms. In practice, just as in electrostatics, it is convenient to distinguish between free, or accessible, charges (ρ_f) and atomic, or bound, charges (ρ_b) by introducing auxiliary vectors—$\boldsymbol{P}$, the polarization, and $\boldsymbol{D}$, displacement—so in magnetism the current density $\boldsymbol{J} = \boldsymbol{J}_f + \boldsymbol{J}_b$ is separated into free, or accessible, currents $\boldsymbol{J}_f$ and bound, or atomic, currents $\boldsymbol{J}_b$. Because the atomic currents are bound in atoms, the net flux of bound current $\boldsymbol{J}_b$ out of any finite volume containing many atoms is zero. Mathematically this is expressed by $\nabla \cdot \boldsymbol{J}_b = 0$, and this makes it possible to express $\boldsymbol{J}_b$ in the form of the curl of another vector ($\boldsymbol{M}$), written $\boldsymbol{J}_b = \nabla \times \boldsymbol{M}$. Ampère's law thus becomes $\nabla \times \boldsymbol{B} = \mu_0(\boldsymbol{J}_f + \nabla \times \boldsymbol{M})$.

Just as the vector $\boldsymbol{P}$ in electrostatics (initially introduced as a way of representing bound charge density by $\rho_b = -\nabla \cdot \boldsymbol{P}$) can be identified with the electric dipole-moment density in a medium, so can the new vector $\boldsymbol{M}$ be identified with the atomic magnetic dipole-moment density. That is, if there are N atoms each with circulating currents resulting in an atomic moment $\boldsymbol{m}$ in unit volume, then $\boldsymbol{M} = N\boldsymbol{m}$.

At this point a new vector ($\boldsymbol{H}$), unfortunately for terminology also known as the magnetic field, is introduced and defined as

$$\boldsymbol{H} = \frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M};$$

This vector satisfies the equation

$$\nabla \times \boldsymbol{H} = \boldsymbol{J}_f; \tag{68}$$

i.e., the line integral of $\boldsymbol{H}$ around a closed path is equal to the linked free, or accessible, current. In vacuum the equivalence $\boldsymbol{H} = \frac{1}{\mu_0}\boldsymbol{B}$ exists. The vector $\boldsymbol{H}$ not only satisfies (68) but also $\nabla \cdot (\boldsymbol{H} + \boldsymbol{M}) = \frac{1}{\mu_0}\nabla \cdot \boldsymbol{B} = 0$, or $\nabla \cdot \boldsymbol{H} = -\nabla \cdot \boldsymbol{M}$. Thus, lines of $\boldsymbol{H}$ are not necessarily closed loops and can terminate as discontinuities in the magnetization $\boldsymbol{M}$ of a body. In some media—*e.g.*, permanent-magnet steels—the magnetization $\boldsymbol{M}$ is mainly determined by the chemical composition and heat treatment of the material. In other cases, $\boldsymbol{M}$ depends on $\boldsymbol{B}$ and therefore also on $\boldsymbol{H}$. The relation between $\boldsymbol{M}$ and the magnetic field is rarely linear (*i.e.*, directly proportional), but for many media over an appreciable range of fields there is an approximately linear relationship between $\boldsymbol{M}$ and $\boldsymbol{B}$ or $\boldsymbol{H}$. For reasons that are mainly matters of custom, this relationship is usually expressed as $\boldsymbol{M} = \chi\boldsymbol{H}$, in which χ is dimensionless (a pure number) and is known as the magnetic susceptibility of the medium. In this case $\boldsymbol{B} = \mu_0(1 + \chi)\boldsymbol{H} = \mu_0\mu\boldsymbol{H}$, and μ is known as the relative permeability or simply as the permeability. Most practical magnetic problems are concerned with cases in which either the dipole-moment density ($\boldsymbol{M}$) has a specific value, as in a permanent magnet, or $\boldsymbol{M}$ is proportional to $\boldsymbol{H}$, and $\boldsymbol{H}$ is generated by true currents in coils in a way specified by equation (68).

Magnetic susceptibility

The force between two accurately constructed circuits carrying the same current (I) can be used to measure the current. This is the basis of the current balance and is the primary measurement used in establishing electrical units and standards.

MAGNETIC EFFECTS OF CHANGING CURRENTS

The theory of magnetism due to steady currents is based on the Biot–Savart law (64), which, together with (65), summarizes the experimental laws found in this branch of magnetism. This law leads to Ampère's law $\nabla \times \boldsymbol{B} = \mu_0\boldsymbol{J}$. It would seem that since the divergence of the curl of any vector is identically zero, it follows directly from this equation that $\nabla \cdot \boldsymbol{J} = 0$—*i.e.*, that the flow of current out of any closed surface must be zero. This conclusion is obviously false. If a charged sphere is placed inside the surface and then pulled through the surface, there has to be a net outflow of charge and a net outward current. Thus Ampère's law can be only an approximation. The generalization of Ampère's law to deal with systems in which charge is accumulating and $\nabla \cdot \boldsymbol{J} \neq 0$ is due to Maxwell.

To see how Maxwell arrived at this generalization, a solid cylindrical conductor with a gap (Figure 33) will be considered, which carries an alternating current $I(t)$

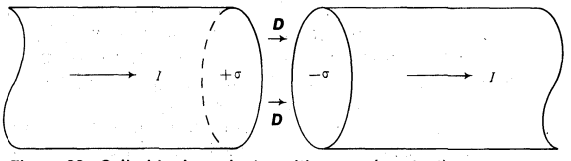


Figure 33: Cylindrical conductor with a gap (see text).

that flows across the gap. There will be a surface charge on the faces of the gap. If I is replaced by the current density J in the conductor and σ is the surface charge per unit area then

$$J = \frac{d\sigma}{dt}.$$

In the gap, lines of what Maxwell called the displacement, $\boldsymbol{D}$, originate on the surface charge on the left face and terminate on the right face and therefore $D = \sigma$. Thus, passing left to right from the conductor through the gap from a region in which $J \neq 0$, $dD/dt = 0$, through a region where $J = 0$, $dD/dt \neq 0$, and then to a region where $J \neq 0$ and $dD/dt = 0$, the sum $J + dD/dt$ remains constant. If the sum $J + dD/dt$ were regarded as the total current density, then $J + dD/dt$ would be a continuous quantity satisfying (in three dimensions) the divergence relation $\nabla \cdot (\boldsymbol{J} + \partial\boldsymbol{D}/\partial t) = 0$; when combined with $\nabla \cdot \boldsymbol{D} = \rho$, however, this equation becomes $\nabla \cdot \boldsymbol{J} + \partial\rho/\partial t = 0$, which can be recognized as the differential statement of the law of charge conservation. The quantity $\partial\boldsymbol{D}/\partial t$ is known as the displacement current (density). To see how Maxwell's hypothesis arises it is useful to consider how the magnetic field lines associated with the divided conductor are obtained by applying Ampère's law. These field lines (the circular lines) are shown in Figure 34; they do not obtrude into the gap. To Maxwell this limitation seemed entirely unreasonable and he therefore hypothesized that displacement current $\partial\boldsymbol{D}/\partial t$ is just as effective as the convection current $\boldsymbol{J}$ in generating a magnetic field. The static form of Ampère's law (66) is therefore to be replaced by

$$\nabla \times \boldsymbol{B} = \mu_0\left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}\right) \tag{69}$$

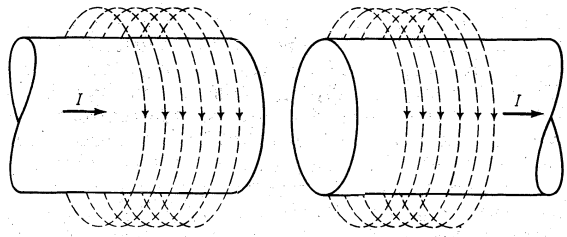


Figure 34: Lines of magnetic force at a gap according to Ampère's law (see text).

as the general law. In vacuum and in the absence of any currents this equation reduces to $\nabla \times \boldsymbol{B} = \mu_0\epsilon_0\partial\boldsymbol{E}/\partial t$, the consequences of which are considered in later sections. It is clear that both currents and changing electric fields generate magnetic effects. It is, however, notable that within a conductor the additional effects due to the displacement current ($\partial\boldsymbol{D}/\partial t$) are almost always negligible. In a medium of conductivity σ, and dielectric constant ϵ, at a frequency ω, the conduction current is $\boldsymbol{J} = \sigma\boldsymbol{E}$, and the order of magnitude of $\partial\boldsymbol{D}/\partial t$ is $\omega\epsilon\epsilon_0\boldsymbol{E}$. The ratio is $(\omega\epsilon\epsilon_0)/\sigma$, and, even in a poor conductor, with $\sigma \simeq 10^6$ mho metre^{-1} and a high value of $\epsilon \simeq 100$, this is only $10^{-15}\omega \simeq 10^{-14}\omega/2\pi$. Thus, at all radio frequencies and lower frequencies, $\partial\boldsymbol{D}/\partial t$ is vanishingly small compared to $\boldsymbol{J}$.

When equation (69) is combined with the other field equations, especially Faraday's law of induction, $\nabla \times \boldsymbol{E} = -\partial\boldsymbol{B}/\partial t$, it shows (see below *Electromagnetic radiation*) that electromagnetic fields propagate with a finite velocity. In vacuum this is $c = (\mu_0\epsilon_0)^{-1/2} \simeq 3 \times 10^8$ metres per second. The two main differences between the fields produced by alternating currents and those produced by steady currents arise from this finite velocity of propagation and from the fact that they can induce potential differences according to Faraday's law.

Velocity of propagation of electromagnetic fields

A disturbance or signal of frequency $\nu = \omega/2\pi$ propagating with a velocity c has a wavelength $\lambda = c/\nu$. In a circuit of linear dimensions (l) comparable with λ there will be an appreciable phase difference between currents in different parts of the circuit because of the finite propagation time. In the opposite case, in which l is much less than λ, the fields near the circuit retain their alternating nature but are not appreciably different from the static fields. The distant fields are quite different and lead to a radiative loss of energy. In the laboratory, circuits are rarely dealt with that have linear dimensions much greater than a few metres, and so the effects associated with the finite velocity of propagation become significant only at high frequencies. For example, even at one megahertz the free space wavelength is 300 metres. At power and audio frequencies, a small fraction of a megahertz, the only new effects associated with alternating, rather than steady, currents are those associated with electromagnetic induction. The most important effects of this type are those that are utilized in transformers and inductors.

In designing a transformer core, efforts are made to minimize energy loss due to eddy currents (circulatory currents induced in metal) in the core, but in other cases use is made of these eddy currents to produce a heating effect. Thus in metallurgy and vacuum-tube manufacture it is often necessary to heat conducting or metallic bodies in vacuum; *e.g.*, in outgassing the anode of a triode. A simple method of doing this is to place a coil carrying radio-frequency alternating current around the assembly and outside the vacuum case (which, through necessity, must not be conducting). The changing magnetic field caused by the current induces eddy currents in the material inside the vacuum case. In general, eddy currents and heating effects occur whenever conducting bodies are placed in the changing magnetic field caused by a changing current.

FORCE, ENERGY, AND POWER ASSOCIATED WITH FIELDS

This section considers one of the most important properties of the electromagnetic field, the way in which it transmits force and energy. The notions of force and energy arise from the study of mechanics, and so it follows that a general treatment of electromagnetic force and energy is based on the equation $\boldsymbol{F} = q\boldsymbol{E} + q\boldsymbol{v} \times \boldsymbol{B}$ for the force ($\boldsymbol{F}$) acting on a particle of charge, q, and velocity, $\boldsymbol{v}$, in an electric field, $\boldsymbol{E}$, and a magnetic field, $\boldsymbol{B}$. The rate at which the two fields do work on the particle is $\boldsymbol{F} \cdot \boldsymbol{v}$ and, therefore, because the magnetic force is at right angles to $\boldsymbol{v}$, only the electric field contributes to the rate of increase of particle energy, which is the product $q\boldsymbol{E} \cdot \boldsymbol{v}$.

For a system of N particles of charge q_k and velocity $\boldsymbol{v}_k$ at positions $\boldsymbol{r}_k$ in a volume V, the total rate at which the fields in V are transferring energy (U) to the particles (p) is $dU_p/dt = \dot{U}_p$; the sum for all N particles is

$$\dot{U}_p = \sum_{k=1}^{N} q_k \boldsymbol{v}_k \cdot \boldsymbol{E}(\boldsymbol{r}_k);$$

and, if the particle distribution is described in terms of a current density $\boldsymbol{J}(\boldsymbol{r})$, this equation can be replaced by the volume integral

$$\dot{U}_p = \int_V \boldsymbol{J} \cdot \boldsymbol{E}\, dV.$$

In a similar way, if $d\boldsymbol{M}_p/dt$ is the momentum of the particles in volume V, it is possible to express $d\boldsymbol{M}_p/dt$ in terms of a charge density ρ and a current density $\boldsymbol{J}$, the volume integral being

$$d\boldsymbol{M}_p/dt = \sum_{k=1}^{N} \boldsymbol{F}_k = \int_V (\rho\boldsymbol{E} + \boldsymbol{J} \times \boldsymbol{B})\, dV.$$

The general results for energy and stress, or force, associated with the fields then follow by combining these equations with the field equations and the laws of conservation of energy and momentum. Though the complete derivation is given in the case of energy, only part of the results are cited in the case of momentum. The reason for this is that the complete momentum formulas involve the use of the stress tensor (a higher order vector), and the result is not sufficiently important here to justify the introduction of a new mathematical technique, tensor analysis.

In vector notation, Faraday's law of electromagnetic induction is $\nabla \times \boldsymbol{E} = -\partial\boldsymbol{B}/\partial t$, and Ampère's law, augmented by the inclusion of Maxwell's displacement current density ($\partial\boldsymbol{D}/\partial t$), is $\nabla \times \boldsymbol{H} = \boldsymbol{J} + \partial\boldsymbol{D}/\partial t$. These two equations can be combined to give the relation

$$\boldsymbol{E} \cdot (\nabla \times \boldsymbol{H}) - \boldsymbol{H} \cdot (\nabla \times \boldsymbol{E}) = \boldsymbol{E} \cdot \boldsymbol{J} + \\ + \boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t} + \boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t},$$

and, if use is made of the formal mathematical identity (symbolized by $\equiv$),

$$\boldsymbol{E} \cdot (\nabla \times \boldsymbol{H}) - \boldsymbol{H} \cdot (\nabla \times \boldsymbol{E}) \equiv -\nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}),$$

then the result will be

$$\nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) + \boldsymbol{E} \cdot \boldsymbol{J} + \boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t} + \boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t} = 0.$$

When this relation is applied to a volume V within a closed surface S with a positive outward normal $\boldsymbol{n}$ and use is made of the relations,

$$\int_V \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) dV = \oint_S (\boldsymbol{E} \times \boldsymbol{H})_n dS$$

and

$$\int_V \boldsymbol{E} \cdot \boldsymbol{J}\, dV = \dot{U}_p \text{ (particle energy)},$$

there results

$$\oint_S (\boldsymbol{E} \times \boldsymbol{H})_n dS + \int_V \boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t}\, dV + \int_V \boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t}\, dV + \dot{U}_p = 0,$$

in which $(\boldsymbol{E} \times \boldsymbol{H})_n$ is the magnitude of the normal component of the vector $\boldsymbol{E} \times \boldsymbol{H}$.

If this result is applied to a volume V, the closed surface S of which coincides with a perfectly conducting metal

wall, then the tangential component of E on the wall vanishes, so that the vector product $E \times H$ has no component normal to S; thus the surface integral vanishes. Further, in the case in which J and D are constant in time, the magnetic field B is constant and therefore $\frac{\partial B}{\partial t} = 0$; hence, only two terms from the above equation remain, namely,

$$\dot{U}_p + \int_V E \cdot \frac{\partial D}{\partial t}\, dV = 0.$$

It then follows, from the conservation of energy, that in this case, the equation $\dot{U}_E = \int_V E \cdot (\partial D/\partial t)\, dV$ must represent the rate of change of the electric-field energy. This result was obtained by a different route in the section *Static electricity*. Next, $\partial D/\partial t$ and J are allowed to vary in time, so that $\partial B/\partial t$ is no longer zero; again, conservation of energy identifies the term

$$\dot{U}_M = \int_V H \cdot \left(\frac{\partial B}{\partial t}\right) dV$$

as the rate of change of magnetic-field energy. Finally, in the most general case in which the surface S is purely a notional mathematical surface, the term $(E \times H)_n$ need not vanish on S. The scalar equation

$$\oint_S (E \times H)_n dS + \dot{U}_p + \dot{U}_E + \dot{U}_M = 0,$$

together with energy conservation, now implies that the surface integral represents the outward flux of energy carried by the field through the surface S. Thus, in a sense, $E \cdot J$, $E \cdot (\partial D/\partial t)$, and $H \cdot (\partial B/\partial t)$ can be interpreted as the rates at which particle, electric-field, and magnetic-field energy densities are increasing and the vector $E \times H$ as the flux of energy, or power flow, per unit area. This vector is usually known as Poynting's vector (after English physicist John Henry Poynting).

Poynting's vector

In a medium with linear constitutive relations—*i.e.*, a medium in which $D = \epsilon\epsilon_0 E$ and $B = \mu\mu_0 H$, with ϵ and μ constant, the two relations $\dot{U}_E = E \cdot (\partial D/\partial t)$ and $\dot{U}_M = H \cdot (\partial B/\partial t)$ can be integrated to give the electric- and magnetic-field energy densities as $\dot{U}_E = \frac{1}{2} E \cdot D$ and $\dot{U}_M = \frac{1}{2} HB$. These results, though often useful, are not as generally valid as the expressions for the time derivatives, which are always correct.

The expressions that arise by considering momentum conservation are altogether more complicated, largely because momentum is a vector, whereas energy is a scalar. There are, however, two results that are relatively simple. In vacuum the pressure p exerted across a surface S with a normal n is, expressed in general form,

$$p = \frac{1}{2}\left(\epsilon_0 E^2 + \frac{1}{\mu_0} B^2\right) - \left(\epsilon_0 E_n^2 + \frac{1}{\mu_0} B_n^2\right),$$

in which E_n and B_n are the normal components of the fields the magnitudes of which are E and B. The momentum g per unit volume carried by the fields in vacuum is

$$g = \frac{1}{\mu_0 c}(E \times B) = \frac{1}{c}(E \times H),$$

in which c is the velocity of light; thus g is equal to Poynting's vector divided by c. If n is taken as the number of photons of energy $h\nu$ (Planck's constant times frequency) that cross unit area per second, the momentum carried across the same area per second is $\frac{nh\nu}{c}$, which can be expressed in terms of the wavelength, λ, as $n\frac{h}{\lambda}$. Thus, a photon of frequency ν and wavelength λ has a momentum h/λ. This is the starting point for the development of quantum mechanics.

Some examples of the use of these results, beginning with Poynting's vector, are the following: A straight wire of circular cross section and radius r, with a resistance R per unit length, has a current I flowing through it. At the surface of the wire there will be an electric field $E = IR$ parallel to the length of the wire and, according to Ampère's law, there will be a magnetic field H parallel to the circumference of the wire, of magnitude $I/2\pi r$. These fields are mutually perpendicular, and Poynting's vector $E \times H$ is directed in towards the centre of the wire. Its magnitude is $I^2R/2\pi r$. But $2\pi r$ is the area of the surface of unit length of the wire, so that the total energy flow per second from the field into unit length of wire is I^2R. The energy flux described by Poynting's vector is, therefore, just equal to the energy dissipation caused by ohmic (resistance) heating in the wire.

This example does no more than confirm that the use of Poynting's vector leads to standard elementary results. In the next derivation, the use of Poynting's vector is essential. If a radio transmitter puts out a total power P evenly distributed over a hemisphere, the energy flux at a distance r is $P/2\pi r^2$ and, if r is reasonably large compared with the wavelength, the radiation is effectively in the form of plane waves. In a plane wave (see below *Electromagnetic radiation*), the fields E and H are perpendicular to each other and to the direction of propagation, so that Poynting's vector is in the direction of propagation and its magnitude is EH. Furthermore, for a plane wave, $H = (\epsilon_0/\mu_0)^{1/2} E$, and so $P/(2\pi r^2) = EH = (\epsilon_0/\mu_0)^{1/2} E^2$. The field strength ($E$) at a distance r is therefore

$$E = \left(\left[\frac{\mu_0}{\epsilon_0}\right]^{1/2} \frac{P}{2\pi r^2}\right)^{1/2}.$$

If, for example, $p = 10^4$ watts and $r = 10^4$ metres, then, since $\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}}$ has a value of approximately 120π ohms, this gives $E \simeq 8 \times 10^{-2}$ volts metre^{-1}. Calculations such as this are essential in determining the signal strength picked up by distant receivers.

Next to be considered is the pressure associated with an electrostatic field in the absence of a magnetic field, or $B = 0$. Thus, from the general equation above, the pressure (p) is

$$p = \frac{1}{2}\epsilon_0(E^2 - 2E_n^2).$$

At a metal surface, the field $E = E_n$ and the field E is zero within the metal. There is, therefore, an unbalanced pressure $p = -\frac{1}{2}\epsilon_0 E_n^2$ acting outwards from the metal surface. Because $\epsilon_0 \simeq 10^{-11}$ farad per metre, this pressure is small, unless E_n is large. Even when E_n is 10^6 volts per metre, the pressure amounts to only five newtons per metre^2, or 5×10^{-5} atmosphere. If a plane electromagnetic wave falls at normal incidence on a metal surface and is reflected, the normal components are $B_n = E_n = 0$, and the pressure exerted on the surface is $\frac{1}{2}\epsilon_0 E^2 + \frac{1}{2\mu_0} B^2$, which is equal to the energy density U just outside the surface. Thus, for normal incidence on the surface the radiation pressure is $p = U$. For incidence at an angle θ to the normal, $p = U \cos^2\theta$, and for thermal radiation, evenly incident from all angles, the average over θ gives $p = \frac{1}{3} U$. This result is the basis for the thermodynamic treatment of radiation. For further details see the article MAGNETISM.

ELECTROMAGNETIC RADIATION

A changing current in a circuit produces a changing magnetic field, and this, in turn, induces a changing electric field. The displacement current represented by the changing electric field leads to a further magnetic field, and the repeated operation of these effects causes an electromagnetic disturbance, or wave, to propagate away from the circuit. The propagation of this wave is analyzed in more detail below; here it is enough to note that it travels with the velocity of light $c = (\mu_0\epsilon_0)^{-1/2}$. Static fields, because of static charges or steady currents, obey an inverse-square law, and so the Poynting vector, or energy-flux vector, $E \times H$, decreases with distance R

Consequence of the displacement current

from the circuit, as $1/R^4$. The integral of $E \times H$ over a closed surface at a distance decreases as $1/R^2$ because the surface of integration is only proportional to R^2 (in fact the integral is zero) and there is no net loss of energy from the circuit to the field. The fall-off is quite different for changing currents and charge distributions, for which the fields have components that decrease only as $1/R$, and, as a result, $E \times H$ is proportional to $1/R^2$, and its integral over a distant closed surface is independent of the distance of the surface. The wave propagating away from the circuit carries with it energy, and the circuit loses energy by radiation.

If a circuit has an alternating current $I \cos \omega t$ (also denoted by the exponential form $Ie^{j\omega t}$, in which $j = \sqrt{-1}$) flowing through it, the radiated power (p) averaged over a cycle will be proportional to I^2, which can be written $P = \frac{1}{2}R_r I^2$. The resistance R_r is known as the radiation resistance of the circuit.

The value of the radiation resistance R_r depends on the frequency ω and the nature, size, and shape of the circuit. If l is the largest linear dimension of the circuit and $\lambda = 2\pi c/\omega$ is the wavelength, the general form of R_r is

$$R_r = \left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}} Gf\left(\frac{l}{\lambda}\right),$$

in which $(\mu_0/\epsilon_0)^{1/2} \simeq 120\pi$ ohms, G is a dimensionless shape factor, and f is a function of $l/\lambda = \omega l/2\pi c$. For example, a circular ring of wire having radius a much less than wavelength λ has a radiation resistance

$$R_r = \frac{\pi}{6}\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}}\left(\frac{2\pi a}{\lambda}\right)^4.$$

In general, R_r is small when the linear dimensions of the circuit are small compared with λ and reaches peak values at about 100 ohms when a linear dimension associated with the current is an integral multiple of $\lambda/2$.

If the circuit has an ohmic resistance R, the fraction of the power supplied to the circuit that is radiated is R_r/R. In the laboratory, most circuits have dimensions that are small compared with a wavelength, and so this ratio is generally negligible, although even a small radiated power may cause instability in electronic equipment or lead to interference with nearby equipment. In antennas the dimensions of the circuit are made comparable with wavelength to ensure a large value of radiation resistance at the same time the ohmic resistance is minimized, so that most of the power supplied to the antenna is radiated, rather than lost as heat.

The radiation from a circuit is not isotropic—that is to say, it is not evenly distributed in all directions, and there is always some direction in which the energy flux is zero. For circuits small compared with a wavelength, the energy-flux distribution almost always follows a $\sin^2 \theta$ law in which θ is the angle measured from the null direction. Antennas for which the linear dimensions are large compared with wavelength have a much more complicated angular distribution of radiation, or polar diagram. One of the main problems in antenna design is to achieve a well-defined beam of radiation in the desired direction. If the antenna dimensions are large compared with a wavelength the problem is essentially the same as the problem of attaining good angular resolution in optics and is subject to the same laws. The minimum angle in which the radiation can be confined is approximately λ/l, in which l is the linear dimension of the antenna structure perpendicular to the beam direction.

There is a direct connection between the properties of a circuit or antenna as a radiator and its properties as a receiving element; for example, the polar diagrams in the two cases are identical and a structure that is an efficient transmitter is also an efficient receiver.

A transmission line of length l, short-circuited at both ends, behaves as a resonant (oscillating) circuit at frequencies in which $l = n\lambda/2$ and n is integral. In general, any more or less linear circuit of length l will also be resonant when $l \simeq n\lambda/2$. At the same time, it will act as an efficient antenna. Thus, if a circuit is shock-excited either by interrupting current or suddenly shorting the circuit, it is likely to excite resonances at just those frequencies that radiate most efficiently. A well-known example of this behaviour is the radio and television interference generated by automobile ignition systems. Because the lengths of wire involved are a few metres, the radiation occurs in just that part of the radio spectrum used for FM radio and television.

These general and qualitative considerations will now be expressed more formally. The first group of results are concerned with the propagation of plane electromagnetic waves. A plane wave is a wave propagating in a definite direction, say the z-axis, the wave fronts, or surfaces of constant phase, being planes normal to the direction of propagation. In a medium of relative permeability μ and dielectric constant ϵ these waves are found to propagate with a velocity

The propagation of plane waves

$$c' = (\mu\mu_0\epsilon\epsilon_0)^{-\frac{1}{2}}. \tag{70}$$

In vacuum this velocity becomes

$$c = (\mu_0\epsilon_0)^{-\frac{1}{2}} \simeq 3 \times 10^8 \text{ metres per second.} \tag{71}$$

Most of the media considered in optics have μ almost exactly equal to unity, and so the refractive index

$$n = \frac{c}{c'} = \epsilon^{\frac{1}{2}}. \tag{72}$$

This relation between the refractive index and the dielectric constant at optical frequencies shows how the dielectric response of the atoms of a material medium influences the propagation of waves.

In a metal of conductivity σ, the wave propagates as

$$A \cos(\omega t - z/\delta)e^{-z/\delta},$$

in which the skin depth is δ—*i.e.*, the distance within which the amplitude decreases by $1/e$ is given by

$$\delta = \left(\frac{2}{\omega\mu_0\sigma}\right)^{\frac{1}{2}}. \tag{73}$$

Equations (70), (71), (72), and (73) refer only to the propagation of waves and not to the behaviour of the fields E and B associated with these waves. The first and most important result is that these fields are at right angles to the direction of propagation and to each other; thus, if the waves propagate along the z axis, E and B lie in the xy plane. Furthermore, if a wave has electric-field components E_x and E_y, it has magnetic-field components $B_x = -1/c\, E_y$ and $B_y = 1/c\, E_x$. Thus Poynting's vector, the energy-flux vector, has a single component:

$$(E \times H)_z = \frac{1}{\mu_0}(E \times B)_z = \left(\frac{\epsilon_0}{\mu_0}\right)^{\frac{1}{2}}(E_x^2 + E_y^2) = \left(\frac{\epsilon_0}{\mu_0}\right)^{\frac{1}{2}} E^2.$$

These results not only give the electromagnetic interpretation of phenomena associated with polarized light (light vibrating in one plane) but also relate the energy-flux vector to the fields.

The second group of results is concerned with calculating the radiation from a circuit or a localized distribution of current. The radiated power is obtained by integrating $E \times H$, or $\frac{1}{\mu_0}E \times B$, over a large sphere or closed surface surrounding the system, and E and B themselves are derived from a quantity called the vector potential A, using the formula

$$B = \nabla \times A$$

and

$$\epsilon_0\mu_0\frac{\partial E}{\partial t} = j\omega\epsilon_0\mu_0 E = \nabla \times B = \nabla \times (\nabla \times A).$$

The vector potential A at a point r_1 and time t_1 is related to the current density J at r_2 at an earlier time $t_2 = t_1 - r_{12}/c$, in which $r_{12} = r_1 - r_2$, by the integral taken over the volume v,

$$A(r_1, t_1) = \int_V \frac{\mu_0 J(r_2, t_2)dV_2}{4\pi r_{12}}. \tag{74}$$

The retarded potential

The volume integral extends over the whole region containing currents. The vector potential A given by equation (74) is known as the retarded potential. The relation between t_1 and t_2 expresses the fact that the effects of a current element $J\,dV$ at r_2 take a time r_{12}/c to propagate to r_1, the value of A determining the fields E and B. In some respects (74) is the most important result of this section. It expresses in an unambiguous fashion the fact that changes in currents at r_2 are not experienced as changing forces on charges and current at r_1 until a time r_{12}/c has elapsed. It is also the basic formula in calculating the fields or radiation caused by a prescribed set of currents in an antenna or circuit. Except for simple systems, it leads to exceedingly complicated expressions. A further general result is that, if J corresponds to an alternating current varying as $e^{j\omega t}$ in a system of finite dimensions (located near the origin of the coordinate system), the part of A at a large distance R that contributes to radiation can be expressed as

$$A(R, t) = \frac{\mu_0 e^{j\left(\omega t - \frac{\omega R}{c}\right)}}{4\pi R} F\left(\omega, \frac{R}{R}\right).$$

Here, R is the magnitude of R, and $F(\omega, R/R)$ is a vector function of the frequency and the direction (but not magnitude) of R relative to some fixed direction related to the current distribution in the source. The form of the function F depends on the dimensions and shape of the source, as well as the strength of the currents. If, for example, the source is an oscillating electric dipole of moment $pe^{j\omega t}$, then F is simply equal to $j\omega p$; but, for more general distributions, F is likely to be a formidably complicated function. An oscillating electric dipole is equivalent to a current $Ie^{j\omega p}$ flowing in a short straight length of wire specified by the vector l, in which $Il = dp/dt = j\omega p$.

This formula can be used to calculate the power radiated when a short length l of wire is exited at its centre with a current $Ie^{j\omega t}$, which charges the stray capacitance, between the ends of the wire. The result for the power radiated into a solid angle ($d\Omega$) in a direction making an angle θ with the axis of the wire is

$$dP = \frac{\omega^2 \mu_0 l^2}{32\pi^2} \sin^2\theta \, d\Omega,$$

and the total power radiated into all angles is

$$P = \frac{1}{2} I^2 \left\{ \frac{1}{6\pi} \left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}} \left(\frac{\omega l}{c}\right)^2 \right\}.$$

The radiation resistance is therefore

$$R_r = \frac{1}{6\pi} \left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}} \left(\frac{\omega l}{c}\right)^2 \simeq 20 \left(\frac{2\pi l}{\lambda}\right)^2 \text{ ohms},$$

in which λ is the wavelength. This calculation is only valid if $2\pi l << \lambda$, and a more elaborate calculation shows that for a uniform straight wire, R_r has a maximum value of 73.2 ohms when $l = \lambda/2$. For larger values of l, the resistance decreases and then rises to a second maximum of 201 ohms when $l = \lambda$. There are further, and larger, maxima when $l = \frac{3\lambda}{2}$, 2λ, etc.

Calculations of the radiation from circuits are appreciably more complicated than for the simple dipole. In practical antenna design, recourse usually has to be made to approximate methods. The basic formula (74) also applies to optical radiation from atoms and gamma radiation from nuclei. In these cases, although the quantum nature of the radiating system introduces some complications, the problem is essentially much simpler because the dimensions of the radiating system are much smaller than the wavelength involved. If the origin is taken somewhere near or within the source region, r_1, can be substituted for r_{12} in the denominator of Equation (74). In this case the integral can be expanded in a power series in r_2; *i.e.*, in terms of the linear dimensions of the system. If these dimensions are denoted by a, the radiation specified by the nth term in the expansion is proportional to $(a/\lambda)^{-n}$ and, since this decreases rapidly with n, only the first nonvanishing term is usually significant. In the optical (light) spectra of atoms, only the electric dipole term with $n = 1$ is usually of any significance. If this term vanishes because of symmetry, the corresponding spectral line will be extremely weak or absent.

Maxwell's field equations provide a complete description of the wave aspects of radiation at, as far as known, all frequencies from dc up to 10^{20} Hz or more but, as the frequency is increased, the quantum of energy $h\nu$ associated with a wave becomes more and more significant. Although, with care, quantum effects can be observed at radio frequencies, at optical and higher frequencies the quantal aspects become more easily observable. At the frequencies associated with high energy reactions, the most obvious phenomena are quantal. It is, however, possible to observe wave effects. For example, even with one MeV gamma rays of frequency 2.5×10^{20} Hz, the polarization of the rays can be measured using a technique not essentially different from that applied to radio waves at 10^9 Hz. A more expansive treatment of this subject may be found in the article ELECTROMAGNETIC RADIATION.

VII. The effects of electricity on matter

PIEZOELECTRICITY

In the direct piezoelectric effect, stress (force) S applied to a crystal produces a polarization P proportional to S; and in the converse effect, an applied field E produces a strain (distortion) σ proportional to E. The change in external conditions of a crystal, that is, the work done by external agencies, is the sum of the stress energy and the strain energy, or $-PdE + Sd\sigma$; and so, neglecting thermal effects, this is also the change in the internal energy U. The change in a function (H), when defined as $H = U - S\sigma$, is therefore

$$dH = dU - Sd\sigma - \sigma dS = -PdE - \sigma dS.$$

Because H is a definite function of the state of the crystal, this gives

$$\frac{\partial P}{\partial S} = -\frac{\partial^2 H}{\partial S \partial E} = -\frac{\partial^2 H}{\partial E \partial S} = \frac{\partial \sigma}{\partial E}.$$

Setting the two partial differentials for P and σ equal to the same coefficient (d)—*i.e.*, $\partial P/\partial S = \partial\sigma/\partial E = d$—they can be integrated to give the basic piezoelectric equations

$$P = dS$$
$$\sigma = dE.$$

Materials that display a strong direct effect also display a strong converse effect and vice versa.

It is sometimes convenient to relate P to the strain and the stress S to the electric field E. In this case

$$P = dK\sigma \qquad (75)$$
$$S = -dKE, \qquad (76)$$

in which K is the elastic modulus, a constant equal to the ratio of stress to strain. In (76) S represents the external stress needed to keep the strain (σ) constant when E is applied; thus if the external stress is actually zero, the relation is $-S/K = dE = \sigma$ by equation (75).

Piezoelectricity arises, in molecular terms, from the displacement of ions in a crystal lattice, which clearly couples the displacement of charge associated with P to the mechanical change in the lattice shape or dimensions associated with the strain. The electric fields that bind ions in position in a crystal lattice are of the order of 10^{10} volts per metre (one volt per angstrom) and a field E might be expected to produce a fractional change in their position; *i.e.*, a strain, of the order of $10^{-10}E$. In other words, the coefficient d in equation (75) will be about 10^{-10} metre per volt. This is, in fact, the value for one important material ADP (ammonium dihydrogen phosphate), but most crystals—*e.g.*, quartz with $d = 3 \times 10^{-12}$ metre per volt—have smaller coefficients because geometrical factors associated with the lattice tend to cancel the piezoelectric effect.

The strain σ in a deformed elastic body is a measure of its fractional change in dimensions and shape. The stress S is a measure of the internal forces transmitted across a plane in the body and mathematically both σ and S are second-rank tensors (a kind of higher order vector; see below). This makes the treatment of piezoelectricity quite complicated. The tensor nature of σ and S and also the vector nature of E and P have so far been ignored in order to keep the equations as simple as possible.

One consequence of the nature of the variables is of considerable importance. If, in an undeformed body, two points A and B have a vector separation r with Cartesian components x, y, and z, and in the deformed body their separation is $r + s$, in which s has components u, v, and w, the strain tensor σ involves nine ratios u/x, u/y, u/z; v/x, v/y, v/z; and w/x, w/y, w/z. The components u/x, v/y, w/z describe simple extensions, whereas the mixed components—*e.g.*, u/y—describe shear strains. Similarly, if the force per unit area acting across a plane with a positive normal n (with components $[k, l, m]$) is p (with components $[q, r, s]$), the stress tensor involves nine ratios q/k, q/l, q/m; r/k, r/l, r/m; and s/k, s/l, s/m. The components q/k, r/l, and s/m correspond to the extensive stress, and the mixed components—*e.g.*, q/l—correspond to shear. Because σ and S are defined in terms of the ratios of the components of two vectors, an inversion of the coordinate system, which takes x to $-x$, y to $-y$, and z to $-z$, does not affect the components of σ or S, but, in a relation such as $\sigma = d\,E$, the components of E change sign. Thus, under inversion, the coefficient d must also change sign. In a gas, a liquid, an amorphous solid, or a crystal, each with a centre of inversion symmetry, both coordinate systems before and after inversion are completely equivalent and d must be the same in both systems. Since now the transformed coefficient $d' = d$ and also $d' = -d$, this means that $d' = d = 0$. Piezoelectricity is therefore confined to acentric crystals; *i.e.*, crystals with no centre of inversion symmetry. Furthermore, if the crystal has almost perfect inversion symmetry, the coefficient d will be small, and in general crystal symmetry will exert a profound influence on the allowed or nonvanishing components of the piezoelectric coefficient d, which, mathematically, is a third-rank tensor with $27 = 3^3$ components.

Limits of piezoelectricity

Although the piezoelectric coefficient d is small, generally less than 10^{-10} metre per volt, the elastic modulus K is large, generally greater than 10^{10} newton per square metre and so, in the relation $P = d\,K\sigma$, the product $d\,K$ is about one coulomb per square metre. An infinitesimal strain will therefore produce an appreciable polarization, and if the crystal is equipped with electrodes, as shown in Figure 35, will lead to a large induced potential of the order of $P/\epsilon\epsilon_0$, in which ϵ and ϵ_0 have the same meaning as before. If a charge q is placed on, say, the upper electrode, it will be related to the displacement vector D in the crystal and D' on the upper face of the electrode; but usually D' will be small and can be neglected, and therefore, it follows that $q = AD$, in which A is the area of the electrode. In the crystal, D contains a component $-P$ resulting from the piezoelectric effect produced by strain and a component $\epsilon_0\epsilon\phi/x$ resulting from normal dielectric polarization; hence,

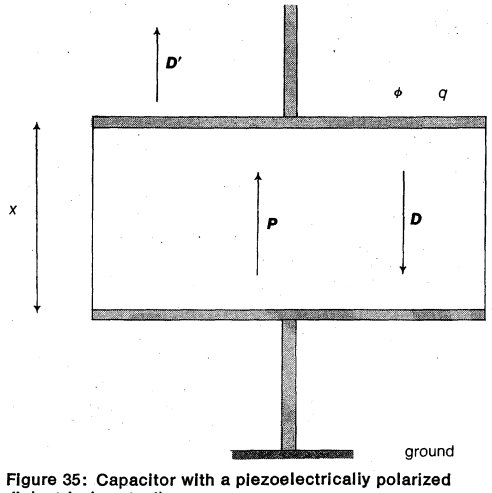


Figure 35: Capacitor with a piezoelectrically polarized dielectric (see text).

$$q = \frac{\epsilon\epsilon_0\phi A}{x} - PA.$$

If the upper electrode is isolated and uncharged so that the charge q is zero, the induced potential is

$$\phi = \frac{x}{\epsilon\epsilon_0}P = \frac{x}{\epsilon\epsilon_0}Kd\,\sigma.$$

But $x\sigma$ is of the order of magnitude of the displacement δ corresponding to the strain σ and so $\phi = \frac{K\,d}{\epsilon\epsilon_0}\delta$.

If, for example, values appropriate to the compound *ADP* are inserted into the last equation—*i.e.*, $K \simeq 10^{10}$, $d \simeq 10^{-10}$, $\epsilon \simeq 50$—then, since $\epsilon_0 \simeq 10^{-11}$, this gives $\phi \simeq 2.10^9\delta$, and a displacement of only five angstroms (5×10^{-10} metre) is required to generate a volt. A displacement of 10^{-4} metre (0.1 millimetre) produces 200 kilovolts. The direct piezoelectric effect is thus very substantial and, for example, even the small displacements of a phonograph stylus attached to a piezoelectric crystal can be easily converted to appreciable voltages.

The converse piezoelectric effect is, by contrast, small. Thus, in a crystal one centimetre thick with 10 kilovolts applied across the crystal, giving a field of 10^6 volts per metre, the strain is only 10^{-4} and the displacement 10^{-4} centimetre. There are a few applications of the static converse piezoelectric effect (*e.g.*, in lasers to alter the mirror spacing), but, in general, most practical applications involve alternating fields at a frequency coinciding with a natural mechanical resonance of the crystal. In this case, the strain for a given field is increased by the quality factor Q of the mechanical resonance. In quartz, because Q usually exceeds 10^4 and can be as high as 10^6, the effective piezoelectric coefficient is increased from 3×10^{-12} metre per volt to about 10^{-6} metre per volt. The resonant quartz crystal can be used as an efficient transducer for converting electrical energy at high frequencies to ultrasonic waves. Because the coupling between the crystal and sound waves in the surrounding medium increases as the square of the frequency, however, the technique is of practical importance only at frequencies above the audible range. The radiated ultrasonic power for a given strain also depends on the acoustic impedance of the medium, and since the maximum strain is limited by the strength of the crystal, not only the efficiency but also the maximum power are only appreciable at high frequencies and in relatively dense media such as liquids.

Pyroelectricity, the appearance of a polarization P proportional to a change in temperature T, can be partially attributed to piezoelectric effects accompanying thermal expansions, but there is also an intrinsic effect; for a change in T also alters the interionic forces and can lead directly to a change in P. Because T is a scalar, the coefficient α in the basic equation

Pyroelectricity

$$P = \alpha T \qquad (77)$$

must be a vector and will necessarily vanish if the medium has a centre of symmetry. Indeed, since equation (77) associates a vector with the crystal, the crystal must not only be acentric but must also belong to one of the polar symmetry classes with a unique axial direction. The effect is generally small, but in a few cases—*e.g.*, triglycine sulfate—the pyroelectric induced voltage can be used to detect thermal radiation. Special camera devices, which detect hot bodies by the bodies' own radiation, have recently been constructed using a pyroelectric screen.

The piezoelectric strain in an applied field leads to a change in refractive index and an indirect electro-optic effect (see next section) but the effect is of little interest.

Finally, it may also be noted that many common materials of scientific interest—*e.g.*, urea, sodium chlorate, and cadmium sulfide—are piezoelectric, and spurious weak resonances that occur in sophisticated electric instruments, such as nuclear quadrupole resonance spectrometers, can often be attributed to piezoelectric effects.

ELECTRO-OPTIC PHENOMENA

The electric field that binds electrons in atoms is more than 10^{10} volts per metre and so the displacement that accompanies the application of an electric field of even 10^6 volts per metre is small and nearly proportional to the applied field. This, as noted above, leads to a linear relation $P = \epsilon_0 \chi E$ between the polarization P in a dielectric and the field E. The relation is not, however, exactly linear and in treating either the effect of strong fields or fine details of the dielectric response it is necessary to include extra terms in the relation between P and E; *e.g.*,

$$P = \epsilon_0(\chi E + \beta E^2 + \gamma E^3 + \text{etc.}). \qquad (78)$$

In solids, the linear coefficient χ is not far from unity and it is reasonable to expect that the nonlinear coefficients β and γ are of the order of $1/E_a \simeq 10^{-10}$ metre per volt and $1/E_a^2 \simeq 10^{-20}$ metre^2 volt^{-2}, in which E_a is the field acting on the electrons in an atom. If a field $E \cos \omega t$ is applied at an optical frequency and a dc field E_0, then the polarization P will contain the following terms,

$$P = \epsilon_0(\chi + 2\beta E_0 + 3\gamma E_0^2) E \cos \omega t;$$

thus the medium will behave as though its optical susceptibility has been changed from χ to

$$\chi' = \chi + 2\beta E_0 + 3\gamma E_0^2.$$

Because refractive index n is related to χ by the relation $n^2 = 1 + \chi$, there will be a change in the refractive index n dependent on dc field. The approximate change δn in n from using the term $2\beta E_0$ is $\delta n = \frac{\beta E_0}{n}$; and if $\beta = 0$ the change made by using the term $3\gamma E_0^2$ is $\delta n = 3\gamma E_0^2/n$. When a coordinate system is inverted, the components of P and E both change signs, and so the coefficient β from equation (78) must also change sign. It follows, as it did for the piezoelectric coefficient d, that β must be zero except in acentric crystals. Thus, in liquids and glasses the only electro-optic effect is that represented by the cubic term γ. The change in n is then quadratic in E_0 and is a so-called quadratic electro-optic effect. Except in certain liquids that exhibit an anomalously large effect, the Kerr effect, to be introduced below, the quadratic effect is small and unimportant. In acentric crystals in which β is not zero, the change in n is linear in E_0 and there is a linear electro-optic effect, named after F. Pockels, a German physicist.

The Pockels effect

If $\beta \simeq 10^{-10}$ metre per volt and $n \simeq 1.5$, then in a field of 1.5×10^6 volts per metre the change δn in n is $\delta n \simeq 10^{-4}$; thus in a crystal one centimetre long traversed by light of wavelength 5,000 angstroms, the change in optical length of the crystal is two wavelengths. Because a change in optical length of half a wavelength is sufficient to produce substantial changes in the optical properties of the system it can be seen that, even though β is small, the linear electro-optic effect is of practical importance. It can be used to modulate beams of light in optical experiments, in optical communication systems, and in optical data-processing systems.

The magnitude of the nonlinear coefficient β that leads to the linear electro-optic effect depends not only on the acentricity of the crystal lattice and the molecular constituents of the crystal but also on the ease with which the atomic electrons are displaced by an applied field. An important empirical rule, which also has a theoretical basis, states that the electro-optic coefficient is proportional to the static dielectric susceptibility and to the square of the optical susceptibility. Thus, in terms of n and the static dielectric constant ϵ, this gives

$$\beta \simeq (n^2 - 1)^2 (\epsilon - 1) \Delta,$$

in which the quantity Δ is a constant. The value of Δ varies by a factor of about 10 from one material to another but, because n can vary from 1.4 to 2.8, and ϵ from three to 3,000, the value of $(n^2 - 1)^2 (\epsilon - 1)$ varies from two to 1.5×10^4, and, as a result, only materials with high values of n or ϵ have appreciable linear electro-optic coefficients. A useful measure of the strength of the electro-optic effect is the half-wave voltage $V_{\frac{1}{2}}$. This is the voltage that when applied across a cube-shaped crystal produces an optical path change of half a wavelength. Its value is $V_{\frac{1}{2}} = n\lambda/2\beta$, and at $\lambda = 5{,}000$ angstroms it varies from a few hundred volts in some titanates and niobates to several hundred thousand volts in, for example, the potassium–magnesium compound $K_2Mg_2(SO_4)_3$. In KDP (potassium dihydrogen phosphate) it is 7,500 V. KDP is an important electro-optic material because it is readily available in large crystals of optical quality and has low dielectric loss even at microwave frequencies. It can be used to construct optical shutters that operate at frequencies up to at least 10^{10} hertz.

The quadratic electro-optic effect, which alone can occur in centrosymmetric crystals, glasses, and liquids is ordinarily too small to be of practical importance. If the coefficient $\gamma \simeq 10^{-20}$ metre^2 volt^{-2}, the field required to produce a change of one part per million in the refractive index is 10^7 volts per metre, which is verging on the limit of practicability. There is one exception to the coefficient being so small. Some liquids (nitrobenzene is an important example) have molecules that are not only highly acentric but also have a large permanent electric dipole moment p. The linear electro-optic effect of the molecules is neutralized by their random orientation, but in a field E at a temperature T the molecules are partially oriented, to an extent pE/kT, and there is then a partial linear electro-optic effect. The net effect is still quadratic in E because the orientation is proportional to E as well. The effective quadratic coefficient that results is approximately $\gamma \simeq \beta p/kT$, in which β would represent the effect for a completely oriented system. This Kerr effect leads to values of γ some thousand times larger than normal. The effect is still small compared with the linear electro-optic or Pockels effect in a good crystal but, because it occurs in a liquid, there is virtually no practical limit to the length of optical path that can be employed. Optical shutters based on the Kerr effect are used in the laboratory and were also employed in early experimental television systems.

In quantum mechanical terms, both the linear optical susceptibility and the nonlinear Pockels and quadratic electro-optic effect are associated with a change in the energy of the atomic ground state in an electric field. A quadratic energy change with field, for example, leads to the linear term χ. These changes in the energies of atomic states also lead to an electric field-induced shift in atomic spectral lines known as the Stark effect (discovered in 1913 by Johannes Stark, a German physicist). The line shift is usually quadratic in the field and is mainly of significance to the spectroscopist because it leads to a broadening of spectral lines from a gas discharge, the broadening being a consequence of strong random electric fields that occur in the discharge. The Stark effect on the first excited state of the hydrogen atom is, however, linear. The transition from the ground state ($S_{\frac{1}{2}}$) to the first excited state ($2P_{\frac{1}{2}}$) is the strongest line in the hydrogen absorption spectrum. The $2P_{\frac{1}{2}}$ level, however, coincides almost exactly in energy with the $2S_{\frac{1}{2}}$ state and the interaction between these two states leads to an anomalously large Stark effect that, except at low fields, is linear with the strength of the electric field. Linear Stark effects can also be observed in molecular spectra, but generally the Stark effect is of much less significance than its counterpart, the Zeeman effect, in which spectral lines are affected by magnetic fields.

The Stark effect

THERMAL EFFECTS

The most widespread and important thermal effect of electricity is ohmic or resistance heating. Electrons moving in a conductor pick up energy from an applied field and by collisions excite random atomic vibrations in the

lattice of the conductor and thus increase its temperature. If the applied field and current are increased, the vibrations may become so violent that they disrupt the lattice and the conductor melts or evaporates. This principle is used in the fuses that protect electrical circuits; a low-melting-point wire is used so that the circuit will be interrupted before the temperature in the wire has become high enough to constitute a fire hazard.

The conversion of electrical to thermal energy in a resistance is complete and so, when electrical heating is employed, all the energy supplied can be utilized. It is this total conversion that makes it possible to measure potential (volts) and current (amperes) by resistance heating. The ampere is defined in absolute terms and the volt is related to the value of the permittivity in vacuum, ϵ_0, obtained from the velocity of light and the comparison of a resistance with a standard capacitance using this value of ϵ_0 and an absolute unit of time. Secondary standards of resistance and voltage of high accuracy are readily available and so the joule ohmic heating in a resistance R, voltage squared divided by resistance (V^2/R), is normally used as the primary thermal standard in physics, chemistry, and engineering.

Reversible thermoelectric effects in conductors are dealt with above, and so only the thermal effects of fields in insulators will be considered here. Irreversible thermal effects occur whenever an alternating electric field is applied to a dielectric the polarization of which lags behind the applied field by a time comparable with the period of the field, or when the period of the field is comparable with the period of the natural electronic or ionic resonance in the medium. The effect is negligible at radio frequencies in quartz and polyethylene, but in some media—*e.g.*, porcelain or polyvinyl chloride plastics—it is appreciable. Dielectric heating can be usefully employed to cure and weld many plastics.

Dielectric media, mainly those comprised of molecules having permanent dipole moments, which exhibit a temperature-dependent susceptibility, show reversible thermal effects in static or slowly varying fields. At higher frequencies, these effects lead to irreversible heating and dissipation. The effect is particularly strong in water at microwave frequencies and is a main contributor to the heating in a microwave oven. Dead organic matter (*i.e.*, food) nearly always contains appreciable amounts of water.

The highest power densities that can be achieved by ohmic or dielectric heating are limited by the properties of solid and liquid materials. Much higher densities can, however, be achieved in plasmas and electron beams. For example, an electron beam of 50 kilovolt electrons carrying one ampere can be focussed onto an area of one square millimetre giving 5×10^{10} watts per square metre. Because thermal radiation at 10,000° K can produce only 6×10^8 watts per square metre; *i.e.*, about 1 percent of this energy density, electron beams (and plasma jets) are preferable for welding, machining, and cutting refractory materials.

Finally, it may be remarked that the heating of a solid body by incident thermal radiation is ultimately caused by the effect of the electric field of the radiation on the electrons or ions in the body.

THE ELECTRICAL PRODUCTION OF LIGHT

For the two most important practical light sources, the incandescent filament lamp and the gas discharge tube, see text article on LIGHTING AND LIGHTING DEVICES. This section is concerned with rather less familiar effects.

Accelerated or decelerated electrons radiate, and if an energetic electron travelling in vacuum is suddenly brought to rest on impact with a solid electrode, it will produce continuous electromagnetic radiation called bremsstrahlung radiation. Bremsstrahlung radiation is not a practical source of light because the effect occurs only with relatively high energy electrons and the resultant radiation lies in the X-ray region. It is, however, one of the two main processes, the other being the excitation of resonance X-radiation from target atoms in the operation of X-ray tubes (see X-RAYS).

Electrons moving in circular orbits in a high-energy machine such as a cyclotron, betatron, or synchrotron are subject to an inward acceleration and therefore radiate. Although the radiation is of some interest as a source to the spectroscopist, its chief importance is that it limits the energy to which electrons can be accelerated in a cyclic machine. Eventually, a point is reached at which the radiation loss just balances the energy input.

Electrons moving with a uniform velocity v (*i.e.*, in a straight line without acceleration) in vacuum do not radiate because v is always less than the velocity of light. If, however, energetic electrons pass through a medium of refractive index n in which the velocity of light c/n would be less than v, they produce Cherenkov radiation (named after Pavel Alekseyevich Cherenkov, a Soviet physicist) at an angle θ to the electron trajectory, given by $\cos\theta = c/nv$. The fraction of the electron energy radiated is small and is confined to wavelengths in which the refractive index n exceeds c/v. Cherenkov radiation is used in high-energy particle physics to record the passage of fast-charged particles.

Cherenkov radiation

Electrons of more than about one kilovolt energy impinging from vacuum on almost any dielectric will excite target atoms and cause some radiation. Usually, the radiation is reabsorbed in the target medium and little escapes, but if a dielectric with the correct energy level structure is used, reabsorption can be eliminated and appreciable efficiency of energy conversion to light, of the order of several percent, can be achieved. Furthermore, the energy level structure of the medium determines the spectral distribution or colour of the light. This is the principle used in the screen phosphors of the cathode-ray tube and the television picture tube. It is also used in nuclear physics to detect charged particles, especially electrons. Electrons impinging on some dielectrics, notably potassium chloride (KCl), can also produce light-absorbing centres and render a screen opaque.

Conduction electrons moving in a solid under the influence of an electric field usually lose kinetic energy in low-energy collisions as fast as they acquire it from the field. Under certain circumstances in semiconductors, however, they can acquire enough energy between collisions to excite atoms in the next collision and cause radiation. A necessary condition for this to occur is that there must be regions, in the semiconductor, of intense electric field. If this field is not to cause destructively large currents to flow, the region must be one in which there are very few mobile electrons (or holes). Such a region occurs in zinc sulfide, a large energy-gap material, and a voltage applied across a thin layer of zinc sulfide powder causes an electroluminescent effect, the detailed mechanism of which is not understood and the efficiency of which is low. At present, electroluminescent panels are of more interest as signal indicators and display devices than as a source of general illumination. A somewhat similar effect occurs at the junction in a reverse-biassed semiconductor, *p-n*-junction diode, *i.e.*, a *p–n*-junction diode in which the applied potential is of a sense opposed to that which produces a substantial current. Electrons in the intense field at the depleted junction easily acquire enough energy to excite atoms. Little of this energy finally emerges as light, although the effect is easily visible under the microscope.

When a junction between heavily doped *n*-type material and less doped *p*-type material is forward biassed, the current consists mainly of electrons injected from the *n* material into the conduction band of the *p*-type material. These electrons ultimately drop into holes in the valence band and release energy equal to the energy gap of the material. In most cases this energy (E) is dissipated as heat, but in gallium phosphide, and especially in gallium arsenide, an appreciable fraction appears as radiation the frequency ν of which satisfies the relation $h\nu = E$. In gallium arsenide, although up to 30 percent of the input electrical energy is available as radiation, the wavelength, 9,000 angstroms, is in the infrared. Gallium phosphide gives visible light but with low efficiency, and alloys have intermediate properties. At present, injection diodes are mainly used as indicator lamps, but, at least in principle,

there is no reason why materials with worthwhile efficiencies in the visible should not be discovered. Gallium nitride, for example, has recently been shown to emit blue-green light, though with low efficiency. The electroluminescent injection diode is also used as a laser, and many proposed optical communication systems are based on this device. Compared with a gas laser or an optically pumped laser, it has the advantage that it can be directly modulated at high frequencies.

Diode laser

CHEMICAL EFFECTS

The direct effects of electric fields on chemical reactions are, on the whole, negligible, and the most important electrochemical effect, electrolysis, involves the motion of charged ions, constituting an electric current. Electrolysis is considered in detail in the article ELECTROCHEMICAL REACTIONS, and in this article only the basic principles are presented. Many substances—*e.g.*, sodium chloride (NaCl), hydrogen sulfate (H_2SO_4), and potassium hydroxide (KOH)—when dissolved in water or other polar solvents, dissociate into positive and negative ions. Water itself is in any case slightly dissociated, to the extent of one part in 10^9, into H^+ and OH^- ions. Under the influence of an electric field, these ions drift in opposite directions, and, for example, in an aqueous solution of H_2SO_4, the positive H^+ ions will accumulate at the negative electrode and the SO_4^{2-} ions at the positive electrode. What happens at the electrodes then depends to a great extent on the material of the electrodes. If the positive electrode is made of a reactive metal, metal ions will go into solution, whereas if it is an unreactive metal such as platinum, the excessive concentration of SO_4^{2-} ions will suppress the hydrogen-ion concentration and cause more water to dissociate to H^+ and OH^-. The OH^- ions are then neutralized at the electrode and the reaction $4OH^- \rightarrow 2H_2O + O_2$ releases oxygen gas. The details of the reactions at the electrodes are not important, but it is noteworthy that the chemical species released is not necessarily related to the ions transported toward the electrode.

Electrolytic conduction differs from electronic in two important ways. First, the charge carriers are massive ions rather than light electrons and the mobilities are low, less than 10^{-4} of the mobility of conduction electrons in a metal. Second, the concentration of charge carriers in aqueous solutions is usually low. As a result, electrolytes are rather poor conductors and electrolytic conduction is primarily of interest for its chemical consequences. It can be applied either to deposit films of metal as in electroplating or in the separation of an element; *e.g.*, nickel from its salts.

However complex the reactions are that occur at the electrodes, electrolytic processes are subject to the laws of thermodynamics, and a number of important conclusions, especially about chemical reaction constants, can be derived from electrochemical measurements. From the point of view of electrical theory, however, the most important results that emerge are Faraday's laws. These laws state that the quantity of a substance released at an electrode is proportional to the total charge passed and that it is also proportional to the chemical equivalent (combining rates, relative to hydrogen) of the substance. The interpretation of these laws in terms of a fundamental unit of atomic, ionic, or electronic charge has influenced the whole subsequent development of physics.

Electrolytic conduction is not confined to aqueous solutions. Many molten salts and solutions of salts in other molten salts are good electrolytes and this is the basis of industrial processes; *e.g.*, the separation of aluminum. Some solids are also electrolytic conductors, especially at elevated temperatures. Both the mobility of the ions and their concentration are usually low and, regarded as a conduction process, the effect is not very substantial. It can, however, also be regarded as a failure in an otherwise good insulation and, in some cases, the products of the electrolytic process can be a serious nuisance.

Although in electrolysis the centre of interest is in the chemical substances released at the electrodes, in some cases, the transport of the ions themselves is important and also subsidiary effects associated with ionic transport—*e.g.*, electro-osmosis and electrophoresis. Before these subjects can be considered, it is necessary to investigate ionic motion in somewhat greater detail.

Ionic motion

The first thing that is noticeable in a table of ionic mobilities in aqueous solution is that ions differing considerably in mass and size have much the same mobility. Thus, for example, the ions Cl^-, Br^-, and I^- have almost identical mobilities although it might be expected that the heavy and bulky I^- ion would have a much lower mobility than the lighter and smaller Cl^- ion. Even more surprisingly, the small and light lithium (Li^+) ion has only about half the mobility of the heavy cesium (Cs^+) ion. This is generally attributed to the fact that ions collect a group of water molecules around themselves, which move as a unit with the ion. This is to be expected because a water molecule with its permanent electric dipole moment is strongly attracted toward the region near an ion, where the electric field is strongest. Further evidence for this view comes from the study of the colours and optical spectra of transition group ions in solution. For example, it is well established that the purple colour of potassium chromium alum crystals is caused by the influence on the chromium (Cr^{3+}) ions of a coordinate octahedron of molecules of water of crystallization. Since this colour persists in aqueous solution, the coordinated octahedron must be retained in solution. Further, if the solution is heated (and the octahedron destroyed), the ions take up a new configuration that gives a green solution. Crystals of chrome alum cannot be obtained from this green solution.

If ions can attach to themselves a considerable number of water molecules, it is obvious that ions can also be attached to solid surfaces; *e.g.*, the walls of the container or material particles suspended in solution. For example, manganese dioxide (MnO_2) particles, as a fine suspension, form a suitable substrate for sulfate (SO_4^{2-}) ions in sulfuric acid solution. A single particle of manganese dioxide (MnO_2) cannot, however, hold many SO_4^{2-} ions, simply because by so doing it would acquire a large negative charge and repel the approach of fresh ions. What actually happens is that the particle acquires almost exactly equal numbers of SO_4^{2-} ions and pairs of H^+ ions. The SO_4^{2-} ions are attached to the particle and the H^+ ions form a neutralizing sheath outside the layer of SO_4^{2-} ions. Thus, the particle is enclosed in an electrical double layer, with its positive surface outward, and complete H_2SO_4 molecules are abstracted from the solution. Many of the otherwise puzzling phenomena that accompany the flow of currents in electrolytes and particle suspensions involve the capacity of ions to attach themselves to solid surfaces or to attach molecules of the solvent to themselves.

If a porous pot is immersed in a beaker of water with two electrodes, one inside the pot and the other outside it in the beaker, a potential difference between the electrodes causes water to flow into the pot, raising the water level and causing a pressure difference. This phenomenon is known as electro-osmosis or electro-endosmosis. It is to be noted that a porous pot is used and not, as in normal osmotic pressure experiments, a semipermeable membrane. The flow is caused by surface effects on the walls of the pores of the pot. The basic process can be understood as follows. Figure 36 shows

Electro-osmosis

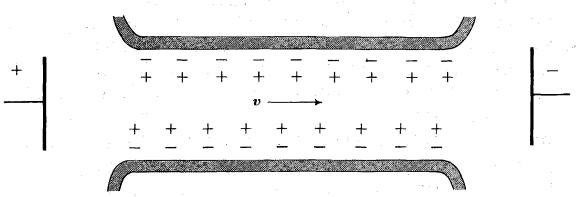


Figure 36: Electro-osmosis in a single pore (see text).

a capillary pore with an electrical double layer of charge adsorbed on its surface. If the liquid is water, the negative ions that are bound to the surface are hydroxyl OH^- ions and the outer positive layer consists of H^+ ions. The

field gradient, due to the electrodes, causes these H^+ ions to glide over the OH^- ions toward the negative electrode and this sweeps the fluid in the centre of the tube along in the same direction with a velocity v for, as the ions move, they make collisions with the non-ionized water molecules of the liquid and therefore transfer momenta to them.

A porous pot is generally made of refractory metal oxides and silicates—*e.g.*, aluminum oxide (Al_2O_3)—which in contact with water tends to form the base $Al(OH)_3$; thus, in this case, it can be seen why a double layer is formed with the OH^- ions relatively firmly attached to the wall. If, on the other hand, the pot is replaced by a porous plug of an electronegative element such as sulfur, this will tend to bind the H^+ ions, leaving the OH^- ions as the outer mobile layer. The flow is then reversed with respect to the potential gradient.

Electrolytes added to the water generally slow down the electro-osmosis. A potassium salt added to the water replaces some of the H^+ ions in the outer layer by K^+ ions and, if these are more tightly bound to the underlying OH^- layer, the flow for a given field is reduced. When the surface layer on the wall is negative, as in Figure 36, the reduction of the flow is determined mainly by the positive ions of the solute and increases rapidly with their valency or positive charge. Thus the barium (Ba^{2+}) ion is more effective than K^+, and a thorium salt containing the thorium (Th^{4+}) ion exerts a very strong effect. The effect is further increased if the positive ion has, in any case, a tendency to adsorb on the wall. This is especially the case with organic ions, which therefore exert an exceptionally strong effect.

The converse effect to electro-osmosis is the production of a streaming potential when liquid is forced through a tube. If in Figure 36 the water is imagined to flow through the tube to the left, caused by pressure gradient applied across the ends of the tube, the positive ions are swept to the left and the electrode marked + in the figure acquires a positive charge and potential. This potential builds up, as ions accumulate near the electrode, until the normal electrolytic conduction current through the tube, in the reverse direction, just balances the rate of arrival of ionic charge. The potential difference (ϕ) between the two electrodes then maintains a steady value, which can be related to the pressure difference (p) between the ends of the capillary.

The result is $\phi = \epsilon\epsilon_0\zeta p/\eta K$, in which η is the viscosity of the liquid, K its electrical conductivity, and ζ the potential difference across the double layer on the surface. Experimentally, streaming potentials of the order of one volt are found to occur for one atmosphere pressure difference, and the values of ζ deduced from the measurements are of the order of 50 millivolts.

Electrophoresis is in many ways similar to electro-osmosis. If small particles of matter are suspended in water in which two electrodes are immersed, the particles are observed to move toward one of the electrodes. The effect is due to the electrical double layer of ions at the surface of the particles. For example, if ferric hydroxide, $Fe(OH)_3$, is suspended in water, there will be some ferric Fe^{3+} ions in solution and these are attracted toward the particles and form the innermost layer of ions. The outer layer consists of OH^- ions. In an electric field there is a net force on the particle with its sheath of iron (Fe^{3+}) ions and this attracts the particle toward the negative electrode. The loosely held OH^- ions slide over the surface of the particle as the particle moves toward the negative electrode.

If σ is surface charge on a particle of superficial area A and u is its velocity, the magnitude of the force is $AE\sigma$ in a field E. The viscous retarding force is of the order of $u\eta A/\lambda$, in which λ is the layer thickness. Thus $u = \lambda\sigma E/\eta$, and again this can be expressed in terms of the potential difference ζ across the double layer as

$$u = \frac{\epsilon\epsilon_0\zeta}{\eta} E.$$

Just as in electro-osmosis, the addition of electrolytes has a strong effect on the motion. In a suspension of gold particles in pure water, for example, the layer of ions next to the gold surface consists of negative OH^- ions. The addition of metal salts containing, for example, aluminum (Al^{3+}) ions leads to a definite reduction or even to a complete reversal of the sign of the charge in this layer.

Different particles in suspension move with different velocities in the same field and this can be used—*e.g.*, in protein chemistry—to effect a separation of the contents of the suspension. The procedure can, for example, be used to separate the constituents of a suspension applied to a strip of filter paper by placing electrodes in contact with the paper. This technique is known as paper electrophoresis.

Paper electrophoresis

BIOELECTRIC EFFECTS

The configuration of enzymes and their biochemical activity depends on the hydrogen ion concentration, and, on a molecular level, this is an electric effect. At the other extreme, large currents destroy tissue by heating and electrolysis. This section deals with an intermediate range of effects associated with nerve, brain, and muscle action, in which small, but macroscopic, potential differences occur and can be influenced by applied potentials. Electric shock is closely related to these processes. For more detail about the biological aspects of electric currents, the reader should consult the article on BIOELECTRICITY.

A beginning can be made by considering how macroscopic potential differences arise in living matter. Biological function depends on chemical reactions in aqueous solution and most of these reactions involve ions. In many cases, they occur at membranes that separate solutions of different ionic composition. If the membrane is permeable to one or more of the ions present in different concentrations on its two sides, the diffusion of these ions down the concentration gradient will lead to charge separation and establish a potential difference across the membrane. In equilibrium, the ionic flows caused by the potential and by the concentration gradients just balance.

If, for one type of ion of charge q, the two concentrations are n_1 and n_2, the potentials on the two sides of the membrane satisfy

$$\frac{n_1}{n_2} = \exp\left(-\frac{q[\phi_1 - \phi_2]}{kT}\right), \qquad (79)$$

in which k is Boltzmann's constant and T is the temperature. This can be expressed as

$$\phi_1 - \phi_2 = -\frac{kT}{q}\log_e\frac{n_1}{n_2};$$

or, changing from $\log_e$ to $\log_{10}$, inserting numerical values for k and T, and setting $q = e$ (the electronic charge), (79) can be expressed as

$$\phi_1 - \phi_2 = -57 \times 10^{-3}\log_{10}\frac{n_1}{n_2}\text{ volts.}$$

The concentration ratios that occur in biological material are usually in the range one to 100 and so the general order of magnitude of these potential differences is less than 100 millivolts.

Usually, the concentrations will finally equalize themselves unless either external influences (*e.g.*, pressure differentials) act on the system or external sources of ions maintain the two concentrations at different values. It is possible, however, even in passive inorganic systems, for potentials of this kind to develop, as will now be seen.

In this example (see Figure 37), a membrane, permeable to sodium (Na^+) ions, chlorine (Cl^-) ions, and water molecules, separates a solution of sodium chloride (NaCl) from a mixed solution of NaCl and NaX, in which X is some large organic ion to which the membrane is impermeable. If $[Na^+]_1$ represents the ionic concentration in region (1) then, the application of equation (79) to the oppositely charged Na^+ and Cl^- ions gives $[Na^+]_1[Cl^-]_1 = [Na^+]_2[Cl^-]_2$. But, because both of the solutions are neutral, $[Na^+]_1 =$

$[Cl^-]_1 + [X^-]_1$, and $[Na^+]_2 = [Cl^-]_2$; from these equations, the ratio of concentrations follows:

$$\frac{[Na^+]_1}{[Na^+]_2} = \left(\frac{[X^-]_1 + [Cl^-]_1}{[Cl^-]_1}\right)^{1/2}.$$

Thus, in equilibrium, there will be a potential difference

$$\phi_1 - \phi_2 = -\frac{57}{2}\,10^{-3}\log_{10}\left(\frac{[X^-]_1 + [Cl^-]_1}{[Cl^-]_1}\right) \text{volts}.$$

The existence of this type of equilibrium, and the Donnan potential (named after Frederick George Donnan, a British chemist) to which it gives rise, follows directly

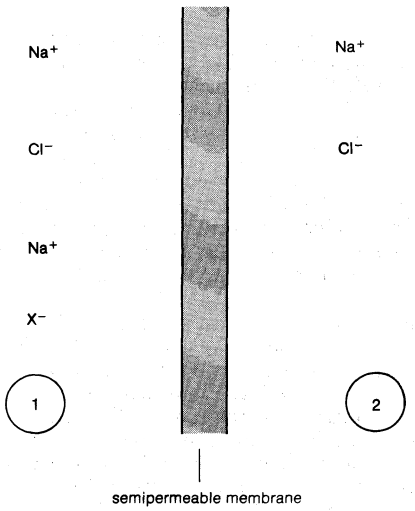


Figure 37: Donnan equilibrium (see text).

from thermodynamic laws, and the potential difference is maintained with no expenditure of energy. In biological systems, however, there is no true thermodynamic equilibrium, and concentration differences leading to potential differences can also be maintained by active processes. Thus, for example, it is possible for a membrane to be permeable to one ion, say the Na^+ ion, only in one direction, provided that energy is continuously supplied to the membrane by biochemical reactions deriving the energy from metabolic processes. Something of this kind occurs in nerves.

Permeability in membranes

A nerve fibre is a tube with an insulating semipermeable wall. The fibre contains in its interior a solution containing predominantly K^+ and negative organic ions together with much smaller quantities of Na^+, Cl^-, and other inorganic ions. The interior fluid is therefore an electrolytic conductor. The fibre is bathed in an exterior fluid containing predominantly Na^+ and Cl^- ions with a much lower concentration of K^+ ions. In some respects, the system resembles a coaxial cable, and, because it is found experimentally that nerve impulses are electrical impulses that advance along a nerve, it is tempting to regard the system as analogous to a coaxial cable in radio. There is, however, one major difference. In a radio cable, the conductors have a low resistance and the propagation of the signals is almost entirely determined by the capacitance (C) and the inductance per unit length. In a nerve, the resistance (R) per unit length is high and the effects of inductance are negligible. This leads to a diffusion equation. A pulse introduced into a nerve fibre is distorted and attenuated to zero within a few millimetres.

If V is the voltage difference between the inner and outer fluids and I the axial current in the inner fluid, the voltage drop per unit length (dV/dx) is RI, whereas the current leakage is $C\frac{dV}{dt}$. This gives the two equations

$$\frac{dV}{dx} = -RI$$

and

$$\frac{dI}{dx} = -C\frac{dV}{dt},$$

which lead to

$$\frac{d^2V}{dx^2} = RC\frac{dV}{dt}.$$

If at $x = 0$ a sinusoidal signal $V_0 \cos \omega t$ is introduced, the signal at position x is

$$V(x,t) = V_0 \cos\left(\omega t - \left[\frac{\omega CR}{2}\right]^{1/2} x\right) e^{-(\frac{1}{2}\omega CR)^{\frac{1}{2}}x}$$

and the phase velocity is $(2\omega/CR)^{1/2}$ which depends on ω. Thus, if a short pulse is introduced into the nerve, the high frequency components of the pulse propagate fastest and are attenuated most quickly. This destroys the shape of the pulse, and, finally, in a distance of the order of $(\tau/CR)^{1/2}$ in which τ is the pulse length, the entire pulse disappears. Because nerve pulses apparently propagate unchanged for considerable distances, this analogy to a coaxial cable is not valid. For a pulse to be maintained, there must be a continuously acting energy source replenishing the pulse energy lost by ohmic dissipation in R.

To return to the electrochemical conditions in a nerve fibre, in the resting condition there is a membrane separating an interior solution, rich in potassium (K^+) ions, from an exterior solution, rich in sodium (Na^+) ions. This condition cannot be a static equilibrium if the membrane is systematically permeable to both ions, for the potential difference associated with the K^+ ions is of the opposite sign to that required to balance the Na^+ concentration gradient. Experimentally, it is found that the inner fluid is about 60 millivolts negative with respect to the outer fluid. This potential difference is almost enough to balance the K^+ concentration gradient, and, further, it varies with the external K^+ ion concentration but not with the Na^+ concentration. It therefore appears that, in the resting condition, the membrane is permeable to K^+ but not to Na^+. This is not, however, sufficient to explain why the potential difference exists, for eventually the K^+ concentration would equalize on the two sides of the membrane. The potential difference is in fact maintained by a continuous, energy-consuming process involving primarily the Na^+ ions. The membrane is not simply passively impermeable to Na^+ but actively pumps Na^+ ions outward. This pumping action not only has to compete with the concentration gradient but also with the potential gradient, and it establishes a slight deficit of positive ions in the interior fluid. This gives it its negative charge and potential. Because the electronic charge is so large, the deficit in positive ions needed is, in chemical terms, small. It is about one part in 10^6 of the total ionic content of the interior fluid.

The resting state

In the resting state, the interior of the nerve has a negative potential of about -60 millivolts; K^+ ions are diffusing relatively freely, though against the potential gradient, through the membrane, and the sodium pump is running. When a nerve impulse (which is a positive potential excursion of the inner fluid) arrives, the impulse, for reasons that are not completely understood, alters the permeability of the membrane and allows the ingress of Na^+ ions down their concentration gradient. This carries the internal potential to a still higher positive potential. The number of ions required is so small that it hardly alters the ionic concentrations. At this stage, the membrane suddenly becomes impermeable to Na^+ ions again, and a net efflux of K^+ ions begins (which returns the inner potential to a negative value) and the sodium pump then brings the nerve back to its resting state. The actual details of these events, which take place during a few milliseconds, are perhaps rather more complicated than this description suggests, but from an electrical point of view, the important consequence is that the energy stored electrically and in concentration gradients by the sodium pump is released and thus augments the energy of the nerve impulse. The mechanism is almost exactly analogous to the operation of a repeater station in a radio pulse-code modulation system. The energy

stored in the repeater is consequently released to form a new pulse in a process that is triggered by the arrival of an input pulse.

In live animals, the initial nerve impulse is generally triggered by chemical action and then propagates along the nerve to a terminal where it triggers further chemical action, thus resulting in a physiological response. Nerve impulses, impulses associated with brain function and impulses associated with muscular contraction can, however, be observed and recorded electrically and also induced by externally applied potentials. Using refined experimental techniques, it is possible to excite single nerves and to produce specific physiological responses, but the simple application of a potential difference across a region of live tissue is more likely to produce a confused response caused by the random stimulation of several nerves at once. Thus, an electric shock in which an appreciable current flows through the body may produce local potential gradients that stimulate either sensory nerves, producing a variety of subjective sensations, or motor nerves, producing uncontrolled muscular contractions. Even if the experimental techniques are crude, however, the experimental electrical stimulation of either muscular tissue, or nerves attached to muscular tissue, is most likely to evoke the typical muscular response; *i.e.*, contraction. This, presumably, was the case in the pioneer observations of twitching frog legs made by the Italian physiologist, Luigi Galvani, which not only laid the foundations of bioelectricity but also influenced the whole subsequent development of practical and theoretical electricity and magnetism.

Attention has been concentrated on nerve transmission because this is the best understood, and most obviously electrical, phenomenon involving electrical effects on living matter. Similar effects occur in muscular tissue, but, on the whole, other biological processes, though controlled by ionic movement and absorption, do not lead to readily observable potential differences and are not affected in any direct way by experimentally applied fields.

BIBLIOGRAPHY

The development of concepts and theories: MICHAEL FARADAY, *Experimental Researches in Electricity*, 3 vol. (1839–55; reprinted in 2 vol., 1965), provides a lucid, immediately comprehensible description of the greatest discoveries and the most important concepts. Also highly readable is EDMUND T. WHITTAKER, *A History of the Theories of Aether and Electricity*, 2 vol. (1910 and later editions), a work of great erudition. HENDRIK A. LORENTZ, *The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat* (1909; 2nd ed., 1953), is the classic exposition of the atomic approach to electricity. Although partially outdated by subsequent developments a remarkably large part of the book is still relevant. A more recent work is LEON ROSENFELD, *Theory of Electrons* (1951). ABDUS SALAM, "Elementary Particles," *Contemp. Phys.*, 1:337–354 (1960), is a readable introduction to the latest ideas. Two advanced texts on charged particles are F. ROHRLICH, *Classical Charged Particles* (1965); and RICHARD P. FEYNMAN, *Quantum Electrodynamics* (1961). WALLACE R. HARPER, *Contact and Frictional Electrification* (1967), gives a modern account of the phenomenon with which the whole science of electricity began.

Basic mathematics: On vectors, BERNARD HAGUE, *An Introduction to Vector Analysis for Physicists and Engineers*, 6th ed. rev. by DANIEL MARTIN (1970), is a good short treatment; DAVID E. RUTHERFORD, *Vector Methods, Applied to Differential Geometry, Mechanics, and Potential Theory*, 9th ed. (1957), is somewhat formal. Several general mathematical texts have suitable sections on vectors. A good example is HENRY MARGENAU and GEORGE M. MURPHY, *The Mathematics of Physics and Chemistry*, 2nd ed. (1956; vol. 2, 1964).

Basic aspects of electricity and magnetism: ARTHUR F. KIP, *Fundamentals of Electricity and Magnetism*, 2nd ed. (1969), is a straightforward text at about the same level as this article. MAX ABRAHAM, *Theorie der Elektrizität*, 8th ed. (1930; Eng. trans. rev. by RICHARD BECKER, *The Classical Theory of Electricity and Magnetism*, 1932 and subsequent reprints), is a rather formal text but not difficult. The treatment of vector methods is notably thorough and the mathematical approach is somewhat more rigorous than usual. ROBERT V. LANGMUIR, *Electromagnetic Fields and Waves* (1961), is a slightly more advanced text, concise but clear. JOHN D. JACKSON, *Classical Electrodynamics* (1962), is a thorough advanced text. It is particularly good on topics with applications in atomic and nuclear physics. An alternate text of similar scope is WOLFGANG K.H. PANOFSKY and MELBA PHILLIPS, *Classical Electricity and Magnetism*, 2nd ed. (1962). ROBERT S. ELLIOTT, *Electromagnetics* (1966), another advanced text, begins each chapter with a colourful, historical survey. SIMON RAMO, JOHN R. WHINNERY, and THEODORE VAN DUZER, *Fields and Waves in Communication Electronics* (1965), is particularly suitable for those whose interests incline more to radio. L.D. LANDAU and E.M. LIFSHITZ, *Electrodynamics of Continuous Media* (1960; orig. pub. in Russian, 1957), is an advanced text and unique in its coverage. It deals authoritatively with innumerable topics associated with conducting and dielectric media. A.S. KOMPANEETS, *Theoretical Physics*, 2nd ed. corr. (1962; Eng. trans. from 2nd Russian ed., 1957; reissued with new introduction as *Physics in the Soviet Union*, 1963), contains a section on electrodynamics that makes a good introduction to more advanced topics. The following books are more specialized: HERBERT FROHLICH, *Theory of Dielectrics*, 2nd ed. (1958), a useful introduction and not too advanced; JOHN H. VAN VLECK, *The Theory of Electric and Magnetic Susceptibilities* (1932, reprinted 1965), still the most important advanced text; HELEN D. MEGAW, *Ferroelectricity in Crystals* (1957), a compact introduction; FRANCO JONA and G. SHIRANE, *Ferroelectric Crystals* (1962), relatively more advanced; N. BLOEMBERGEN, *Nonlinear Optics* (1965), a difficult text on a subject of current interest; and J.D. CRAGGS and J.M. MEEK, *High Voltage Laboratory Technique* (1954), a useful, straightforward introduction to the subject. The remaining books are not recommended as texts so much as sources of information. They all contain solved problems in electromagnetic theory of considerable use in practical and theoretical work. JAMES CLERK MAXWELL, *A Treatise on Electricity and Magnetism*, 2 vol. (1873); JAMES H. JEANS, *The Mathematical Theory of Electricity and Magnetism*, 5th ed. (1925, reprinted 1960); WILLIAM R. SMYTHE, *Static and Dynamic Electricity*, 3rd ed. (1968); JULIUS A. STRATTON, *Electromagnetic Theory* (1941).

Direct electrical currents: Few books devote more than a brief section to direct current circuits. General network theorems are usually given in the context of ac circuits although they also apply to dc circuits. A clear and useful introduction to circuit analysis is ERIC A. FAULKNER, *Principles of Linear Circuits* (1966). For a more detailed treatment one has to refer to older texts such as ANDREW GRAY, *Absolute Measurements in Electricity and Magnetism*, 2nd ed. (1921, reprinted 1967); and DONALD G. FINK and JOHN M. CARROLL (eds.), *The Standard Handbook for Electrical Engineers*, 10th ed. (1968), a useful source of information.

The conduction of electricity: There is a general discussion of conduction in BETTY I. BLEANEY and BREBIS BLEANEY, *Electricity and Magnetism*, 2nd ed. (1965). More detailed discussions of conduction in solids will be found in the many texts on solid state physics. The three listed here are in increasing order of difficulty: ROBERT L. SPROULL, *Modern Physics*, 2nd ed. (1963); CHARLES KITTEL, *Introduction to Solid State Physics*, 4th ed. (1971); and JOHN M. ZIMAN, *Principles of the Theory of Solids* (1964). The following books deal with more specialized topics: FRANS M. PENNING, *Electrische Gasontladingen* (1955; Eng. trans., *Electrical Discharges in Gases*, 1957), an excellent simple introduction to this subject; ALFRED VON ENGEL, *Ionized Gases*, 2nd ed. (1965); WILLIAM SHOCKLEY, *Electrons and Holes in Semiconductors, with Applications to Transistor Electronics* (1950), the classic work on the subject; ROBERT A. SMITH, *Semiconductors* (1963); ALISTAIR C. ROSE-INNES and E.H. RHODERICK, *Introduction to Superconductivity* (1969), a readable introduction; A.B. PIPPARD, *The Dynamics of Conduction Electrons* (1965), an advanced and specialized text; D.K.C. MACDONALD, *Thermoelectricity* (1962); S.M. SZE, *Physics of Semiconductor Devices* (1969), deals with a number of topics—*e.g.*, conduction processes in semiconductors and photoeffects; JOHN R. PIERCE, *Theory and Design of Electron Beams* (1949), a clear account of the principles that govern the operation of electron tubes employing electron beams —deals with the effects of both space charge and thermal velocities and the basic principles of electron optics.

Alternating electric currents: RAMO, WHINNERY, and VAN DUZER (*op. cit.*), has a good introduction to the use of complex numbers in ac theory and also deals with waves on transmission lines. BLEANEY and BLEANEY (*op. cit.*), has sections on ac theory, measurements, and filters. FAULKNER (*op. cit.*), is a useful introduction to more advanced mathematical methods and the theory of transients. BERNARD HAGUE, *Alternating Current Bridge Methods*, 6th ed. rev. by THOMAS R. FOORD (1971), is comprehensive. F.N.H. ROBINSON, *Noise in Electrical Circuits* (1962), gives an introduction to concepts and principles; ALDERT VAN DER ZIEL, *Noise* (1954), gives a fuller treatment.

DAVID P. HOWSON and R.B. SMITH, *Parametric Amplifiers* (1970), gives a treatment of one particular type of nonlinear circuit and its applications.

The primary effects and properties of electric fields and currents: LANDAU and LIFSHITZ (*op. cit.*), provides the best coverage of field energy and stress; in particular, their treatment of the relation between thermodynamics and electromagnetism is superior to anything to be found elsewhere. Books on relativity abound. A good example, which does not ignore the relation between relativity and electromagnetism, even though it is at an elementary level, is CLAUDE KACSER, *Introduction to the Special Theory of Relativity* (1967). ROLF HAGEDORN, *Relativistic Kinematics* (1963), is an advanced text mainly concerned with applications in particle physics. It does, however, contain an exceptionally clear exposition of the basic theory. JOHN F. NYE, *Physical Properties of Crystals* (1957), treats piezoelectricity, pyroelectricity, and the electro-optic effect. This is not a difficult book but it does introduce and use tensor notation. WARREN P. MASON, *Piezoelectric Crystals and Their Application to Ultrasonics* (1950), is a useful work on piezoelectricity. AMNON YARIV, *Quantum Electronics* (1967), deals with the electro-optic effect, also light emission from semiconductors. SZE (*op. cit.*), may also be consulted for light emission from semiconductors. HENRY F. IVEY, "Electroluminescence and Related Effects," *Adv. Electronics Electron Phys.*, Suppl. 1 (1963), deals at length, but without using advanced theoretical concepts, with this topic. Black body radiation is dealt with in most texts on modern physics. MAX BORN, *Moderne Physik* (1933; Eng. trans., *Atomic Physics*, 8th ed., 1969); and SPROULL (*op. cit.*), are good texts.

The effects of electricity on organic matter: FRANCIS O. RICE and EDWARD TELLER, *The Structure of Matter* (1949), is a useful introduction to general ideas about atomic physics. HUGO R. KRUYT and J.T.G. OVERBEEK, *Inleiding tot de physische chemie*, 15th ed. (1958; Eng. trans., *An Introduction to Physical Chemistry for Biologists and Medical Students*, 1960), is also a useful introduction to electrochemistry for physicists. DUNCAN A. MACINNES, *The Principles of Electrochemistry* (1939, reprinted 1961), is an adequate introduction despite the date of its original publication. R.V. COXON and ROBERT H. KAY, *A Primer of General Physiology* (1967), gives the biological background for bioelectric effects. ALAN L. HODGKIN, *The Conduction of the Nervous Impulse* (1964); and BERNARD KATZ, *Nerve, Muscle, and Synapse* (1966), deal more specifically with nerves.

(F.N.H.R.)

Electric Motor

An electric motor is a device for converting electrical energy to mechanical energy. It is the complement of the electric generator, a device that converts mechanical energy into electrical energy, the two devices together providing the link between electrical and mechanical systems. In general, generators may act as motors and, conversely, motors may act as generators; they differ only in some construction details and in their auxiliary equipment.

HISTORY

The history of the electric motor begins with Hans Christian Ørsted (1777–1851), who discovered that electricity in motion produces a magnetic field. In searching for a demonstration of the converse of Ørsted's discovery, Michael Faraday in 1821 built an apparatus that demonstrated the principle of the electric motor. In 1823 Peter Barlow and William Sturgeon in England wrapped a current-carrying conductor around an iron bar to produce a strong electromagnetic force, and Joseph Henry in the U.S. learned that insulated wire wrapped several layers thick around a piece of iron made a more powerful electromagnet. In 1833 a young American blacksmith named Thomas Davenport first saw an electromagnet. He began experimenting with electromagnets and batteries, and, after considerable effort, invented an electric motor. In 1835 Davenport showed his invention to Henry, who gave him a certificate attesting to the importance and originality of the device. Though Davenport obtained a patent, his motor was never a very great commercial success.

The first electric motor of commercial significance was demonstrated in Vienna in 1873 by Zénobe Théophile Gramme, a Belgian-born electrical engineer. Developer of the ring armature, which enormously improved the efficiency of early electric generators, he had discovered that if a large dynamo were electrically connected with a smaller one, the smaller dynamo would revolve and act as an electric motor. Hence, Gramme's motor was a combination of two direct-current dynamos, acting as generator and motor, respectively. In 1888 Nikola Tesla invented the first alternating-current electric motor, the prototype for most of the electric motors in use today.

BASIC PRINCIPLES

Electrical and mechanical coupling methods

The coupling method between electrical and mechanical systems can be (1) electromagnetic, in which forces are established as a result of an interaction between magnetic fields created by stationary and moving parts; (2) electrostatic, in which forces are produced by the interaction of electric fields created by the stationary and moving parts (see ELECTRICITY); (3) magnetostrictive, represented by the microscopic change in size experienced by ferromagnetic materials when subjected to an external magnetic field (see MAGNETISM); (4) piezoelectric, represented by the deformation of certain crystals when subjected to an electric field (see PIEZOELECTRIC DEVICES); and (5) magnetohydrodynamic, or the production of a force due to the interaction between magnetic fields and electrically conductive fluids and gases (see MAGNETOHYDRODYNAMIC DEVICES). By far the most commonly exploited phenomenon is the electromagnetic, because it makes possible the conversion of the most energy per unit weight, volume, and cost. Electric motors, which utilize this phenomenon, may range in output from a fraction of a watt (0.0001 horsepower) for use in instrument control systems and servomechanisms, to several thousand horsepower, for pumps and wind-tunnel drives.

Two general principles involved in motor and generator operation are those of Ampère and Faraday. Ampère's law says that a conductor will experience a force if a component of current in that conductor flows at right angles to a magnetic field. Faraday's law states that a potential difference will exist between the ends of a conductor if that conductor is given a component of motion perpendicular to the magnetic field.

The rotor and stator

To exploit these principles and convert electrical energy into mechanical energy, motors must consist of two parts—a rotor and a stator. The rotor is usually round and is the moving portion. It contains conductors to establish and shape magnetic fields that will interact with other magnetic fields produced by the stator. The rotor also may have a contacting device (slip rings or commutator) to connect it electrically with the external circuit, together with a supporting shaft and bearings, an integral blower for cooling, etc. The stator consists of similar magnetic materials and electrical conductors to establish and shape the magnetic fields, as well as the frame to support the whole machine. The manner in which the conductors are arranged (and in which the magnetic fields are controlled) determines the type of motor and its characteristics.

INDUCTION MOTORS

In an induction motor, one winding—a conductive path, usually wire—induces a current into another winding. Usually the stator winding is connected to an external power source; the other, on the rotor, consists of continuous wire loops. The magnetic field established when currents flow in these conductors is concentrated in a narrow air gap between the stator and the rotor. The purpose of the windings connected to the external energy source is to produce a rotating field that interacts with the induced currents in the rotor conductors, thereby establishing forces on those conductors.

To set up a rotating field, there must be at least two windings carrying currents whose phase—the time when the amplitudes are a maximum—is not the same. Figure 1A illustrates two alternating currents whose amplitudes —the magnitudes of the wave, or peak value measured from zero—are the same but whose phase relationship is

90°. Figure 1B shows schematically a stator on which four coils have been wound; coils 1 and 3 are so connected that if the current i_1 is in the direction shown, the polarities of the magnetic field at the stator surface will be north at the upper and south at the lower coils. When the current i_1 reverses, so do the polarities of the fields. Coils 2 and 4 are also connected so that a current i_2 produces a similar field. As shown in Figure 1A, at a given instant of time t_1, i_1 is a maximum and i_2 is zero. Under these conditions the field situation is as shown in Figure 1C. A quarter-period later, at t_2, the current i_1 is zero, and i_2 has risen to a positive maximum. The resulting field conditions are now as shown in Figure 1D. At a third instant, t_3, i_1 has advanced to a maximum, but in the negative sense, while i_2 has fallen back to zero from a positive maximum. The field situation is now that of Figure 1E. Finally, at t_4, i_1 has again become zero and i_2 has reached a negative maximum, and Figure 1F shows the resulting fields. The north pole is found to travel around the periphery of the stator, making one complete revolution while the current goes through one cycle. Hence a revolving field is produced, and its speed depends solely upon the geometry of the coils and the time required for one cycle. For a 60-hertz power supply, 60 revolutions will be made in one second, or 3,600 revolutions in one minute. The speed of the revolving field is known as the synchronous speed.

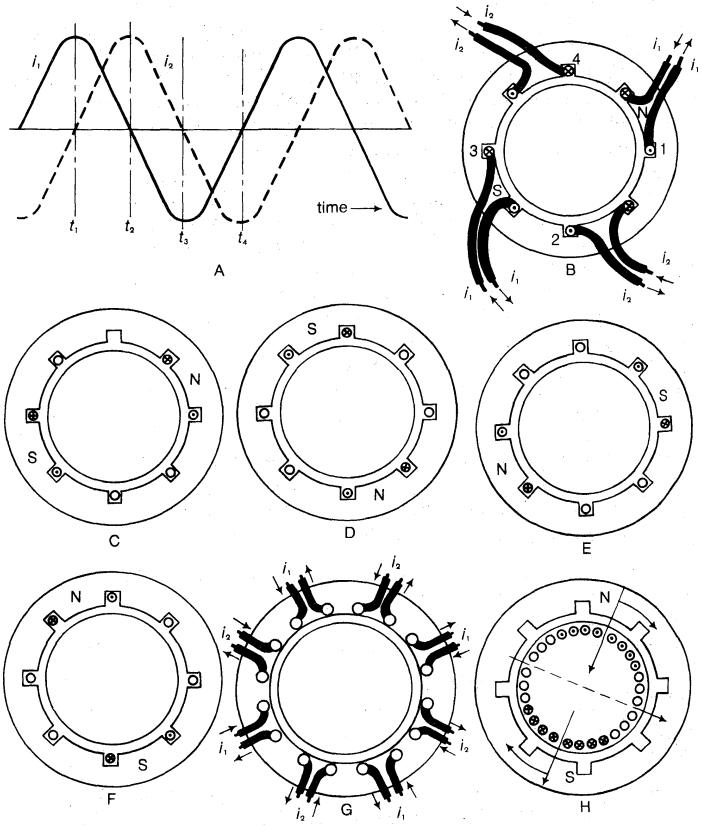


Figure 1: Induction motor fields and their interaction with rotor and stator conductors. ⊙ indicates that the flow of current is directed toward viewer, perpendicular to page; ⊗ current is directed away from viewer (see text).

If each coil is reduced in size so that it spans just one-half the stator surface (Figure 1G), the revolving field will travel only half as far for each cycle of the supply current, and the speed of the revolving field will become 1,800 revolutions per minute for a 60-hertz supply. The coil arrangement of Figure 1B is said to produce a two-pole motor, since at any instant there is but one north and but one south pole. The arrangement of Figure 1G, on the other hand, will produce two north poles and two south poles at any instant and therefore is a four-pole stator. It is not uncommon to have as few as two or as many as 100 poles in a motor. The speed of the revolving magnetic field in revolutions per minute in any motor is $120\,f/p$, in which f is the frequency in hertz and p is the number of poles.

The four-pole stator

In a practical motor the stator will have more coils than indicated here, and these will be distributed around its periphery in equally spaced slots. This arrangement makes better use of the available space than does the concentrated-coil arrangement described above, but the result is still the same—a revolving field is established.

If a set of conductors is uniformly spaced around the periphery of a rotor, as shown in Figure 1H, and if the rotor is stationary and a two-pole revolving field is located in the position shown, a voltage will be induced (according to Faraday's law) in a direction out of the page in the upper conductors and into the page in the lower conductors. If these conductors are now short-circuited through a slip ring—a device for making electrical connections between stationary and rotating contacts—currents will flow in the same directions as the induced voltages. These currents in turn will produce their own magnetic field, and its direction will be shown by the dashed line in Figure 1H. The north pole of the rotor and south pole of the stator will attract each other, and a torque will be exerted on the rotor tending to turn it in the direction of the rotating field.

If the rotor is travelling at the same speed as the revolving field, there will be no induced voltage in the conductors, no current flow, and no torque. Hence, at synchronous speed the torque drops to zero. If the rotor travels faster than synchronous speed, the direction of the induced voltage and the current will be reversed, thus reversing the torque; the motor is now serving as an induction generator.

Construction features. The rotor and stator are built up of laminations of 0.25 percent to 3.50 percent silicon steel, from 0.014 to 0.025 inch (0.36 to 0.63 millimetre) thick. The laminations are sometimes coated with an insulating varnish to reduce the magnitude of currents induced in the laminations (eddy-current losses). Uniformly distributed slots are punched in the steel to receive windings.

The coils (usually the stator coils) which are connected to the external energy source may be preformed and insulated and then inserted into the slots, or the conductors may be placed directly into the slots and the insulation applied afterward. The former technique is presently used for motors rated at more than 600 volts or about 100 horsepower.

Types of secondary windings

The conductors that form the secondary winding may be of three types: wound, squirrel-cage, or solid. To make the wound type, a two-phase winding (having a phase difference of 90 electrical degrees, or one-quarter cycle) or three-phase winding (differing in phase by one-third of a cycle, or 120 electrical degrees) is placed in the rotor slots; and some of the terminals are brought out to slip rings, the insulated brass or bronze rings on which carbon brushes ride, to provide an electrical connection to the rotor coils.

The squirrel-cage winding may be either cast from aluminum or copper or built up from copper or brass bars welded to a common end ring. To cast such a winding, a stack of rotor laminations is inserted in a centrifugal casting machine and molten aluminum or copper is introduced. The result, after the metal solidifies, is an integrally cast set of longitudinal conductors and two end rings. The cast winding generally is used for standard production line motors in sizes under 100 horsepower. For larger motors the welded bar type is more common.

An example of the solid or continuous winding is the drag-cup motor, in which a nonmagnetic metal is rotated in a magnetic field to generate a torque or voltage proportioned to its speed. A solid copper sheath shrunk on the rotor steel will also provide a continuous conductor. Continuous-winding rotors are used principally for servomotors.

Polyphase induction motors. The three-phase stator connection is by far the most common, because most in-

dustrial power distribution systems are three-phase. In this type of motor the revolving field is produced by three different currents whose phase relationships are 120° instead of the 90° shown in Figure 1A.

Both squirrel-cage and wound-rotor motors may be used in polyphase service; however, the latter are usually reserved for speed-control applications because of the added cost of manufacture. Squirrel-cage motors may also be classed in terms of their starting current and their speed-torque curves.

Single-phase induction motors. In describing how a revolving field was produced it was pointed out that a necessary condition was the existence of at least two currents out of phase in time and flowing in coils displaced in physical position from one another. While it is possible to satisfy the spatial requirement when a single-phase supply is used, it is more difficult to satisfy the time relationship for the currents flowing in the two sets of coils. The most common ways of solving this problem are described below.

Capacitor-start, capacitor-run motor. For this type of motor the stator is built as for a two-phase motor—*i.e.*, with windings spaced 90° apart in a two-pole case (Figure 2A). The main winding is connected directly across the

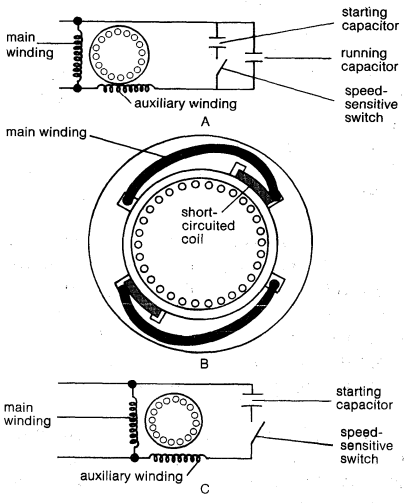


Figure 2: *Single-phase motor circuits.* (A) Capacitor-start, capacitor-run; (B) shaded pole; (C) capacitor-start (see text).

power line, and an auxiliary winding is connected to the line through a suitable capacitor, a device that stores an electric charge, to provide a means of starting. The electric field in the capacitor tends to overcome the effect on the circuit of the magnetic field in the motor so that the current flowing in the auxiliary winding (containing the capacitor) is out of phase with that in the main winding and the necessary theoretical conditions are satisfied. One difficulty with this arrangement is that the 90° phase relationship is obtained only during operation at the rated speed. In order to obtain satisfactory characteristics during the starting period, another capacitor is placed in the circuit to retain the 90° phase relationship; the second capacitor is then taken out of the circuit at about 80 percent of rated speed through the use of a centrifugal switch.

Shaded-pole motor. If a permanently short-circuited coil is placed over a portion of the main field coil, thus "shading" it as shown in Figure 2B, the result will be a field similar to that produced by two out-of-phase currents. Because the winding cannot be displaced 90° and the currents cannot be 90° out of phase in time, such a motor has a low speed–torque curve and is used for the most part to run fans, phonographs, and household appliances.

Other types. Another group of motors employs a revolving field only during the starting period and then runs on a pulsating field, which may be regarded as two counter-rotating fields of equal magnitude. If the rotor response was the same to both fields, the net torque would be zero and the motor would not run at all. The rotor impedance can be made higher to the field rotating in the reverse sense than to the field rotating forward, however; thus, less reverse current flows in the rotor bars and a net forward torque results. At a standstill, on the other hand, the forward and backward torques are equal, and the motor will not start unless some special provision is made. For example, an auxiliary field may be put into the circuit for starting and then switched out when the motor attains about 80 percent of its rated speed; or internal connections may be provided that in effect make the motor a different type for starting (*e.g.*, the repulsion-start motor, a single-phase motor in which the stator winding is subsequently connected to the source of power and the rotor winding to the commutator), after which the connections are changed automatically so that the motor will run on the pulsating field.

Of the auxiliary field types, the capacitor-start motor is very common; its circuit is shown in Figure 2C. In the resistance split-phase motor the capacitor is omitted, and consequently there is lower starting torque, since the currents cannot be shifted 90° out of phase. Repulsion-motor principles are also used. During the starting period the motor operates as a repulsion motor until a centrifugal switch short-circuits the repulsion winding. Thereafter the machine operates primarily as an induction motor.

Applications of capacitor-start motors

Capacitor-start motors are used for refrigerators, compressors, and other equipment in which there is a high starting torque requirement. Resistance split-phase motors are used for light tools and other equipment starting unloaded. Repulsion-start motors have been largely replaced by capacitor-start, capacitor-run motors.

Servomotors. The motors used in control systems may be operated from two-phase or single-phase supplies. Figure 3 shows the usual method of energizing a servomotor through an amplifier (a device that draws power from a source other than the input signal and that produces as an output an enlarged reproduction of its input). The reference-phase voltage is constant, but the control phase has a variable voltage impressed upon it from the amplifier, which in turn receives its signal from the rest of the control system. If the rotor resistance is relatively high, the motor will exhibit variable speed characteristics.

These motors sometimes have drag-cup rotors, but they are more commonly of the squirrel-cage type. They are limited to small outputs because much of the power input must be dissipated as heat in the rotor.

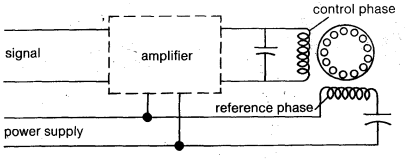


Figure 3: Elements of a typical servomotor circuit.

Linear motors. High-speed ground transportation needs have stimulated the development of the linear induction motor. Linear induction motors may take several forms. Most often encountered is the double stator with a continuous rotor as shown in Figure 4. When used on a rail vehicle the continuous sheet "rotor" is suitably fastened to the roadbed, and the "stator" is suspended between the wheels in the centre of the vehicle. The polyphase power supply is generated either by a source of power on the vehicle or picked up by trolley from a suitable line paralleling the track. Besides applications in ground transport, linear motors have been used for belt conveyers, shuttles in textile looms, curtain rods, aircraft launchers, and electromagnetic pumps in which the solid "rotor" is replaced by a conducting fluid (such as the sodium–potassium alloy, NaK, or mercury).

From a theoretical point of view, a linear motor may

be regarded as a cut-and-unrolled cylindrical motor, with suitable adjustments in the theory to account for the transients occurring at the two ends of such an unrolled cylindrical motor. In practical linear machines, these end effects account for about 10 percent reduction in the force produced.

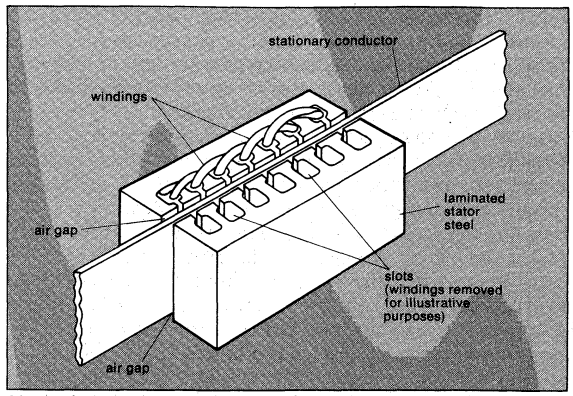


Figure 4: Semi-cutaway view of a linear motor of the type used experimentally for propelling rail passenger vehicles.

The tubular motor

Another type of linear motor is the tubular motor, which resembles the linear accelerator (see ACCELERATOR, PARTICLE). These motors use concentric coils and, when excited by a polyphase current supply, a travelling wave is developed. This wave interacts with any conductor, producing a force that tends to make the conductor travel at the same velocity as the travelling magnetic wave. Such machines are useful as electromagnetic cannons or missile launchers.

Speed control. Induction motors basically are constant-speed devices and are used most frequently as such. If, however, speed control is desired, several methods are available: frequency control, pole changing, addition of rotor or secondary resistance, stator voltage control, and counter-rotor voltage control.

In the discussion on the revolving field it was pointed out that the speed of the field is a function solely of the power-supply frequency and the design of the winding. Thus, if the frequency is varied, the synchronous speed of the field will change and, depending to some extent on the load, the shaft speed will also change. This method requires some external device to alter the ordinarily unvarying frequency of the power supply. With the advent of a high-current silicon controlled rectifier (SCR), primary frequency control has become extensively employed for speed control purposes.

The pole-changing method involves connecting the individual stator coils (as in Figure 1G) externally. Thus, adjacent coils may be of either the same or opposite polarity, depending on how they are connected. If adjacent coils are of the same polarity, the result will be half as many magnetic poles as when the coils are of opposite polarity, and the synchronous speed will be doubled. However, this method makes it possible to change the motor speed only in discrete steps (usually 2:1); it does not provide for continuous variation.

Control of motor speed by varying the power-supply voltage is not effective in alternating-current induction motors (except in servomotors), though some speed variation is possible within narrow limits.

Secondary or rotor resistance control can be applied only to a wound-rotor motor. One difficulty is that the losses in the external resistor may become so substantial that it is desirable to offset them by employing counter-voltage methods. These involve the use of auxiliary machines and are justifiable only when the rotor loss is sufficiently important economically to warrant the investment in additional equipment.

Starting and protection. Induction motors are extremely rugged and in the smaller sizes (under 25 horsepower) usually will withstand sudden application of full voltage at standstill without damage. For larger motors, however, the electrical supply system must be protected against sudden surges arising from the large starting currents produced. All induction motors (except servomotors and those used in speed-control service) during the starting period draw from four to six times the normal running current. When the electrical system will not withstand such currents there are four techniques that may be used: reduced-voltage, resistor or reactor, Y-delta, or part-winding starting.

Starting techniques

For reduced-voltage starting an autotransformer—a single-coil transformer in which primary and secondary are part of the same coil—is used to lower the applied voltage to 65–80 percent of normal. Because the current is proportional to the square of the voltage, an 80 percent reduction in voltage will mean that only 64 percent of the full-voltage current will be drawn. Autotransformers, however, are designed to be used only for only very short periods of time.

The second method involves simply the provision of resistors and reactors that can be switched into the supply lines to limit the current during starting.

For motors operating on three-phase power, the windings may be connected either with both ends across the line (delta) or with one end on the line and the other connected to the remaining two windings (Y). If a motor designed with delta-connected windings is reconnected Y, the voltage across the windings is reduced to 57 percent of normal and the current will be only one-third that drawn at full voltage.

Part-winding starting is possible if connections are provided so that only part of the motor winding is in the circuit during starting, the remainder being switched in as the motor picks up speed.

Protecting induction motors is primarily a problem of preventing damage to the windings from excessive heat. On motors under five horsepower, internal temperature-sensitive switches are sometimes provided to open the main line if the temperature rises above a certain value. Temperature detectors embedded in the windings are used on large motors.

SYNCHRONOUS MOTORS

Synchronous motors, like induction motors, operate on the principle of the revolving magnetic field, produced usually by the stator; the rotor of the synchronous motor is designed so that it can produce a constant field. If, for example, a two-pole rotor of this type is excited with direct current through slip rings, a north and a south magnetic pole will be produced on the rotor. If the rotor is travelling at the speed of a revolving stator field such as that shown in Figure 1 (C, D, E, F), the two fields will lock together and the rotor will be pulled around at the same speed as the revolving field.

Construction. In most synchronous motors the winding that produces the revolving magnetic field (known as the armature winding) is on the stator, and the design of such a stator is essentially the same as for an induction motor. (For a synchronous motor that has its armature winding on the rotor, the design is much the same as for a wound-rotor induction motor.) A unidirectional field, to interact with the revolving field produced by the stator armature winding, is consequently produced on the rotor by exciting a direct-current winding from a direct-current source.

Motors using direct-current fields must have some means of starting. In most motors this is accomplished by installing a squirrel-cage winding in the pole faces of the rotor. Such a motor operates on induction motor principles up to about 95–97 percent of synchronous speed, at which point the direct-current field is applied. The motor then locks into synchronism—that is, an exact relationship is established and maintained between motor speed and power supply frequency. Motors with this type of field winding usually are more costly to build than general-purpose induction motors in sizes up to 300–500 horsepower.

In small motors it is possible to eliminate the need for a direct-current power source by building the field struc-

ture from the same metals used for permanent magnets. The field structure then is given a permanent residual magnetic field that locks in with the revolving field at synchronous speed. So-called hysteresis motors have a lower power-to-weight ratio than the direct-current excited types described above, but they are useful in some low-load applications in which constant speed is important—as gyro-spin motors in navigational equipment, in phonographs and tape recorders, and in clocks.

Hysteresis and reluctance motors

Another type of field construction is found in the reluctance motor, in which no rotor winding is used at all, the toothed design of the rotor producing the synchronous motor characteristics. If the rotor lags behind the revolving field, the average magnetic energy stored in the air gap around the rotor decreases. This produces a greater positive ratio between the stored magnetic energy and the displacement angle, and a greater torque is established, pulling the motor back into synchronism. The power-to-weight ratio here is even more unfavourable than in the hysteresis motor, but the reluctance motor is economical to construct and is therefore often used to power certain instruments and electric clocks.

Polyphase synchronous motors. Like most induction motors, synchronous motors are designed to operate on three-phase power in sizes above one horsepower. Furthermore, in such sizes they are always direct-current excited. The source of direct current may be rectifiers, direct-current generators (exciters) driven from the motor shaft, or separate motor-generator sets.

One important feature of synchronous motors is that control of the field current affords a means of power factor control. When the direct-current field of a synchronous motor is overexcited, the motor must draw a demagnetizing component of current from the power lines to compensate for the overmagnetization that takes place. The demagnetizing current drawn by the synchronous motor will balance some of the magnetizing current drawn by other induction machines on the same power line. This component of alternating current transmits no energy but does cause additional loss in all associated equipment, such as transmission lines and transformers. In most cases, power company rates are adjusted to reflect the cost of supplying magnetizing current, and it is often economically feasible to correct for this condition in a manufacturing plant rather than to pay the power company to do so. Thus, a synchronous motor may serve the dual purpose of driving a load at constant speed and at the same time reducing the net magnetizing current drawn from the power lines. In fact, in some cases it is economically justifiable to operate a motor solely for the purpose of power factor correction. A motor designed for this service has no output shaft and is referred to as a synchronous capacitor.

Single-phase synchronous motors. As was the case for induction motors, a revolving field is needed, and to produce it a capacitor may be inserted in the second phase of a two-phase motor, or a shaded-pole structure may be used. Thus a revolving field is established, although it is inferior to that provided by a polyphase supply.

There is relatively little demand for direct-current, field-excited motors in sizes under one horsepower, so virtually all small motors are hysteresis or reluctance types. Such motors rarely develop more than 20 percent to 50 percent of the power that could be expected from a magnetic circuit of the same size in an induction motor.

Electric clock motors

Of special interest is the single-phase synchronous motor used in electric clocks. A two-pole rotating field is set up by means of shaded-pole construction, in which one or more auxiliary short-circuited stator windings are provided and displaced magnetically from the main winding. The resulting rotating field reacts with one or more thin disks of hardened steel especially chosen for their desirable magnetic properties and housed in a sealed metal container. The rotating field starts the motor by eddy-current action in the container and brings the motor up to speed, usually 3,600 revolutions per minute (rpm). As the motor approaches synchronous speed, the rotating field magnetizes the disks and they then are locked in with the revolving field. A sealed gear train reduces the speed to one revolution per minute at the motor shaft. While these motors have efficiencies of less than 1 percent, they are very inexpensive to build and operate.

Speed control. Synchronous motors are less amenable to speed control than are induction machines. Only two methods are suitable: frequency control and pole changing. The principles are the same as for induction motors.

Starting and protection. Small synchronous motors of the hysteresis and reluctance type are started on full voltage and are protected (if at all) by a thermal cutout in the same fashion as induction machines.

Larger machines using direct-current fields have more troublesome starting problems than any other class of electric motors. Not only must the current inrush to the motor from the power source be controlled, as was the case for the induction motor, but the field must be protected from high voltages arising from transformer action. Then, when the rotor has been accelerated up to synchronous speed, direct current must be applied to the field at the proper instant. On the other hand, synchronous motors may be designed to have almost any type of starting torque characteristic without sacrificing good running features. Starting torques of from 400 percent to 600 percent of full load may be obtained, with corresponding starting currents 250–600 percent of normal. The method of limiting current inrush to the armature winding is the same as for induction motors.

During the starting period, the field winding is generally short-circuited. When the machine has accelerated to 95–97 percent of synchronous speed, direct current is introduced to the field and the motor locks into synchronism. Application of the direct-current field current at the proper moment can be done automatically by taking advantage of two motor characteristics: as the motor nears synchronous speed, the frequency of the current induced in the rotor circuit approaches zero; and as the rotor "slips" with respect to the revolving field, there are corresponding pulsations in the line current. These conditions can be detected either by a frequency-sensitive relay in the field circuit or by a special relay in the line, and the relays can then initiate the switching on of the direct-current field current.

The protection of synchronous motors, while more complicated than in the case of induction motors, centres on the same basic problem: prevention of damage to the windings from heat caused by excessive currents. Specifically, the problems encountered and their solutions are (1) excessive stator current, usually detected by external sensors or embedded temperature detectors; (2) excessive field current, detected by external sensors; (3) excessive field voltage, which can be controlled by short-circuiting the field winding through a resistance during starting or by the use of discharge gaps between the slip rings; (4) squirrel-cage over-heating, controlled by external timers; and (5) phase failure (causing abnormally high current in the squirrel-cage winding), detected by external sensors.

COMMUTATOR (DIRECT-CURRENT) MOTORS

Commutator motors excited with direct current (called direct-current motors) were the forerunners of induction and synchronous machines, because many cities had well-established direct-current distribution systems for both lighting and power while alternating-current systems were still speculative. Hence the direct-current member of the commutator motor family was well established when Nikola Tesla invented the induction motor. Modern commutator motors find their main use in speed-control applications.

Basic principles. Unlike induction and synchronous machines, commutator motors (with a few exceptions) make use of stationary rather than revolving fields—in fact, the sole purpose of the commutator is to control the position of the rotor field irrespective of the speed of rotation or the frequency of the current in the windings. For the purpose of explaining how such motors work, it is most effective to develop the principles for direct-cur-

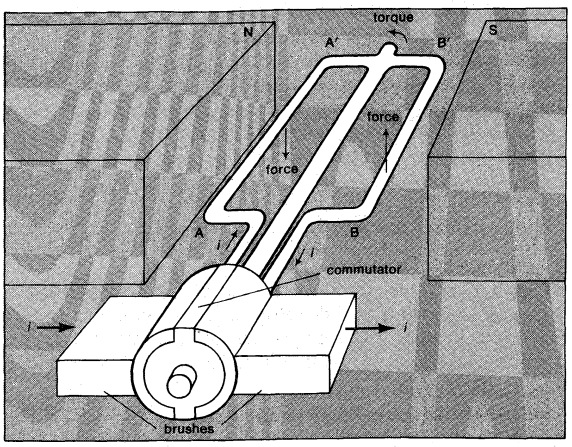


Figure 5: Operating principles of a direct-current motor (see text).

rent motors and then extend those concepts to alternating-current motors.

In Figure 5 a coil is mounted on a shaft so that it is free to rotate in the magnetic field. If current i is caused to flow in the directions shown, a downward force will be evident on the left side of the coil and an upward force on the right side, and the coil will tend to rotate in a counterclockwise direction. To maintain this rotation, however, the current i must be reversed by the time the conductor AA′ comes under the south pole. This is accomplished by the commutator, which consists of an insulating member, carrying two copper segments, that rotates with the coil. Riding on the commutator but fixed to the motor frame are two brushes, usually made of graphite. These provide the connection to the external circuit. When the current i is forced into the left brush, it flows through the copper commutator bar and thence into the coil side AA′. It returns through B′B to the other commutator bar segment and out the right-hand brush. When coil side AA′ has moved around to BB′ the current in the coil will have been reversed.

In an actual motor, greater field strength in the vicinity of the conductor will be achieved if the magnetic circuit cross section is like that of Figure 6. Also, in an actual motor the number of coils is multiplied so that there may be from three (in very small, low-voltage motors such as those used in models) to 300 or more (in very large steel mill motors). The number of commutator segments also will increase, for there must be one for every coil.

When the current in a rotor (armature) coil is suddenly reversed by the commutator, the coil is momentarily short-circuited, and the voltage induced in it by the changed current may cause sparking at the brushes and consequent eroding of the commutator. In motors under one horsepower this is not a serious problem, but in larger motors it is customary to reduce sparking by inducing a voltage equal and opposite to that caused by the change in current. Shifting the brushes will do this, but a different angle of shift is needed for each load condition. The use of an interpole, or commutating pole, whose winding carries the same current as the armature, is a more satisfactory way of inducing an opposite voltage. The location and relative size of the interpoles is shown in Figure 6.

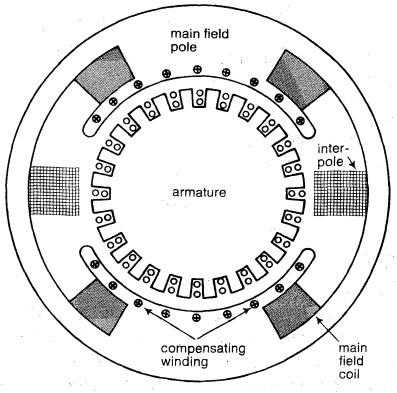


Figure 6: Simplified cross section of direct-current motor (see text).

In large motors that must operate under rapidly changing load conditions, an additional winding may be used to reduce the problems created by rapidly changing armature current. Such a winding is called a compensating winding and is placed in the pole faces as shown in cross section in Figure 6. This winding is in series with the armature conductors, but it is arranged so as to oppose the field established by the rotor conductors.

The compensating winding

The actual construction of a commutator motor is similar to that of induction and synchronous machines. Thin steel laminations are stacked together to form the armature and the field poles. Coils are usually preformed, although for some small motors the coils are wound on automatic winding machines that insert the wire turn by turn into the slots. The commutator is built up of segments of copper and then either cast in an insulating compound or fastened together with compression clamps.

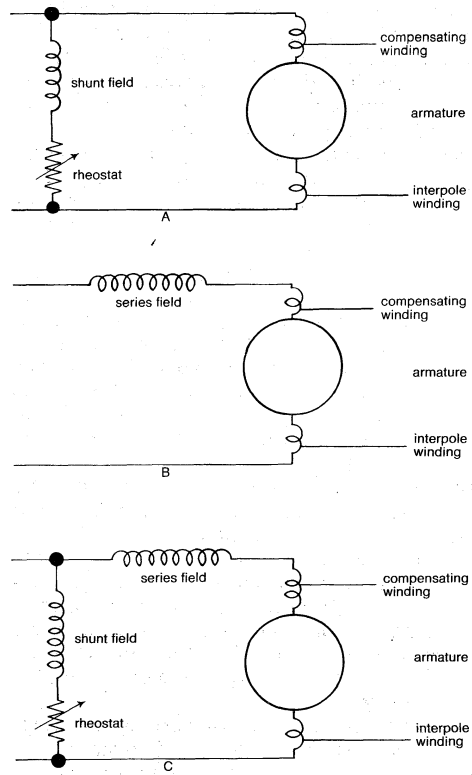


Figure 7: *Connections for direct-current motor.* (A) Shunt; (B) series; (C) compound (see text).

Direct-current motor characteristics. The operating characteristics of direct-current motors are determined solely by the manner in which the main field is excited.

In the shunt motor, the main field coil is placed across the direct-current line (Figure 7A). This type of excitation tends to give a constant magnetic field, and the speed–torque characteristic would fall off only because of the resistance drop in the armature circuit. However, the armature field arising from the currents in the rotor conductors (armature reaction) tends to demagnetize the main field to some extent. The result is that the shunt motor usually has a slightly concave upward speed–torque characteristic.

A series motor (Figure 7B) has its field winding in the same circuit with the armature. This produces a speed-torque curve with a high no-load speed. Because of their very high starting torque, these motors are useful for traction service.

Compound motors (Figure 7C) have both shunt and series windings, giving them composite characteristics. Usually the series field has small effect under a normal load condition. For example, a motor whose series field aids the shunt field tends to run as a shunt motor under normal load, but it also will develop higher starting torques than a shunt motor because of the contribution of the series winding at standstill.

ALTERNATING-CURRENT COMMUTATOR MOTORS

An alternating-current commutator motor presents two problems that do not arise in direct-current motors: the phase relationships between currents in different parts of the motor may not permit interaction and there are now voltages present due to transformer action.

In the series alternating-current motor, the phase problem is not present because the currents in both the armature and field are the same. Hence the armature current reverses at the same instant the field current reverses and a unidirectional torque is produced. These motors are very common in such household appliances as vacuum cleaners, mixers, and hand-held drills. These small motors for domestic use are sometimes called universal motors, since they may be used on direct-current or on any alternating-current frequency to 60 hertz. In larger sizes, series motors are used for alternating-current railway service, especially on 25- and 16⅔-hertz power.

Speed control. Mention has been made of the use of commutator motors for alternating-current speed control; however, the direct-current machine is more commonly used for variable-speed applications.

The simplest way to control the speed of a shunt motor is to vary the field current or the armature voltage. This can be accomplished with a variable field resistor (marked rheostat in Figure 7A) and a variable armature resistor. If both of these are set for zero resistance, the motor will operate at its base speed. Decreasing the field current (*i.e.*, increasing the field resistance) but keeping the armature voltage unchanged will cause the motor to increase in speed. Because of the centrifugal forces involved, the maximum speed that can be achieved in this way is about double the base speed. If the field resistor is set at zero and if the armature voltage is decreased by inserting more resistance in series with the armature, the speed will be decreased.

Direct-current motors may be controlled very effectively from alternating-current supplies using rectifiers and SCR's in various circuit arrangements.

Starting and protection. Except for very small sizes (⅛ horsepower), all direct-current motors must have reduced-voltage starters. Often this is a simple rheostat in the armature circuit; as the motor comes up to speed, the external resistance is gradually reduced to zero. In many cases this is done automatically by timers or current- or voltage-sensitive relays.

Overcurrent protection, a means of keeping the windings from being overheated, is common. Thermal relays, which remove the motor from the line when the motor is too warm, are often used. To protect the commutator from flashover—a disruptive discharge through air—there is sometimes provided current trip set to open the circuit at 200–300 percent of rated current.

Overspeed is a problem not often encountered in synchronous and induction motors, but it can be serious in many commutator motors. Centrifugal switches are commonly used to detect an overspeed condition and shut off or reduce the power.

Field failure on shunt motors produces very heavy armature currents and a tendency for the motor to accelerate dangerously. A field-protection relay therefore is provided to shut off power if the field current drops to zero.

BIBLIOGRAPHY. D.G. FINK and J.M. CARROLL (eds.), *Standard Handbook for Electrical Engineers*, 10th ed. (1968); S.D. USLAN (ed.), *Basic Electricity/Electronics*, vol 5, *Motors and Generators—How They Work* (1964); G.R. SLEMON, *Magnetoelectric Devices: Transducers, Transformers, and Machines* (1966); for advanced treatment, see H.H. WOODSON and J.R. MELCHER, *Electromechanical Dynamics* pt. 1 (1968); and S.A. NASAR, *Electromagnetic Energy Conversion Devices and Systems* (1970).

(R.M.S.)

Electric Power

Electric power, which is basic to modern life, is largely responsible for the high standard of living and advanced level of industrialization of much of the world. The consumption of electric power, which has expanded steadily in the 20th century, is expected to continue doubling roughly every ten years.

The modern electric-power industry was founded on a series of inventions and developments for converting the mechanical energy of falling water and the heat energy of fossil fuels (coal, petroleum, natural gas) into electrical energy.

Three distinct functions are involved in supplying electricity to the consumer: generation, transmission, and distribution. These functions correspond to the manufacture, transportation to market, and retail distribution of many other products, the chief differences being that electricity moves from the point of generation to the point of ultimate use in a continuous flow at a speed approaching that of light and that it cannot be stored in its original form.

Energy produced at central stations

Generation, the production of electrical energy from mechanical energy, takes place at central stations that may be either thermal power plants, in which the energy source is steam-driven turbines; or hydroelectric plants, in which the energy source is water-driven turbines. Transmission is the moving of this electrical energy from generating plants to bulk delivery points called substations, whence distribution delivers it to the consumer.

Generation and transmission facilities are usually economically interdependent in that the choice of location, design, and scale of generating capacity must take into account the associated cost of transmission and vice versa. Responsibilities for generation and transmission customarily go together. Distribution can be, and frequently is, conducted efficiently as a separate operation.

For many technical and economic reasons, the most appropriate arrangement for electric service to all customers in a given area is provided by a strongly integrated complex of generation and transmission facilities with sole responsibility for meeting all of the power requirements.

In distribution, the area economically covered by a single complex can be relatively small. Real distribution economies depend on the load density, while the most economical generation and transmission depend upon integration over a large interconnected area.

This article is divided into the following sections:

I. Historical background
  Before World War II
  After World War II
II. Power generation
  dc power generation
  Thermal power generation
  Hydroelectric generation
III. Transmission
  Types of transmission systems
  Transmission line design and construction
IV. Distribution
  Voltage considerations
  Networks
V. Electric power systems
  The electric utility
  System characteristics
  Interconnection
VI. International aspects
  Multinational interconnections
  World electric power

I. Historical background

BEFORE WORLD WAR II

The history of electric power goes back to 1878, when a plan to transmit electric energy to customers for the purpose of lighting was first proposed by an English inventor, St. George Lane-Fox, and by the U.S. inventor Thomas A. Edison. The first public power station was

placed in operation in London on January 12, 1882, and the second on September 4 of that year in New York City. The systems produced direct electric current (dc) of low voltage; service areas of these early central stations were limited by the short distance over which this low-voltage dc power could be economically distributed. Transmission losses made service to another city prohibitive, and not uncommonly a single city was served by a dozen or more separate systems. Consequently, although such systems played an important early role, more suitable methods were soon sought.

Development of alternating-current systems. In 1881 a Frenchman, Lucien Gaulard, and an Englishman, John D. Gibbs, had obtained English patents for a "series alternating current system of distribution," in which the current alternated from positive to negative and back a number of times every second. These patents were purchased in 1885 by an American inventor and manufacturer, George Westinghouse. William Stanley, a young U.S. inventor working with Westinghouse, perfected a transformer based on these patents and developed an alternating-current (ac) constant-potential generator. In 1886 Westinghouse and Stanley demonstrated the practicability of alternating current in Great Barrington, Massachusetts, by transmitting electric power 4,000 feet (1,200 metres), using one transformer (a device for changing ac voltage from one value to another) to increase the voltage at the source to 3,000 volts and another at the receiving end to reduce this to 500 volts. The primary advantages of the alternating-current distribution system were thus first demonstrated.

Practicability of alternating current demonstrated

No truly successful ac motor then existed. In 1888 Nikola Tesla, a Croatian-born inventor, proposed a new system known as the polyphase ac system. Its heart, or essence, was an induction motor, characterized by a rotating magnetic field. Westinghouse realized the fundamental importance of Tesla's ac system and acquired patents. The first public impact of this new system was at the World's Columbian Exposition in Chicago in 1893, where a two-phase generator, providing two voltages 90° out of phase with each other, supplied electric power for motors and lamps.

In Europe, the world's first three-phase 100-kilowatt generators (see below) were built at Lauffen power station in Germany, and the first three-phase ac transmission, between Lauffen and Frankfurt, began in 1891.

The International Niagara Commission, charged with finding the best means of transmitting hydroelectric power from Niagara Falls to the city of Buffalo, a distance of 22 miles (35 kilometres), selected alternating-current transmission in 1893. Generators were installed at Niagara Falls in 1894, and ac transmission to Buffalo began in 1896.

In the 1880s the merits of dc and ac systems were hotly debated. Though the dc system had already been somewhat developed and the ac system was new, the inherent superiority of ac was soon recognized and it rapidly became dominant. By 1898 alternating-current transmission had been sufficiently developed that a 30,000-volt, three-phase, 75-mile (120-kilometre) line from Santa Ana, California, to Los Angeles was placed in service; and in 1899 a 40,000-volt, 70-mile (110-kilometre) line from Colgate Hydro, California, to Sacramento was opened. Development of higher voltage transmission was restricted by the limitations of the pin-type insulator then available, but in 1907 the U.S. electrical engineers Edward M. Hewlett and Harold W. Buck developed the first suspension insulators, making practical higher voltages. By 1920 voltages up to 132,000 were common, and some lines were being operated at 150,000 volts.

In the 1890s steam-reciprocating-engine-driven ac generators predominated, but they had many deficiencies; they were too large and heavy, and the rotation was pulsating. In 1903 a 5,000-kilowatt steam-turbine-driven ac generator was put into service in the United States, and its output gradually increased. The steam-turbine generators of that time also had certain limitations, however, primarily limited output capacity and high fuel consumption. As a result, hydroelectric generation came to be favoured for large-scale electric-power generation. Developments were rapid. From the 1920s many large-scale hydroelectric plants were constructed, and transmission lines capable of operating at 220 kilovolts (a kilovolt is 1,000 volts) or more were constructed in many parts of the world. In 1922 Pit River transmission line and Big Creek transmission line, both designed to operate at 220,000 volts, were placed in operation in California. In 1934 power generated at Boulder (later Hoover) Dam on the Colorado River was transmitted to Los Angeles, a distance of 270 miles (430 kilometres), at 287,000 volts.

High-voltage operation. Corona (electrical discharge from the surface of lines operating at high voltages) with its related phenomena is an important problem at 220,000 volts, and conductor diameter was increased to as much as an inch (2.5 centimetres) and more, irrespective of current, to minimize the loss of power resulting from this effect. During World War II, Germany made fundamental experiments both on ac extra-high-voltage (EHV) transmission (380 kilovolts ac) and EHV dc transmission (440 kilovolts dc). In EHV ac, to solve the problem of corona at 380 kilovolts, the Germans devised the multiconductor system, using four conductors. In EHV dc they manufactured the world's highest voltage mercury-arc rectifiers (for converting ac to dc) and constructed an underground dc cable transmission line some 60 miles (100 kilometres) long, between the Elbe hydroelectric station and Berlin. Tests were discontinued at the end of the war, but their experiences greatly affected technological development in EHV transmission.

AFTER WORLD WAR II

Large-scale projects. Post-World War II developments were characterized by the emergence of large capacity machines, especially steam turbines. The maximum sizes of steam-turbine generator units in the United States (the world's largest) were 260 megawatts (260,000,000 watts) in 1956, 450 megawatts in 1960, 650 megawatts in 1963, and 1,000 megawatts in 1965. These large units reduced construction cost per kilowatt for steam power stations; thermal power generation increased because of the low price of oil and coal and the exhaustion of favourable sites for hydroelectric generation in most developed countries (except the Soviet Union and Canada). This situation changed the role of hydroelectric generation in many countries. Formerly hydro had been the principal power source and steam power an auxiliary; then thermal power took the base load and hydropower the extra, or peak, load. Pumped hydro for peak use was common in the 1970s in many countries.

Soviet and Canadian hydro projects

Despite this trend, many large-scale hydroelectric power projects have been instituted in the Soviet Union and Canada. In the Soviet Union the world's largest hydroelectric station, Krasnoyarsk, with its total generating output of 6,000 megawatts (12 hydroturbines of 500 megawatts each) began operation in 1971, and the 6,500-megawatt Sayan–Shushenskoye station with its ten 650-megawatt hydroturbines was under construction in the early 1970s. In Canada, the Churchill Falls project, with an ultimate capacity of 5,200 megawatts, and the James Bay hydroelectric project (14,000 megawatts) were under consideration.

In the developing countries, the early 1970s saw many large-scale hydroelectric projects already completed or under construction: Kariba Dam (final 1,500 megawatts) in Rhodesia, Aswān High Dam (2,100 megawatts) in Egypt, Furnas (final 1,200 megawatts) in Brazil, and Guri (final 6,000 megawatts) in Venezuela.

After World War II, widespread fear that such conventional fuels as coal and oil would be exhausted led to the promotion of nuclear power as a new energy source. Thus, gas-cooled reactors were developed in the U.K., and light water-cooled reactors in the U.S. It was difficult, however, for nuclear power to compete in cost with power generation by conventional fuels, because from the end of the 1950s oil and coal continued to be abundant and cheap. By the mid-1960s, however, nuclear power had become competitive with conventional thermal power.

Nuclear power generation is described below (for further information see NUCLEAR FISSION).

Extra-high-voltage transmission. After World War II it became evident in many countries that, with the tremendously increasing demand for electric energy, high-voltage transmission would be needed to transmit ever larger amounts of power, and studies and experiments on demonstration lines with voltages of 500,000 and beyond were undertaken. In Great Britain a 275,000-volt transmission system (supergrid) was constructed to transmit large amounts of power from one part of the system to another, and the original extensive 132,000-volt British grid, construction of which had begun in 1926, was almost relegated to the role of subtransmission, or delivery to a distribution network. The first higher voltage line became operative in Sweden in 1954, a 380,000-volt line from the hydroelectric station at Harsprånget, 22 miles (35 kilometres) north of the Arctic Circle, to southern Sweden, 600 miles (970 kilometres) away. This line marked an advance in extra-high-voltage (EHV) transmission technology, with the first practical use of bundle conductors (actually double conductors in this case), previously studied by the Germans. The line's success revealed the possibility of raising transmission voltage even higher, to say 1,000,000 volts, without incurring corona loss. This line also first used large-capacity series capacitors on EHV lines, making long-distance transmission economical.

Later, in the United States, Canada, and Australia, 330,000-volt to 345,000-volt lines came into use, while Germany, France, Sweden, the Soviet Union, and other countries began 380,000-volt transmission. Next, 500-kilovolt lines were developed in the Soviet Union and the United States. In 1965 lines in Canada, from Manic dams to Montreal via Quebec, began operation at 765,000 volts, the world's highest voltage transmission line up to the early 1970s. Studies and experiments in many countries, however, aimed at 1,000 kilovolts and higher.

New interest in direct current. Direct-current transmission systems, after more than half a century, also began receiving renewed attention because of the development of high-voltage mercury-arc rectifiers. In 1954 an experimental dc submarine cable, operated at 100,000 volts and carrying 20,000 kilowatts of power, was put into service between the mainland of Sweden and the island of Gotland, 60 miles (100 kilometres) away. This installation was another triumph for Sweden in the EHV transmission field, and a Swedish manufacturer provided the high-voltage mercury-arc rectifiers for nearly all subsequent dc transmission projects in the world.

Submarine cables

Submarine cables under the English Channel from Dungeness in Kent, England, to Le Portel, a suburb of Boulogne, France, a distance of 30 miles (48 kilometres), began operation in 1961, carrying 200,000 volts dc with a total capacity of 160,000 kilowatts. In 1965 other systems began operation, with voltages ranging from 200 to 500 kilovolts and transmitted powers from 200 to 600 megawatts. Most of these projects used undersea cable, but the Volgograd–Donbass line in the Soviet Union (800 kilovolts, 750 megawatts, 290 miles [470 kilometres]) and Pacific Intertie in the United States (800 kilovolts, 1,440 megawatts, 825 miles [1,330 kilometres]) use overhead lines only.

New developments in generation and transmission. In the power-generation field, fast breeder reactors, which produce more fuel than consumed; magnetohydrodynamic power generation, in which the thermal efficiency may be more than 50 percent; and, finally, nuclear fusion reactors (see below) will emerge.

In the electric-power transmission field, cryogenic (supercooled and superconducting) and microwave systems (transmitting power by radio waves) appear to be feasible in the future, while 1,000,000-volt transmission seems feasible using present-day technology.

Pollution considerations. New problems, primarily environmental, emerged, however, which the utility industry must solve. From the middle of the 1960s, gaseous effluents, such as sulfur dioxide and dust, from fossil-fuelled power stations caused much public concern. Water from thermal stations that has been heated in cooling the condensers (*i.e.*, thermal pollution) has also caused alarm. Also, gaseous and liquid effluents from nuclear stations contain slight amounts of potentially dangerous radioactive substances, and the possibility of a major accident, releasing large amounts of radiation, drew critical attention. Pollution problems are described more fully below and in the article POLLUTION CONTROL.

Operating frequencies. In the history of the power industry, frequencies of ac generators (the number of oscillations per second of the ac output) were very diverse, ranging mainly from 16⅔ hertz (cycles per second) to 133 hertz. Generators with different frequencies, however, cannot be operated in parallel. After many years of efforts at standardization, two frequencies, 50 hertz and 60 hertz, remain. In the United States and Canada, 60 hertz is commonly used for general utility purposes; while on the European continent and in Great Britain, 50 hertz is the rule. Japan is an exception, both 50 hertz and 60 hertz being used nearly half and half. To link regions of different frequencies, a frequency changer using high-voltage mercury-arc rectifiers is needed. Lower frequencies such as 25 hertz and 16⅔ hertz are used for electric railways in many countries.

North American and European systems

II. Power generation

There are only two basic sources for driving electrical generators: hydro and thermal. Hydroelectric power is derived from generators turned by falling water. Most other electric power is produced by generators turned by steam pressure; the steam may be produced either by burning fossil fuel or by a nuclear reactor.

Hydroelectric power has many advantages. It is continuously renewable because of the recurring cycle of rainfall, runoff, evaporation, and transpiration. Hydroelectric power plants require neither fuel nor cooling water and create no thermal pollution problem. Their maintenance is relatively easy, and their plants, often located at remote sites, can be designed for automatic or remote-control operation. Their water-turbine generating units are more reliable than steam-electric units because they operate at relatively low speeds and are not subject to temperature stresses.

On the other hand, the steam power plant is not dependent on rainfall or water availability and can supply power during all seasons of the year. Site selection is much more flexible, and consequently the transmission line from the plant is usually shorter. Steam-electric units can be designed to burn several kinds of fossil fuels and to change from one fuel to another on short notice to take advantage of price changes.

In the United States, coal is burned to produce more than 60 percent of the electricity from fuel, while natural gas supplies more than 20 percent and oil about 10 percent. In the United Kingdom and West Germany, the share of coal is higher (more than 80 percent), whereas in Italy and Japan most of the electrical energy from fuel (more than 70 percent) comes from imported heavy oil.

Thermal-power countries

The proportions of the two forms of power also vary from country to country, depending on the respective availability of fuel and waterpower. In the United Kingdom almost all electricity comes from fuel (98 percent); in the United States, the Soviet Union, and West Germany, thermal power contributes more than 80 percent, and in Japan nearly 70 percent. These are called thermal-power countries. Canada, Norway, Switzerland, and Sweden are termed hydropower countries. In Norway 98.5 percent of electric power is hydro-generated, in Switzerland more than 90 percent, in Sweden 85 percent, and in Canada 80 percent. France and Italy stand between, with approximately equal quantities of hydro and thermal power generated.

In the early days of the power industry, hydropower was cheaper than thermal power and played the main role in supplying electricity, with thermal power an auxiliary. Since the 1930s, however, and especially since World War II, technology in thermal-power generation has progressed tremendously; since the late 1950s, fuel has been abundant and relatively inexpensive. Thus the

cost of thermal-power generation has decreased. Furthermore, in many developed countries, favourable hydro sites are becoming exhausted. Thus the proportion of thermal power to the total power generated rises year by year in many developed countries.

Though nuclear power has considerable potential, its actual contribution to total electric-power generation in the early 1970s was very slight (less than 2 percent) in many countries, except the United Kingdom, where its contribution was more than 10 percent (12 percent in 1969). At the same time, construction of nuclear power stations in such countries as the United States, West Germany, and Japan was being carried out vigorously, and by 1980 the share of nuclear-power generation was expected to amount to more than 20 percent.

DC POWER GENERATION

For industrial use, dc power has special advantages in some motor applications and in electrochemical and electrolytic processes. This power is generated locally in industrial plants, sometimes by engine-driven or turbine-driven dc generators but more frequently by motor-generator sets that combine an ac motor and a dc generator.

Since the late 1930s, dc power has been generated by such electronic devices as mercury-arc rectifiers that convert ac power into dc power (see ELECTRON TUBE). These generally have important advantages over motor-generator sets, such as lower construction cost, less weight, and less area required for installation. Maintenance and repair costs are usually simpler, and start-up is easier. Consequently, motor-generator sets have gradually become obsolescent.

In the 1950s the silicon-controlled rectifier (see also SEMICONDUCTOR DEVICES) emerged, and in the 1960s it supplanted the mercury-arc rectifier in industrial applications, except for high-voltage dc transmission, in which high-voltage, high-power, mercury-arc rectifiers are used. In the early 1970s, large-capacity silicon-controlled rectifiers were being employed experimentally in high-voltage dc power-transmission projects.

THERMAL POWER GENERATION

In the generation of electric power from fuel, the heat energy of the fuel is first converted into mechanical energy by a prime mover and then into electrical energy by a generator. Until about 1900 the only prime movers available to the industry were slow-speed steam engines, the ratings of which were limited to less than 2,500 horsepower. The introduction of the steam turbine broke the prime-mover bottleneck by permitting much larger concentrations of output in a single unit.

Continuous technical progress with boilers, steam turbines, and turbo-generators steadily improved the efficiency and reliability of the steam-electric process, enabling the industry to increase steaming capacity of the boiler and to increase the temperature and pressure of steam available at the turbine, which in turn led to larger turbine-generator sets and overall improvements in plant design. In the turbines the advances included fabrication of longer blades and higher efficiencies through improved steam condition; in the generators, they included increases in capacity in relation to generator size through hydrogen cooling and, later, water- and oil-cooling.

The thermal efficiency of steam electric generation, which was as low as 5 percent in 1900, increased to more than 40 percent in the 1960s. In terms of fuel consumption, this increase meant that an eightfold reduction had been achieved since the beginning of the century. These achievements are dramatically reflected in the rapid increase in the size of individual units. Figure 1 shows the history of increase of unit capacity of turbine-generator sets in the United States; they were the world's largest ones then available.

Steam temperature and pressure at the turbine throttle continued to increase steadily until the beginning of the 1960s. The steam condition of the Eddystone, Pennsylvania, station, in the United States, is the world's highest,

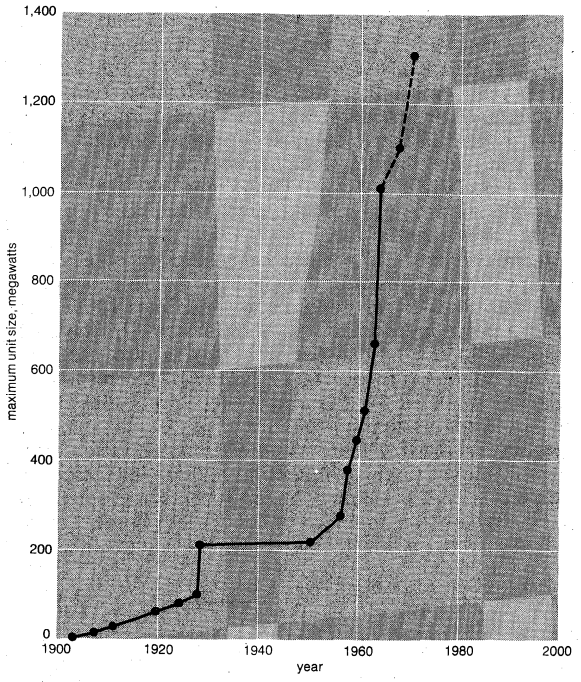


Figure 1: Growth in the unit size of electrical generating plants driven by steam power (U.S.).

with a temperature of 1,200° F (650° C) and a pressure of 5,000 pounds per square inch (350 kilograms per square centimetre). Since then the steam temperature has shown a downward trend, because adoption of temperatures as high as 1,200° F raised difficulties in finding satisfactory metals. In the early 1970s, the general practice was to use steam at temperatures of about 1,000° F (540° C) and at pressures of 3,500 pounds per square inch (245 kilograms per square centimetre).

This retreat halted the trend toward higher efficiency by improving the steam condition and caused designers to strive for higher efficiency by enlarging the individual unit. The largest units in operation in the early 1970s were in the U.S.; the No. 3 unit of the Ravenswood plant in New York, 1,000 megawatts (1965), and the No. 3 unit of the Paradise, Kentucky, plant, 1,130 megawatts (1970). The largest unit under construction was also in the U.S., the No. 1 unit of the Cumberland, Maryland, station, which is rated at 1,380 megawatts (see Figure 1). It appears highly probable that 1,500-megawatt or 2,000-megawatt units will be manufactured in the 1970s. Future developments in high-temperature metallurgy may make relatively cheap metals available for use at high temperatures, again permitting improvements of the steam condition for high efficiency.

Use of the binary cycle

Two other methods can be utilized to improve the efficiency of the thermal power station. One is to employ a binary cycle, the other is through the combined operation of a gas-turbine steam-turbine cycle. The binary cycle is a combination of a mercury-vapour cycle and a steam cycle. High-temperature mercury vapour can be attained at relatively low pressure, in contrast to steam; the increased initial temperature increases the overall efficiency. By the early 1930s, several power stations used binary cycles. Facilities were complicated, however, and many new technical problems arose in the design of mercury turbines and boilers, and so in the end construction of the binary-cycle plant using mercury was abandoned. For a description of the combined operation of gas-turbine steam-turbine cycle, see below and also the article STEAM POWER.

Some types of thermal power stations produce both electricity and heat. Heat in the form of process steam is used in the industrial plant, and heat in the form of steam and hot water is used for heating residences.

Where the industrial factory needs both electricity and process steam for treatment, drying, and heating of raw materials, it is frequently preferable to have a high-

pressure boiler producing high-pressure steam for electricity generation and to obtain process steam from bleeding steam or from the exhaust from the turbine, rather than to install separate boilers for electricity generation and for process steam. This type of dual-purpose plant is used in many countries for the paper and food industries, oil refineries, and chemical plants. Many countries, especially the Soviet Union and West Germany, have plants that produce both electricity and by-product steam for residential use.

Steam turbines. The electrical generator of a steam-turbine system is usually designed for a turbine speed of 3,600 revolutions per minute and sometimes 1,800 revolutions per minute in the 60-hertz system. In the early days, 1,200 revolutions per minute was also used, but this speed was rare by the early 1970s. Corresponding turbine speeds in the 50-hertz system are 3,000, 1,500, and 1,000 revolutions per minute.

Steam turbines for thermal power stations fall into two classifications: single cylinder and multicylinder (see also TURBINE). In the single-cylinder turbine the steam expands from boiler pressure to condenser pressure in one cylinder. Such a turbine is shown in Figure 2A. This

Reprinted from *Principles of Electric Utility Engineering* by C.A. Powel by permission of The M.I.T. Press, Cambridge, Massachusetts. Copyright 1955 by M.I.T.

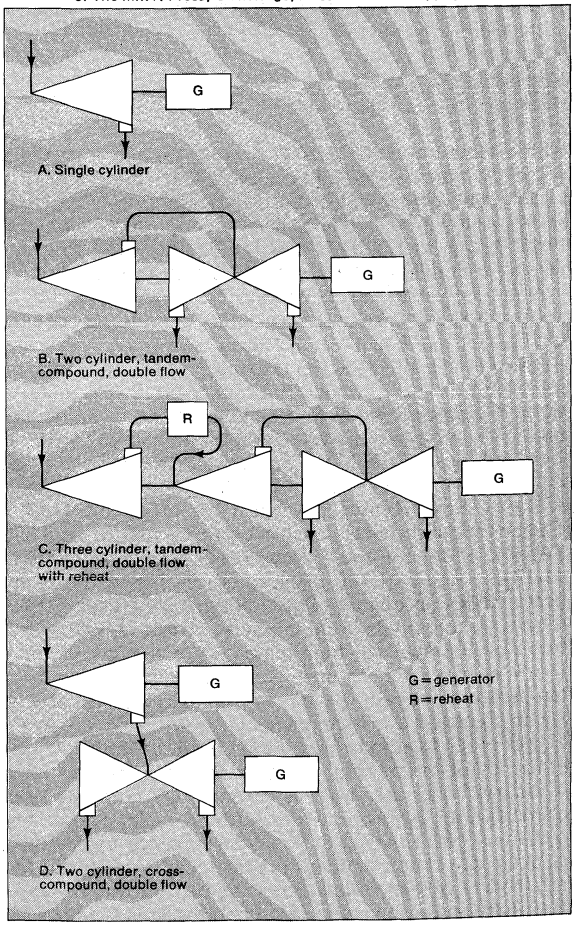


Figure 2: Typical turbine arrangements in steam plants (see text).

kind is always connected to a single generator. In the multicylinder turbine the steam expands from boiler pressure to condenser pressure in two or more cylinders, which may be tandem-compound (single shaft) or cross-compound (multi-shaft). Figure 2B shows a two-cylinder tandem-compound turbine. Steam from the high-pressure cylinder is conducted into the low-pressure cylinder, which has a double exhaust. With a given maximum peripheral speed, it is the blade length in the last row of blades that determines the turbine output. Practically, the length of the last blade is limited by the mechanical strength of the metal of which it is fabricated.

Multi-cylinder turbines

Figure 2C shows a tandem-compound, three-cylinder, double-flow turbine. The steam on leaving the first cylinder is returned to the boiler and reheated (R) to its initial temperature before being fed into the second medium-pressure cylinder.

Cross-compound turbines also consist of a high-pressure cylinder and one or more low-pressure cylinders, but each is connected to its own generators as shown in Figure 2D. Sometimes a high-pressure turbine generator runs at 3,600 revolutions per minute and the low-pressure unit at 1,800.

As mentioned above, the length of the last-stage blade in the low-pressure turbine is the limiting factor for output increase. The 1,100-megawatt turbine at the Paradise plant can be expressed as CC4F-52″; that is, a cross-compound four-flow turbine with the last blade being 52 inches (1.3 metres) long. A large-capacity turbine is often of the cross-compound type (multishaft); the largest under construction in the early 1970s was 1,380 megawatts. The largest capacity of the tandem-compound turbine (single shaft) was around 800 megawatts.

Electric generators. Generators in thermal power stations are two-pole and four-pole machines (3,600 or 1,800 revolutions per minute at 60 hertz), direct-connected to the steam turbines. In cross-compound units, two or more generators constitute one unit and are treated electrically as a single machine.

Rating electric generators

Generators are rated in kilowatts and kilovolt-amperes at a certain power factor and short-circuit ratio. The rating and dimensions of generators are affected by all these factors and by the pressure of the hydrogen used for cooling. Standard generators are rated at 0.85 power factor, 0.8 short-circuit ratio, and 0.5 pound per square inch (0.035 kilogram per square centimetre) hydrogen pressure. The power factor is the ratio of the real power to the apparent power, that is, kilowatts to kilovolt-amperes. The short-circuit ratio is the ratio of the field current required to generate rated voltage at no load to the field current required to circulate rated armature current through the short-circuited armature, both at rated frequency.

The effect of the short-circuit ratio on generator rating is not as simple as that of power factor, but in general the higher the short-circuit ratio, the larger the machine; so that for a given physical size, the lower the short-circuit ratio, the higher the generator's rating. In practice, the short-circuit ratio is seldom less than 0.8. When a generator is short-circuited, the initial current is very high but drops rapidly and after one or two seconds stabilizes at a final value determined by its synchronous reactance. Synchronous reactance is the value of reactance that exists in the machine after all transient effects have died out.

The purpose of a high short-circuit ratio is to keep the machine in synchronism with the system during a disturbance, because the higher the short-circuit ratio, the lower the synchronous reactance. With improved voltage-regulating systems, however, there is a tendency to adopt a short-circuit ratio lower than 0.8, and a value as low as 0.6 has been adopted in some large-capacity generators.

The steam-turbine-driven generator has a far higher speed of rotation than a hydroturbine-driven generator; its rotor is cylindrical, and usually the shaft is horizontal. As the output of a turbo-generator becomes higher, the cost per kilowatt decreases. Thus there is a general trend to adopt higher output machines. Under ordinary conditions, however, it is preferable to adopt standard-output machines, because the cost is low, delivery time is relatively short, and they are rather flexible in application. The limiting factor in increasing unit size is the rotor. Because of the high rotational speed, centrifugal force limits the diameter of the rotor, necessitating an increase in length for higher ratings. To attain high strength, the rotor is usually forged in one piece, out of carbon steel for small machines, nickel steel for medium sizes, and high-tension steel for high-capacity units.

Developments of materials and fabrication techniques have contributed much toward the larger capacity ma-

chines, but most important have been improvements in generator cooling techniques. Either natural cooling or forced cooling by fans was once employed, but from the early 1950s hydrogen-cooling was widely used, phenomenally increasing capacity. Later, the introduction of direct cooling methods with water or oil made it possible to design and fabricate even higher capacity machines.

Nearly all turbo-generators above 20-megavolt-ampere capacity are hydrogen-cooled, thus necessitating building them with gas-tight housings and facilities to prevent formation of explosive mixtures with air; the advantages are such that the extra complication is justified. The effect of hydrogen-cooling on the rating is to reduce the losses for a given temperature rise. Because of the lower density of hydrogen, windage losses (frictional losses caused by gas around the armature) in the generator are reduced to one-tenth of what they would be with air at atmospheric pressure. Hydrogen has a thermal conductivity seven times that of air, and the rate of heat transfer in hydrogen is 35 percent greater. The effect of corona is negligible in hydrogen, insulation retains its flexibility longer, and, because hydrogen will not support combustion, the fire hazard is reduced.

In the past, machines were designed for operation at a gas pressure of 0.5 pound per square inch (0.035 kilogram per square centimetre), a design permitting a rating about 25 percent higher than for the same machine with air-cooling. Later, hydrogen pressure was raised to 15 pounds per square inch (one kilogram per square centimetre), giving about a 15 percent output increase, and then up to 30 pounds per square inch (2.1 kilograms per square centimetre) with a further 10 percent increase in output. Gains fall off at higher pressures, so that 30 to 50 pounds per square inch seems to be an optimum.

Direct cooling

A new design, the so-called direct-cooling, or inner-cooling, method, has emerged. The rotor conductors are hollow, and high-pressure hydrogen (45 to 90 pounds per square inch, or 3.15 to 6.3 kilograms per square centimetre) circulates through them. In the stator (stationary part of the turbine), conventional copper strands are used, but between them a thin-walled copper tube runs the length of the slot, and hydrogen is forced through it. Thus, in both the rotor and stator the hydrogen removes the heat directly from the copper conductor instead of through the insulation. The output of a machine is about doubled by this method of cooling. As a medium for this direct-cooling method, water is also used. The Table shows the general trends in cooling methods used for different machine sizes.

Cooling Methods at Different Machine Sizes

rating (megavolt-amperes)	stator	rotor
Up to 200	hydrogen cooling	hydrogen cooling
Up to 350	hydrogen cooling	direct cooling (hydrogen)
Over 400	direct cooling (water)	direct cooling (hydrogen)

Site selection. A most challenging problem repeatedly facing each electric utility is the location of large thermal power stations to provide the desired increase of electric-power supply at the lowest cost. Important factors in locating the plant include fuel supply, load distribution, load growth, availability of land, foundation conditions, and availability of cooling water. Utilities have also had to consider thermal- and air-pollution problems. Generally, availability and cost of fuel are basic to the location of a large generating station.

The number of favourable sites available for large generating stations is decreasing with the increase in population, expansion of the economy, and the more active interest of the general public as well as local public agencies in community matters, and thermal and air pollution. Consequently, electric utilities try to use each site selected for the largest generating plant that can be economically justified.

Most thermal plants are located on rivers, lakes, or the seashore because of the need for an adequate supply of cooling water to condense the steam leaving the turbines. The amount of cooling water required depends upon the size of the plant, its thermal efficiency, and the permissible temperature rise of the cooling water passing through the condenser. For a 1,000-megawatt generating station, a flow of 1,500 cubic feet (42 cubic metres) per second is required, and the minimum river flow would have to be about 4,000 cubic feet (110 cubic metres) per second to limit the temperature rise of the river water to 7° F (4° C).

A shortage of cooling water can be met in two principal ways: by recirculating the water in reservoirs or ponds and by the use of cooling towers, both evaporative and nonevaporative. Each method requires a supply of make-up water to replace water evaporated and otherwise lost. For a 1,000-megawatt generating station, the water requirement for a cooling pond would be 10 to 20 cubic feet (0.3 to 0.6 cubic metre) per second; for an evaporative cooling-tower system, 20 cubic feet (0.6 cubic metre) per second; and for a nonevaporative cooling tower system, about two cubic feet (0.06 cubic metre) per second.

When an ample supply of cooling water is available, the once-through circulation method for condenser cooling is usually the most economical; the recirculation system involving reservoirs and ponds ranks next in economy. The cooling-tower system is relatively expensive, especially the nonevaporative, or dry, cooling-tower system.

Air pollution concerns in site location

Air-pollution problems also pose severe limitations in locating large plants. Special studies must be made of atmospheric conditions in the area, the sulfur content of the fuel burned, and the degree to which the area is developed. Even with modern boiler design, dust collectors, and proper plant layout and stack height, the station size may sometimes be limited by the tolerable pollution level.

Air pollutants of major concern at fossil-fuelled power plants include fly ash and polycyclic hydrocarbons in the particulate group, and the oxides of sulfur and nitrogen in the gases. Modern power plants are normally equipped with mechanical and electrostatic precipitators to provide a substantial degree of control over fly ash. The sulfur-derivative gases are generally the most troublesome pollutants from fossil-fuelled plants, sulfur being present in almost all fossil fuels. Most of the sulfur appears in the stack gases as sulfur dioxide in a concentration that is a function of the sulfur content of the fuel and the amount of excess air in the stack emission.

Removal of sulfur dioxide from combustion gases by techniques known in the early 1970s was difficult and expensive, and experimental work was underway for better methods. As a control, utilities have constructed very high stacks at thermal power stations, in some instances as high as 825 feet (250 metres) or more, to allow the wind to dilute the stack gases. This method has disadvantages, however, because higher stacks can cause nuisance problems covering wider areas.

The search for fuel of low sulfur content was active in the 1970s in many developed countries, but the supply is limited. Instead of extracting sulfur dioxide from stack gas, a technique of extracting sulfur directly from oil was being developed with promising results, but the cost was uncertain. Similar experiments were being conducted on coal.

Gas-turbine stations. The gas turbine, which is widely used as an aircraft engine, was first applied in electric-power production when a 4,000-kilowatt machine was built in Switzerland in 1939. In the 1960s the growing use of turbojet engines for aircraft brought both capacity and performance improvements, and gas turbines found growing application in the electric utility field.

The principal advantages of gas turbines are low construction cost, short construction period, flexibility in the choice of site location, quick start-up, and ease of automation. There are two types of gas turbines, the open-cycle, in which the output is exhausted directly to the atmosphere, and the closed-cycle, in which the exhaust is recirculated back to the input. Although the

open-cycle has lower thermal efficiency, it is preferred because the closed-cycle is more costly and more complicated in layout and in operation. A capacity of 1,000 megawatts is difficult to attain in a single gas-turbine unit, but the unit capacity of gas turbines is higher than that of diesel engines, the output of which is limited to the order of several megawatts.

Single units of the open-cycle type were available in the 1970s in ratings up to about 30 megawatts, and gas-turbine units of larger ratings had been developed by using several aircraft jet engines as gas generators. One design employs several jet engines in tandem and produces an output of 120 megawatts. Another design consists of ten cylindrically mounted aircraft gas turbines operating as gas generators exhausting simultaneously into a single-stage power turbine driving a 100-megawatt generator.

Because the construction period of gas-turbine stations (12 to 18 months) is far shorter than that of fossil-fuel steam-turbine stations (three to five years) and of nuclear power stations (four to seven years), and because the cooling water requirements are far less, gas-turbine stations are popular for building up a standby reserve of generating capacity to use in meeting peak load demands. Gas turbine units can be started, synchronized, and fully loaded in two to five minutes.

Disadvantages of gas turbines

Gas turbines have some disadvantages. Open-cycle units have low thermal efficiencies, around 22 to 24 percent, and are limited to high-grade liquid fuels or gas, which are costly in many areas. As a result, energy costs are high. In most cases, therefore, use of gas turbines was limited to annual plant factors not exceeding 5 to 10 percent; that is, total operating hours of a gas turbine per year did not exceed 500 to 1,000 hours. Construction delays with thermal and nuclear power stations in the U.S. in the early 1970s contributed to the introduction of many gas turbines in the power system. In some installations the plant factor was as high as 20 to 25 percent in 1969 to cover what might otherwise have been a severe supply shortage. The rate of increase was expected to taper off as thermal and nuclear stations under construction and being planned were brought into operation.

Another method of application for gas turbines, however, involves combining them with steam turbines to increase overall plant efficiency as much as 4 percent. Exhaust gas of the open-cycle gas turbine is as hot as 750° F (400° C), and also retains 80 percent of the oxygen in the air. This exhaust gas is used as the preheated air to be injected into the furnace of a steam boiler. The first application of this combined cycle to a utility system was at the Horseshoe Lake station, Oklahoma City, Oklahoma, in 1963. Capacity of the combined unit was 245 megawatts and proved the applicability of the technique. Later, this combined cycle was used in plants in several countries.

Diesel-engine generators. The diesel engine as a power source for generating electricity has a number of advantages: relatively small size, light weight, high reliability, quick start-up (60 to 90 seconds), and high thermal efficiency (30–38 percent). Unit capacity, however, is small, ranging from several hundred to several thousand kilowatts and so cannot be used as the primary generating source for central power stations. Because it is a reciprocating engine, its rotational speed fluctuates during one revolution. To reduce this effect, heavy flywheels are needed.

Nuclear power generation. Like fossil-fuel-fired plants, nuclear power stations generate heat to produce steam to drive turbine generators. The difference is that, whereas conventional thermal plants use heat produced by combustion of fossil fuel, nuclear plants produce heat by fission of nuclear fuel in a reactor. Basically, therefore, a nuclear reactor is substituted for a fossil-fuel boiler (see also NUCLEAR FISSION).

In 1964 the order for the nuclear station at Oyster Creek, New Jersey, was the first to be placed on a competitive basis with fossil-fuelled plants and introduced competitive nuclear power. Later, orders from utilities for nuclear plants in the United States grew phenomenally, and nearly 100 reactors with an aggregate capacity of roughly 85,000 megawatts were ordered by 1971. This success exerted worldwide influence; United States-type reactors were introduced into the power systems of other countries, especially in West Germany and Japan.

Reactor types. Fuels, moderators, and coolants are the major components of reactors, and from their combination can be theoretically considered the many varieties of nuclear reactor types. The number of reactor types actually developed, however, is rather limited, namely light-water reactors in the United States, gas-cooled reactors in the United Kingdom and France, and heavy-water reactors in Canada. Of these three types, the most developed is the light-water reactor. Gas-cooled reactors developed in the United Kingdom (often called Magnox type) have long had excellent operation. In total generated kilowatt-hours by reactors in the early 1970s, the United Kingdom surpassed all other countries. The modified version of the Magnox type, the advanced gas-cooled reactor, was under construction in the United Kingdom in the early 1970s. In 1971 France, which also had developed gas-cooled reactors, adopted instead the United States light-water technology. Canada, uniquely, has consistently developed the heavy-water type, but the total number and capacity of such reactors already developed and under construction was still modest in the early 1970s.

High-temperature gas-cooled reactors, advanced versions of the gas-cooled reactor, were being enthusiastically developed by the United States, United Kingdom, and West Germany. Fast breeder reactors are deemed the ideal reactors for the future, and much developmental effort was being devoted to them by the developed countries. In the early 1970s, leaders in nuclear reactor development were the United Kingdom, the Soviet Union, and France, followed by the United States, West Germany, Japan, and Italy.

Special features. In nuclear power generation, fuel consumption is tens of thousands of times less than in conventional steam power generation, with obvious savings in transportation and storage costs. In general terms, the nuclear power plant has a higher construction cost and a lower fuel cost than a thermal power plant of comparable size; fixed cost is one-third and fuel cost two-thirds, whereas with a nuclear plant the fixed cost is two-thirds and fuel cost is one-third. To obtain cheap electricity from the nuclear plant, construction cost must be kept down, and to this end, larger sizes were being planned and built. The average capacity of nuclear reactors ordered in the early 1970s ranged from 700 to 800 megawatts, and reactors with a capacity as high as 1,100 to 1,200 megawatts were not uncommon.

Fossil fuel versus nuclear power

The spent fuel, or burned-up fuel of nuclear power stations, contains newly born nuclear fuel, plutonium, and partly burned uranium and also highly radioactive fission products. To extract plutonium and partly burned uranium for re-use and to separate fission products for disposal, the burned-up nuclear fuel must be sent to a spent-fuel reprocessing plant. The normal operation of a nuclear power plant also produces minute quantities of gaseous, liquid, and solid radioactive wastes, some of which are released to the environment. National legislation in all nuclear-power countries is designed to control the resultant radiation problems. Operational experience in the 1970s showed the radiation dosage in all localities well below the maximum tolerable values as specified by the International Commission on Radiological Protection (ICRP). The safety review of nuclear power stations, which should be verified by authorized bodies, covers the whole phase of design, construction, and operation. The reason for this strict measure lies in the tremendous amount of fission-product inventory in the reactor. There is a possibility, however remote, that a widespread radiation hazard could occur if a significant part of the fission-product inventory should be released by accident.

Governments and industries have acted in three principal ways to avoid serious consequences: (1) enacting safeguards to prevent or reduce to minor levels accidental release of hazardous radiation and radioactive products

within the reactor structure; (2) building containment structures and other safeguards to confine accidental release of radiation inside the plant; and (3) enacting zoning measures to prevent radiation and radioactive products that may escape inadvertently, despite multiple safeguards, from causing damage to persons and to property.

Nuclear plant insurance

Liability insurance is a special problem for nuclear power plants. Accidents have been few and damage small, so that there is little statistical basis for establishing risks. Meanwhile, conceivable liability risks have been considered to exceed the financial capabilities of the private insurance business. In 1957 and again in 1967 the United States Congress authorized the federal government to provide liability protection up to $500,000,000 for any single nuclear accident. Utilities are required to secure the prescribed amount of private liability insurance up to a limit of $60,000,000, depending on the size of the plant. In other countries, liability insurance for nuclear power plants is treated similarly, but the amount of insurance varies.

The number of economically exploitable uranium mines is limited, and the main uranium-producing countries are the United States, Canada, Australia, and South Africa. World assured reserves of uranium that can be exploitable at less than about $10 a pound of U_3O_8 is estimated at about 800,000 tons. Many countries were actively prospecting for additional uranium reserves in the 1970s.

Natural uranium reactors—that is, reactors that do not use enriched fuel—were being actively developed in only a few countries, principally Canada. The United Kingdom and France had shifted from natural to enriched uranium, which was provided solely by the United States and which was used in almost all reactors.

Thermonuclear power. Nuclear power generation uses the physical process called nuclear fission. In the thermonuclear reaction, or fusion, very light nuclei, such as deuterium or tritium (both isotopes of hydrogen), join together with the release of tremendous quantities of energy (see NUCLEAR FUSION). This process, which takes place in the thermonuclear, or hydrogen, bomb, had not yet been controlled in the early 1970s so that it could be used to generate electric power, but the potential exists. The future of fusion reactors, however, was far from clear; optimistically it was felt that electric power generation by fusion might be possible in 1990 or 1995. Most authorities felt that the fusion reactor would be the technology of the 21st century.

Advantages of the fusion reactor

The advantages of the fusion reactor would be very great. First, fuel resources would be no problem, because deuterium exists in almost unlimited quantities in ordinary water, and tritium can be obtained from lithium, a relatively abundant element. Second, radioactive waste is far less from the fusion reactor than from the fission reactor. Third, a serious accident in the fusion reactor holds no danger of discharging large amounts of radioactive material into the environment, as do fission reactors. The successful fusion reactor would therefore be an ideal energy source, and research and development were actively being pursued in the 1970s in many countries, primarily the United States, Soviet Union, United Kingdom, West Germany, France, and Japan.

HYDROELECTRIC GENERATION

Falling or flowing water formed the earliest of the natural sources of energy to be used for power, but, as has been noted, most of the waterpower in use in the second half of the 20th century was developed from the first decade of the century. Conditions favouring such development included the availability of sites, the higher efficiency in power generation promised by hydroelectric sources as compared with steam plants, and the advent of the technology for high-voltage, long-distance transmission. The hydraulic turbine made it possible to use any head (height of water above the turbine) from 30 to 5,000 feet (nine to 1,500 metres) with high efficiency and at speeds that permitted use of a relatively inexpensive electric generator. Developments in electrochemical and electrometallurgical processes requiring large amounts of cheap electricity created a demand that could under favourable circumstances be satisfied more readily from waterpower plants than from other sources.

Eventually, however, most favourable hydroelectric sites were exhausted. Meantime, construction cost of steam plants was reduced, their efficiency improved, and their fuel costs ameliorated by new transportation techniques, such as pipelines. The trend was consequently reversed, and steam power plants began to overtake and surpass hydroelectric stations in numbers and capacity. In the second half of the 20th century, almost all hydroelectric construction was government-sponsored, with the projects including such other considerations as navigation, flood control, irrigation, city water services, and conservation of resources.

To increase the load factor of thermal power stations, it is necessary to provide peak power by other means. For this purpose pumped-storage facilities are useful.

Hydroelectric site selection. The possibilities of hydroelectric development depend on weather conditions and geographical features. Mountainous regions subject to heavy rainfall close to industrial regions are most favourable for hydroelectric development.

Water may be collected and stored at a high elevation and led through tunnels or pipelines to a station at a much lower elevation, or dams may be erected across a river to raise the level of the water. The power developable by a volume of water is directly proportional to the working head, so that a high-head installation requires a smaller volume of water than a low-head installation to produce an equal amount of power.

Head types

The distinction between high head, medium head, and low head is not clearcut, but generally heads from 20 to 100 feet (six to 30 metres) are referred to as low head; from 100 to 660 feet (30 to 200 metres) as medium head; and those over 660 feet as high head.

Low-head plants are usually associated with dams and high-head plants with penstocks, conduits for carrying water downstream from a considerable height. Between are medium-head plants that may use either scheme. Pipelines and turbines for high-head plants are smaller and the installation in general is cheaper. The volume of water that must be stored during wet weather to allow continuous operation throughout the year is also smaller.

High-head sites offer many advantages, but they are not essential for the successful development of hydroelectric energy. Most of the world's waterpower sites that are reasonably close to industrial communities are in regions in which the stream gradients are medium. The total power developed from medium- and low-head plants is far greater than that from high-head plants.

Water flow in the river is always changing; the ratio between the flow at floodtime and the dry weather flow usually lies between 10 and 100. A hydroelectric plant the capacity of which is only equal to the dry weather flow can provide a continuous output year round, but its capacity is very small, and most of the river flow is wasted. It is usually more sound economically to install a hydro plant of a capacity two or three times greater than that corresponding to the dry weather flow and add a steam plant capable of maintaining the output when the water supply is inadequate.

Hydroturbines. There are five types of turbines for hydroelectric station use: Pelton wheel for high-head use; Francis turbine for medium-head use; Deriaz turbine for medium- and low-head use; and Kaplan and propeller turbines for low-head use (see also TURBINE). Selection of the appropriate type depends largely on the available head. In general, Kaplan turbines are used for heads up to 100 feet (30 metres), Francis turbines for heads from 70 to 1,000 feet (21–300 metres), and Pelton wheels for heads between 300 and 3,000 feet (90–900 metres). In the overlapping ranges, the choice depends upon the size of the installation and other considerations.

Hydroturbines are divided into two classes: impulse (in which the turbine blades reverse the direction of the fluid flow) and reaction (in which the blades travel at about the same velocity as the fluid flow). Pelton wheels are impulse turbines, and the other four types are reaction turbines. Francis and Deriaz turbines are of the mixed flow

type with inward radial flow and axial discharge; the former have runners (rotating parts of the turbine) with fixed blades, the latter with adjustable blades. Propeller and Kaplan turbines are of the axial flow type; the former have runners with fixed blades, the latter with adjustable blades.

The design of a waterwheel for any conditions of head and output is based on an empirical value known as the specific speed. Under conditions of the same output and same head, Kaplan and propeller turbines have the highest speed, Francis and Deriaz turbines are next, and Pelton wheels have the lowest speed.

Turbine efficiencies

The efficiencies of hydroturbines at full load are high and run between 90 and 94 percent for all types; but under part load, efficiencies vary considerably. At one-quarter load, for example, efficiencies of Francis and propeller turbines, with fixed blades, fall within 60 to 70 percent, whereas those of Kaplan and Deriaz turbines, with movable blades, and of Pelton wheels range from 80 to 85 percent.

In case of a fault in the electric power system (*i.e.*, transmission line), the load on a hydroelectric generator may drop suddenly from full load to no load. Because the gate controlling the water flow to the turbine is wide open, a large quantity of water is admitted, and when the load is removed the turbine speed rises until the gate is closed. The runaway speed, the maximum speed that the unit would attain under these conditions, is about two times normal for Francis turbines and three times normal for Kaplan turbines. It is controllable to some extent, however, and the rotors of the generator are usually designed to withstand 180 to 200 percent of normal speed regardless of the types of hydroturbines employed.

The Pelton wheel is a slower-running machine than the reaction turbine and is better suited for very high heads. In addition, because the water is discharged through its one, or at most two, nozzles, these may be of reasonable size when handling the small water volumes normally available in high-head plants.

Pelton wheels, most of which have a horizontal shaft and one or two nozzles, are built in sizes up to 40,000 kilowatts. Speed usually is controlled by a deflector that rapidly cuts off the water jet from the wheel, acting with a central needle in the nozzle that slowly reduces the flow through the nozzle while the deflector returns to its normal position.

The reaction turbine with its full peripheral admission of water is well adapted to using the large water volume necessary in low-head plants. Its higher speed of rotation enables the cost of the electrical generators to be reduced.

Francis turbines are the most extensively used, because of the very wide range for which they are suited, from the lowest economically usable head to high head. They can have either horizontal or vertical shafts, the latter for machines of larger capacity. Very large Francis turbines have been installed with outputs exceeding 200,000 horsepower from a single runner, and outputs of 500,000 horsepower were under construction in the 1970s.

Deriaz turbines, with movable blades, provide improved part-load efficiency for medium-head machines. This design has proved very useful for pumped storage applications, in which significant variations in head and flow for turbine and pump operations exist.

A fixed-blade propeller turbine has higher specific speed and so saves in plant cost compared with the Francis turbine. The low efficiency of the propeller turbine at part-load operation can be compensated for in a multi-unit station by changing the number of operating units as the load fluctuates.

The Kaplan turbine, with adjustable blades, has overcome the deficiencies of the propeller turbine and is very useful for wide flow changes and large head variation.

Turbine generators. Electric generators used with water turbines are of two types: the two-bearing design, in which the thrust bearing and one guide bearing are located above the rotor, with the second guide bearing below the rotor; and the umbrella design, in which a single combination thrust and guide bearing is located below the rotor. The umbrella design has many advantages over the two-bearing design, including reduced cost because the second bearing is eliminated, and less head room is required under the crane for dismantling.

Umbrella generator design

All modern large- and medium-capacity waterwheel generators are installed with a totally enclosed recirculating ventilation system that gives longer insulation life because of absence of dirt, reduced noise level, and lower powerhouse temperature. In small sizes, simple open-type machines are often used.

Of the two common types of ac generators, the synchronous (fixed field, constant speed) and the induction (rotating magnetic field) types (see ELECTRIC MOTOR), the former is used exclusively in many systems and predominates in all systems. The induction type is more rugged, however, and requires fewer controls and no direct current for excitation. Thus the induction generator is sometimes used in remote-controlled and automatic stations.

Automatic stations. The automatic generating station is especially suited to systems in which many small power stations are used. In such circumstances the operating staff expense at each station would be prohibitive, but making each station automatic and linking all into a common distributing system makes the operation economic.

The waterwheel of such an installation has no governor, the gates being preset at some opening corresponding to the desired output of the generator. Suitable control circuits shut down the station in case of overload, overspeed, or any other kind of trouble. Normally the starting and stopping of each generator set is accomplished automatically by the response of float switches actuated by the change in the level of water above the dam. In remote-controlled stations, provisions are made for controlling these operations, as well as the gate openings of the individual turbines, from the central station.

In small stations, induction generators are sometimes used because of their simplicity. They require no exciter, no voltage regulator, and no synchronizing equipment. The disadvantage is that they take their excitation from the system in the form of reactive current, thus lowering the system power factor. They are usually small in number, however, and their advantages often outweigh this disadvantage.

Pumped storage plants. In ordinary hydroelectric power stations, hydraulic turbines and electric generators are the main mechanisms, whereas in pumped storage plants there are also hydraulic pumps and electric motors that drive the pumps.

In such plants, at times of maximum river flow or during off-peak periods, water is pumped into a storage reservoir on a neighbouring height, or simply back from a lower pool (tailrace) into the headwater pool; and at times of low river flow or during peak power periods, this water is used to generate electricity. The storage may be for lengthy periods to offset seasonal shortages or for short periods to compensate for daily peaks.

The world's first pumped storage plant, the Lettern plant near Zürich, Switzerland, began operation in 1892. The station at Vivarone, Italy, in 1912 was the first pumped storage plant to adopt the combined motor-generator system (using the same electrical machine both as a generator and as a motor). These plants are for seasonal regulation.

Later, in order to use effectively the reserve of thermal power generation during the late evening and early morning hours, and to cope with the changing electric load, pumped storage plants for daily or weekly regulation were constructed.

Pump-turbine units

In 1931 the pump turbines emerged, combining the functions of pump and turbine; in such pumped hydro plants as Lago Baltone (Francis turbine), Italy, and Baldeney (Kaplan turbine), West Germany, they were adopted for the first time. Later, in the United States, large-capacity Francis-type pump turbines with high efficiency both as pumps and as turbines were developed and have been widely used.

In a pumped storage plant, the generation of two kilo-

watt-hours of electric energy requires about three kilowatt-hours of off-peak energy because of losses from pumping, power generation, and transmission to and from the load centre. Despite the inefficiency in pure energy terms, however, such facilities are of great practical efficiency because they allow the use of otherwise wasted off-peak power.

III. Transmission

TYPES OF TRANSMISSION SYSTEMS

Direct current systems. The first electric systems operated on two-wire dc. In 1883 a three-wire dc system was introduced and thereafter commonly adopted. The three-wire system has an economical advantage over the two-wire system, as three conductors in the former system can carry the load equivalent of two circuits of the latter—that is, of four conductors. Its inherent disadvantages, however, led very early to replacement by the more flexible ac system. The only exception was the Thury system in France, which provided high-voltage direct current by connecting a number of direct-current machines in series at the sending and receiving ends. Several generators were placed in series to provide 57,000 volts dc for a 112-mile (180-kilometre) line from Montiers to Lyon, which was constructed in 1905. This line remained in service until 1937.

Because of its inherent simplicity and economy, dc transmission made a comeback, as has been seen, when mercury-arc rectifiers became available to convert the direct current into alternating current (necessary because of the number of individual consumer devices requiring the ac form of power to operate). In 1937 the General Electric Company in the United States operated an experimental dc line in western Massachusetts at 27,000 volts, using mercury-vapour converter tubes. In 1939 Brown Boveri Company operated an experimental dc line at the exposition in Zürich, Switzerland, using mercury-arc rectifiers; this experiment served as the prototype of later dc transmission systems.

During World War II, the Germans constructed an experimental dc transmission line from the Elbe power station to Berlin, using a voltage of ± 200 kilovolts (400 kilovolts between terminals), and transmitting 60 megawatts of power. Operation was discontinued at the end of the war.

After the war, dc transmission was actively studied in Sweden for transmitting large blocks of power from north to south. In 1954 the world's first dc transmission line, between the Swedish mainland and Gotland, began operation, using a submarine cable about 60 miles (100 kilometres) long. The power transmitted was 20,000 kilowatts at 100 kilovolts. The excellent performance record of this scheme demonstrated the feasibility of the new mercury-vapour tubes for high-voltage rectification (ac to dc) and inversion (dc to ac).

In 1961 a second underwater dc transmission project was placed in service between England and France, transmitting 160 megawatts at ±100 kilovolts. Other dc projects using underwater cables soon followed, including the Konti-Skån (between Sweden and Denmark); the Sardinia–Italy; the New Zealand (between the South and North islands); and the Vancouver Island project of Canada, all of which were in operation in the early 1970s. Projects using underground cables and dc transmission, all in operation in the early 1970s, included the Kashira–Moscow line, in the Soviet Union, and the Kingsworth–London line, in the United Kingdom. Overhead dc transmission lines are used on the Volgograd–Donbass line, in the Soviet Union, and the Pacific Intertie, in the United States.

Relative costs of ac and dc

Insulation requirements for a nominal voltage are less with dc than with ac, and only two conductors are needed per dc circuit as compared with three for ac. The consequent saving in materials accounts for the cost of a dc line being about two-thirds that of an ac line with comparable capacity. The terminal facilities, however, for converting between ac and dc are a major cost. While high-voltage ac transmission requires transformers at each terminal, dc requires similar transformation plus a reactive power supply, converters, and related controls. The savings in dc line cost and lower transmission losses must offset the additional cost of the terminals to make dc transmission as economical as ac. As a result, overhead dc lines must be several hundred miles long to obtain an economic advantage.

An insulated cable has the characteristic of storing electrical energy, because of the high electrical capacity or capacitance between the conductor and ground or between the conductor and the conduit in which it is placed. The current magnitude required to charge this capacitance is very nearly proportional to the length of the cable. With ac, the full current-carrying capacity of the cable may be required merely to charge the insulation, especially if the cable is fairly long. This requirement imposes a limitation on the length of underground and underwater circuits when ac is employed. In contrast, a dc circuit maintains an essentially constant voltage and thus needs no charging current in the normal operation so that no similar distance limitation applies. In addition, the insulation can be subjected to much higher voltage when used for dc. For these reasons, a given cable can carry two to ten times as much dc power as ac. Therefore, for underground or submarine cable lengths of 25 miles (40 kilometres) or more, dc transmission is more economical than ac.

Alternating-current systems. An ac circuit has three characteristic properties: resistance, inductance, and capacitance. Strictly speaking, conductance (leakage over the insulators, corona discharge) must also be taken into consideration, but this property can often be ignored in well-designed lines. In a transmission system the three main circuit properties are distributed along the line. A detailed discussion of how these properties are taken into consideration in studying the characteristics of an ac circuit is included in the article ELECTRICITY.

For long transmission lines, or in systems in which high accuracy is desired in short lines, special equations have been developed to determine the voltage and current at any point on the line and thus determine system characteristics. For shorter lines, system characteristics can be calculated by approximation. At frequencies of 50 to 60 hertz, up to a maximum distance of 100 miles (160 kilometres) for overhead lines and 15 miles (24 kilometres) for underground, a line with distributed line constants can be treated as a line with concentrated or lumped constants.

The capacitance per unit length in underground cables is about 30 to 100 times greater than that in overhead lines. As pointed out above, this characteristic means that the charging current in underground cables 20 miles (32 kilometres) long may be as high as that in 1,000 miles (1,600 kilometres) of overhead transmission lines. As a result, the current-carrying capacity of cable lines may be reduced by this current, and losses in the transmission of energy may be substantial. When dc is used, however, the high line capacitance has no effect on the power transmission, so long as the steady conditions are maintained.

In interconnected systems involving many transmission lines and generating stations, it is very difficult and time consuming to calculate electrical characteristics (line current, power flow) on the basis of the special equations mentioned above. To speed up these calculations, network analyzers were developed in the 1930s and are extensively used. Power system models are set up on these analyzers, and by applying voltages at points corresponding to generating stations, the current, voltage, and power at any desired point in the system can be readily determined.

Use of digital computers

Since World War II, digital computers have largely replaced network analyzers, permitting solution of system problems with higher precision and with more flexible system arrangements.

Three-phase systems. A single-phase transmission line consists simply of two conductors. Three conductors make possible a three-phase transmission line, in which the voltages and currents in the three wires are out of step by one-third cycle, or 120° in electrical degrees. Such a

system can transmit as much power as three separate single-phase systems with six conductors.

Generators and motors are also built as three-phase units. The constant flow of power, an attribute of three-phase systems, makes them operate much more smoothly and efficiently than they would if the flow of energy were pulsating, as in a single-phase circuit. Also, magnetic interaction between the phases, or windings, of a three-phase motor makes it inherently self-starting (see ELECTRIC MOTOR).

There are two representative connection schemes in the three-phase system. One is the Y, or star, connection, and the other the delta (Δ) connection. In the Y-connected arrangement (see Figure 3, top), the common connection is the neutral. In the balanced three-phase system, the current in the common conductor (or neutral conductor) is always zero, and that conductor can be eliminated entirely. In actual practice, the balance is not always perfect, and some provision must be made for current in the neutral conductor. If the current likely to be encountered is small, the neutral may be connected to earth (grounded), and this small current may be permitted to flow into the earth. In many three-phase distribution systems, especially in the United States, Y connection with a separate neutral conductor is often used. In the delta-connected three-phase system (see Figure 3, bottom), the fault current to ground in case of a single-phase fault is limited to a smaller value determined by the capacitance of the system. This property in turn makes the voltages of the unfaulted phases higher than in the Y-connected system. The delta-connected system is used mostly in distribution lines and in low-voltage transmission lines. A similar arrangement to the three-phase circuit can be established with four phases or even more, the phase differences between successive voltages being equal to one-fourth of a cycle if there are four circuits, and so on. The three-phase system, however, is the most efficient and is almost universally used to the exclusion of all others.

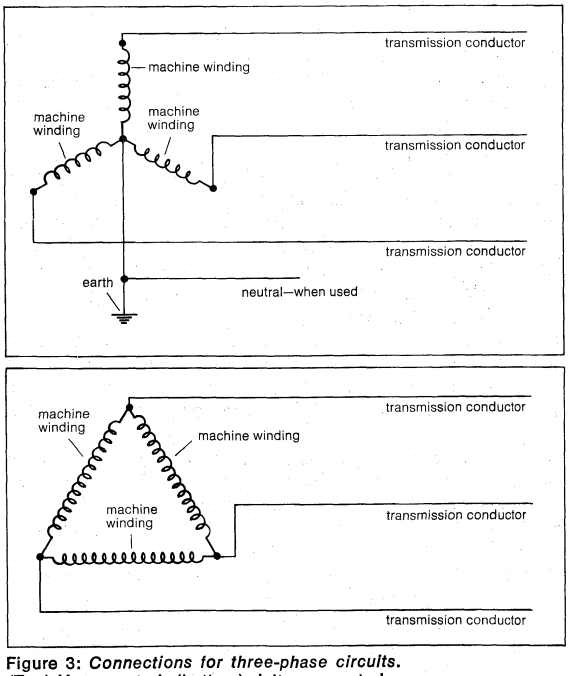


Figure 3: *Connections for three-phase circuits.* (Top) Y-connected; (bottom) delta-connected.

TRANSMISSION LINE DESIGN AND CONSTRUCTION

After a power-plant site has been selected, surveys are made to determine the best routes for the transmission lines. Before construction, studies determine the voltage, conductor spacing, number and type of insulators, conductor size and material, pole or tower design, and the protective features to guard the line against damage from lightning or other hazards. On large systems it is necessary to check normal and abnormal operating conditions, and extensive studies on network analyzers are often necessary.

Digital computers have been used to optimize transmission line design because of the complex relationships between the electrical and mechanical characteristics of transmission lines and their effect on economical design. Because nearly 200 separate items of design, material, and labour cost are used, digital computers can effect large savings in the construction of long-distance extra-high-voltage transmission lines. Several different types of conductor arrangement and supporting tower structures are sketched in Figure 4.

From *McGraw-Hill Encyclopedia of Science and Technology*, 3rd ed. (1971); used with permission of McGraw-Hill Book Co.

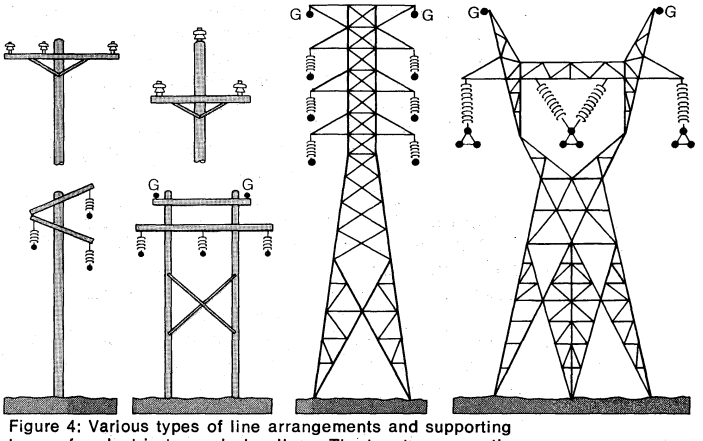


Figure 4: Various types of line arrangements and supporting towers for electric transmission lines. The two towers on the right are steel structures; the others are made of wood poles. G indicates a ground or neutral conductor (see text).

Choice of transmission voltage. The choice of transmission voltage depends on the amount of power to be transmitted, and on the transmission distance.

Effects of higher voltage

For a given amount of transmitted power, adoption of higher voltage reduces the line current, leading to smaller cross-section line conductors and reduced cost. Other items, such as insulator cost, rise because of the adoption of higher voltages, but, generally speaking, transmission line cost decreases with an increase in transmission voltage. In the substations, however, transformers, circuit breakers, and other components rise in cost with an increase in transmission voltage. The substation is the point at which the high voltage from the transmission line is reduced in voltage and redistributed over intermediate transmission lines. The transmission voltage is thus determined by the balance of the gain in transmission lines and the loss in the substations. In long-distance transmission lines, therefore, adoption of higher voltage is justified.

In the early 1970s, 100, 140, 220, 275, 380, 500, and 735 kilovolts were standard transmission voltages, and transmission at 1,000 kilovolts and more was under study.

Conductors. Materials generally used for transmission line conductors are hard-drawn copper and aluminum. A composite conductor called ACSR (aluminum conductor, steel reinforced) consists of aluminum wound around a steel core that provides strength.

Among many requirements for conductor materials, the following are important: (1) high electrical conductivity, (2) high mechanical strength, (3) low specific gravity (weight per unit volume compared with water), (4) low corrosion in air, and (5) ease of jointing.

If the electrical conductivity of soft-drawn copper is taken as 100 percent, that of hard-drawn copper is 97 percent and that of aluminum is nearly 60 percent. In ACSR are combined the high mechanical strength of steel and the relatively high electrical conductivity of aluminum.

The required cross-sectional area of a conductor is determined by the current flowing through it. Losses from the resistance of the transmission line (ohmic losses) can be calculated as the product of the resistance of the line

and the square of the line current. The resistance of the line is inversely proportional to its cross-sectional area.

The transmission line conductor is heated by the ohmic losses, and this heat is dissipated into the air, bringing the conductor to its equilibrium temperature. At high temperatures the mechanical strength of the conductor is reduced, and so its equilibrium temperature must not exceed a certain critical level, generally around 212° F (100° C); to allow a margin of safety the design is usually for about 194° F (90° C). The current required to produce this temperature is the upper limit of line current, called the current-carrying capacity of the conductor. Aluminum alloys with high mechanical strength at relatively high temperature have been developed, making possible a conductor with a higher current-carrying capacity than ordinary aluminum.

Corona and radio interference. As has been seen, if the voltage between conductors in a transmission line exceeds a critical value, a hissing sound comes from and a glow appears around the conductor, a phenomenon known as corona. The critical voltage depends upon the diameter of the conductor, the roughness of its surface, its distance from other conductors, and atmospheric conditions. Corona is caused by intermittent ionization of the air surrounding the conductor, producing high-frequency waves and causing energy loss as well as interference in nearby radio and television receivers. The only practical way to minimize corona is to increase conductor diameter. At sea level, for example, a minimum conductor diameter of about one inch (25 millimetres) is necessary for a 230-kilovolt line, about 1.5 inches (38 millimetres) for 330 kilovolts, and about 2.5 inches (63 millimetres) for 500 kilovolts. Sizes must be increased at higher altitudes, insulation strength of the atmosphere decreasing there.

Aluminum has become preferred for extra-high-voltage transmission line conductors, despite its conductivity being lower than that of copper. The poorer conductivity of the aluminum actually works to its advantage, because a larger diameter (especially with its internal steel core) is required for high voltage.

At 500-kilovolt transmission, the conductors must be 2.5 inches or more in diameter to avoid corona. Such conductors weigh nearly 4 pounds per foot (6 kilograms per metre) and are thus difficult to manufacture, ship, and install. A scheme to alleviate these difficulties and provide other benefits, known as bundling, consists of installing two or more cables (called subconductors) of much smaller diameter spaced one foot (30 centimetres) or more apart, in place of a single large conductor. The electrical stress on the air is reduced to the equivalent of the single large-diameter conductor. This arrangement, known as a bundle conductor, decreases line impedance and increases electrical load capability by 25 percent or more. In Figure 4, the right-hand tower uses bundle conductors, each bundle consisting of three conductors.

Advantages of bundling

A quadruple conductor scheme for 400-kilovolt transmission was studied by the Germans during World War II, but installation of a double conductor occurred first on the 380-kilovolt Harsprånget transmission line in Sweden, in 1952. Since then nearly all transmission lines with voltages ranging from 300 to 735 kilovolts have used the bundle conductor, and future 1,000-kilovolt transmission systems will undoubtedly adopt it also.

Insulators. For overhead transmission lines the conductors are always bare metal. High-voltage conductors are suspended from the poles or towers by strings of insulators, each consisting of several disks of porcelain or glass. As voltages increase, it is merely necessary to add more units to the string. At lower voltages, an insulator of porcelain or glass (a pin-type insulator) suffices.

Insulators must be strong enough to support the weight of conductors and to resist the forces of wind and ice. A type of design in widespread use in the EHV transmission lines is the V string support, using two insulators at an angle of nearly 90° (Figure 4 right). While this design requires additional insulators, the conductor is restrained from swinging in a crosswind, so that reduction in the required size of the tower structure and the width of right-of-way result, which may result in larger savings than the additional insulator cost.

Insulators must withstand not only the normal operating voltages but wherever possible abnormal voltages of short duration caused by lightning or by the operation of circuit breakers. Insulating properties are degraded by dirt and moisture, but the minimum insulation value must prevent electrical breakdowns for normal voltages, even during unfavourable weather. Severe deterioration occurs when salt is deposited on the insulators, and thus the number of disks in each insulator must be increased when the transmission line is installed near a sea coast.

Fortunately insulator failure nearly always occurs as an electric arc discharge through the air around the insulator or over its surface rather than by puncture of the insulator itself. If the line is disconnected from service rapidly by circuit breakers when an electrical fault occurs, the insulator may suffer little permanent damage and the line frequently may be restored to service immediately.

The insulation between conductors on overhead lines consists solely of the air between them; spacing must be adequate to prevent the conductors from coming in contact or from coming so close that a flashover occurs between them.

Towers and poles. Supporting structures for transmission lines are steel towers or poles of steel, concrete, or wood. Steel towers are generally used with voltages of 100 kilovolts or more, but wooden poles may be used up to 220 kilovolts if climate and other local conditions are favourable. Several tower designs are shown in Figure 4.

Effect of climate on tower design

Climatic differences require design variations in transmission towers. In some regions, lines must resist hurricanes or typhoons, while in others they must withstand snow and ice storms. The severity of lightning differs widely from area to area. The topography that a line traverses may vary from open plains to forested mountains. Construction requirements and maintenance, particularly where access is difficult or corrosion is a problem, are important factors in the selection of tower material and design. Conductor configurations at the tower are horizontal, vertical, and sometimes triangular.

A single-circuit transmission tower carries three conductors, a double-circuit tower six (see Figure 4). In large city suburbs, where acquisition of right-of-way for transmission lines is difficult, there may be three-circuit and four-circuit towers.

Spacing between towers is determined mainly by a compromise between two factors. If the span is made larger, the total number of towers needed is smaller, apparently leading to lower construction costs. With increased spacing, however, the sag of the conductors between towers becomes larger. The conductor height above ground is always limited by national regulations, and so to minimize the effects of sag the height of the towers must be increased, leading to a weight and cost increase. Span length is compromised between these two factors. Standard spans widely adopted are: for steel tower lines, between 150 and 300 metres (500 and 1,000 feet); for steel pole lines, between 100 and 150 metres; for wooden pole lines, between 50 and 100 metres. Where highways, rivers, or telephone lines are crossed, special long spans are often necessary.

The major requirement in designing transmission towers is mechanical strength. Besides the weight of the conductors, allowance must be made for wind pressure and, in snowy regions, the weight of snow and ice on the lines. When ice and snow fall off, the line jumps and may contact other conductors, creating a short circuit. To avoid this, consideration is given to the conductor configurations at the towers. Sometimes a large current is circulated through the conductor for ice-melting.

Growing popular awareness of the aesthetics of man-made structures has led to a widespread adoption of high-strength steel alloys to make towers less bulky and of painted towers that blend into the terrain. Existing facilities are being used more intensively to avoid problems in the acquisition of new rights-of-way, such as rebuilding lines to operate at higher voltages and substituting multi-circuit structures for single circuits. Fundamental-

ly, however, low-cost underground transmission lines appear to be the real answer; research and development work in the early 1970s was in progress in several countries.

Lightning protection. Transmission lines are susceptible to lightning, and a direct stroke may produce a serious short circuit unless special protection is provided. Statistical data on lightning in many areas of the world are available, and the number of lightning faults that will probably occur during a year can be forecast. A moderately high level of insulation is provided to prevent insulation failure against all except the more severe lightning strokes, and in especially dangerous locations added protective measures are used. Insulating transmission conductors at each pole or tower against an electrical breakdown that might result from a severe stroke of lightning, however, is usually uneconomical.

Circuit breakers. When insulation failure or other accident occurs on a transmission line, the line must be disconnected quickly from the unaffected portion of the system. Switches for this purpose, called circuit breakers, operate to close and open circuits under load when conditions are normal and to interrupt circuits during abnormal conditions.

The speed factor in circuit breakers

As systems became extended and the problem of system stability became better understood, it was recognized that speed in opening a faulty circuit was a major factor in increasing system stability. Design efforts therefore have been made to shorten the interruption time. In 1926 the interruption time was reduced from 45 cycles to 25 cycles, in 1932 to eight cycles, in 1954 to three cycles. In the 1960s two-cycle circuit breakers were available, and one-cycle breakers were under development in the early 1970s in Japan. Circuit breakers in turn are controlled by precise protective relays that locate failures and cause only the needed circuit breakers to operate. The relays can function in as short a time as one cycle, and thus a section of a transmission line having a failure can be isolated in less than four cycles, or $\frac{1}{15}$ of a second.

Three types of high-voltage circuit breakers are in common use: oil-filled, air-blast, and gas (sulfur hexafluoride)-blast circuit breakers. Because the current cannot be interrupted instantaneously, an electric arc is drawn between the two contact points of the circuit breaker as they separate. In the oil-filled breaker a small amount of the surrounding oil is vaporized, producing a high pressure that drives quantities of unvaporized oil into the path of the arc. Moving electrons, the main elements of the electric current in the arc, become attached to the vaporized oil particles, and their mobility is so restricted that the arc is extinguished. In the air- or gas-blast breaker, the arc is quenched by the force of an air or gas jet.

For a given set of system conditions, the current resulting from a fault can be determined fairly accurately, but the difficulty lies in forecasting future circuit and generating-station changes that, during the expected life of the breaker, materially change the duty (rupturing capacity) it is expected to perform. One method frequently used to keep down breaker investment is to move breakers, as the system grows, from the heavier duty locations to the fringes, where the interrupting requirements are lighter.

Protective relays. Protective relays are usually a complicated assortment of devices used to protect power systems against outages. The original purpose of relays was to protect apparatus, a purpose now secondary to maintaining continuity of service. Relays must be set for short-circuit protection and not for overload; overload protection is the responsibility of the station operators (see below).

Relays receive operating energy from current and from potential transformers connected to the circuit to be protected. The potential transformer reduces the system voltage to 120 volts, and the current transformer reduces the normal maximum line current to five amperes, two values for which relays can be readily built.

Overcurrent relays. The simplest and oldest relay is the ordinary overcurrent relay, still the principal functional unit of protection schemes. The overcurrent relay operates when excessive current passes through the system and normally has a tripping time inversely proportional to the magnitude of the fault current.

Differential relays. For removing faulty generators or transformers from the circuit, differential relays are universally used. The protective schemes are based on the principle of balancing the secondary currents in the current transformers at the equipment terminals so that under normal conditions, or during external faults (that is, faults that are outside the protected zone), a balanced current circulates in the transformer secondaries. A protective relay connected in parallel with this balanced circuit receives current only when a fault occurs in the equipment or inside the protected zone of the transmission line.

Directional relays. For situations in which the current can flow to a fault from two directions, the overcurrent relay must be given a sense of direction if the loss of circuits is to be restricted to the faulty section, a characteristic provided by equipping the relay with an element responsive to the direction of the power flow.

Distance relays. Most power systems incorporate extensive transmission networks with several sources of power. The relaying of such systems with simple overcurrent and directional relays, even where possible, would involve keeping a fault on a line for long periods. Faulted sections of transmission lines must be switched out in the shortest possible time. Relatively complicated and costly relay schemes have therefore been developed. Distance relays measure either the impedance or the reactance of the circuit between the relay and the point of fault as a measure of distance. The greater the distance to a fault, the longer it takes the relay to disconnect (trip) the circuit, the tripping time being proportional to the distance to the fault.

Application of relays. Relays act to open appropriate circuit breakers on either side of the faulty machine or line section by energizing the coils that trip (open) the circuit breakers. Relays have three broad classes: those that protect against individual machine and transformer faults; those that protect against faults in buses (uninsulated bars or rods for interconnecting equipment in generating stations or substations); and those that protect against transmission circuit faults. Differential relays can be used to protect against bus faults by balancing all the incoming currents against all the outgoing currents. If there is a fault in the protected zone, the incoming and outgoing currents will not be equal, and the differential relay will operate.

Protection against transmission line faults is obtained mostly by using distance relays, which have, however, the disadvantage that they cannot be designed to cover the whole section to be protected, because the relay is not accurate enough to allow precise location of a fault just inside or just outside the protected section. Some 10 to 20 percent of the length of the section at either end, therefore, does not have the benefit of high-speed protection. This problem has been solved by superposing a carrier current signalling system on the distance relay. The carrier signal is used to block the tripping of distance relays in the unfaulted sections. Microwave relaying has also been used.

Electronic relays. All the relays mentioned above are called electromagnetic relays, because the operating principle of these relays is based on electromagnetic induction. Electronic relays, which came into use in the late 1960s, have the advantages compared with the electromagnetic relay of faster operation (there are no moving parts and contacts), smaller volume, and less power needed for operation.

Underground cables. Underground cables were used for street lighting when Thomas A. Edison began the operation of dc distribution systems in New York City in the 1880s. The structure of the cable was simple: a copper conductor was wrapped with jute and placed in an iron tube, called the Edison tube. Later, in 1884, use of overhead lines was prohibited in New York City, thus leading to further development of underground cables.

Early dc cables

Still later, ac transmission was introduced, voltages became higher and higher, and new methods for insulating the cables were introduced.

In 1917 Luigi Emanueli, an Italian engineer, established the fundamental principle of oil-filled cables, and in 1922 M. Höchstädter developed modern-type cables. Cables operating at 220 kilovolts were laid in Milan in 1932 and in Paris in 1936. In 1965, 345 kilovolt cables were laid in New York City, and these are the highest voltage underground cables in the world. Cables operating at 380 kilovolts are used at hydroelectric stations, such as Harsprånget, Sweden, but the cable length is quite limited. Cables of 500 kilovolts were under development and extensive testing was under way in the early 1970s, mostly in the United States.

The cable most commonly used because of low dielectric losses and cost is the paper-insulated, oil-impregnated type. Single conductor cables are built up by winding layers of paper tape on a stranded conductor and impregnating them with oil. A tight-fitting extruded lead sheath is applied over all. Three-conductor cables may consist of three round conductors, each insulated with paper, with a jute filler to round out the cable. Around all this is more paper insulation, and a lead sheath over the whole. These cables may be either belted or shielded. In the belted cable, the conductors are insulated with paper, cabled together with a filler piece, and wrapped with more layers of impregnated paper (the so-called belt) before the lead sheath is applied. In the shielded cable, a thin metallic shielding tape is wrapped around the individual conductor insulation to control electrostatic stress, eliminate corona formation, and improve thermal conductivity. In general, single-conductor, paper-insulated cable is used up to 69 kilovolts and three-conductor cable (shielded) up to 46 kilovolts. When the voltage exceeds 69 kilovolts, special designs become necessary.

Besides paper, cable insulation may consist of rubber, varnished cambric, or synthetic compounds. Experience has shown that increasing the thickness of the paper insulation is no guarantee of improved performance. One of the earliest ways of producing better cables, and one still popular, is the use of the Pirelli oil-filled, paper-insulated cable, in which the conductor strands are wound around an open helical spring to form a hollow conductor. Oil under pressure is kept on the cable by means of reservoirs so that no voids can form in the insulation. The oil is relatively thin and remains fluid at all temperatures. This type of cable has been in service in Paris since 1936 at 220 kilovolts.

Another widely used cable for high-voltage service is the Bennet pipe-type cable, consisting of three single-conductor paper-insulated cables laid in a steel pipe six to nine inches (15 to 23 centimetres) in diameter. The pipe is kept filled with oil at a pressure of 200 pounds per square inch (14 kilograms per square centimetre). Nitrogen gas is sometimes used instead of oil.

A third type is the compression cable, like the Bennet except that the three cables retain their sheaths (lead or polyethylene), and pressure is exerted on the sheaths. This external pressure on the noncircular sheath acts on the paper insulation to prevent the formation of voids. An advantage of this type of cable is that threading it into the pipe does not require such meticulous care as with other types.

Economic and aesthetic factors

Because the cost of underground cables is much higher than that of overhead transmission lines, it is not generally economical to place cables underground except in very high load density areas. Elsewhere, however, utilities are often required to place transmission circuits underground for aesthetic reasons. The ratio of construction cost between underground cable lines and overhead lines becomes steadily higher with the increase in the transmission voltages: four to six times at 69 kilovolts and eight to 15 times at 220 kilovolts and upward.

Obtaining rights-of-way for overhead lines in heavily populated areas is becoming increasingly difficult, making it urgent that the utilities develop techniques for economical underground transmission. Several new techniques were under study in the 1970s, including the increase in rating of underground cables simply by forced cooling. Another was the use of sulfur hexafluoride gas instead of oil-impregnated paper; these are gas-insulated pipe-type cables, having very small capacitance.

Cryogenic (low-temperature) cables were also being actively investigated in the 1970s. When certain metals are cooled sufficiently, they become superconducting; that is, the resistance of the conductor becomes virtually zero. Possibly as much as 10,000 megawatts of power can be transmitted over a single circuit of superconducting cable (see also CRYOGENICS, APPLICATIONS OF). As mentioned above, dc transmission is being actively investigated as a technique for transmitting large blocks of power.

IV. Distribution

VOLTAGE CONSIDERATIONS

In early systems, only one step of voltage transformation occurred at each end of a transmission line; but as systems grew, it became necessary to use two or more steps of voltage reduction at the receiving end. The high voltage needed for long-distance transmission was too high for lines on city streets, and yet the secondary voltage was too low for supplying electric energy to consumers over a wide area. Intermediate voltages were therefore introduced into the system via distribution circuits. In many systems one intermediate level is insufficient, and two or more levels are used.

Intermediate voltage

As systems were interconnected, the transmission circuits of the original system often became the distribution circuits of the combined system, so that the distinction between distribution and transmission became vague. Transmission, however, in general implies larger blocks of power delivered at higher voltage to a few main substations, while distribution implies smaller blocks of power delivered at lower voltages to many smaller substations or directly to consumers.

Levels. Three voltage levels are normally involved in the distribution system: (1) subtransmission voltage, used in connecting the supply point on the transmission network to the distribution substations; (2) primary distribution voltage, used on feeder lines from the distribution substations to the line transformers; and (3) utilization voltage, or secondary distribution voltage, for lines from the neighbourhood transformer to the customer. Subtransmission circuit voltage varies roughly from 20 to 70 kilovolts, that of primary distribution circuits from three to 20 kilovolts, and that of secondary distribution circuits from 100 to 500 volts.

The voltage for residential services in the United States is uniformly 120 volts, and in Europe it is mostly 200 or 220 volts, a difference that leads to considerable variation in distribution methods.

In the United States the voltage is reduced in a distribution substation to 2,400 to 13,800 volts, and a second reduction is made to utilization voltage (120 volts) by means of a large number of small transformers mounted on poles, or situated in vaults, known as distribution transformers. In Europe the subtransmission or primary distribution voltage is brought into a kiosk, or vault, and reduced to 220 volts, and a relatively large area is covered at 220 volts. Distribution transformer output in the United States ranges generally from five to 100 kilovolt-amperes, whereas in Europe it is generally from 50 to 600 kilovolt-amperes.

The steady increase in load per customer requires constantly higher capacity circuits. To increase capacity, the voltage is usually increased from the initial level to a higher one as the system grows. In the United States primary distribution circuits of 2,300 or 4,600 volts were considered adequate in the early 1950s, but 12,000, 33,000, and even 69,000 volts were standard in many areas by the early 1970s. Commercial and industrial loads are often supplied at the higher voltages, and high-rise buildings usually are supplied through high-voltage circuits of 480 to 12,000 volts; but the single-family residential service has not changed from 120/240 voltages, which are supplied usually from one phase of the three-phase system, 120 volts by a two-wire system and 240 volts by a three-wire system with midpoint grounded.

Use of three-wire service in homes

Most older residences have only a two-wire, 120-volt supply, but the increased use of household appliances has led to the practice of putting three-wire, 240-volt service into practically all new residences. In these installations fractional horsepower motors are supplied from the 120-volt lighting circuits, while cooking stoves, water heaters, and large single-phase motors up to five horsepower are supplied from the outside wires of the three-wire system at 240 volts.

In Europe, 220 volts are supplied from a single-phase two-wire system. In the United States single-phase three-wire system, the voltage above ground in the normal operating condition is limited to 120 volts, whereas in the European system the full 220 volts always appears on the ungrounded wire.

For power use, 240- and 480-volt, delta, three-phase voltage is provided, and for combined light and power use, 120/208-volt, three-phase, four-wire, star (Y) voltage is used in the United States. The latter is the common practice in metropolitan secondary network systems. In Europe, a 220/380-volt three-phase four-wire Y system is commonly used.

Regulation and regulators. Quality service requires maintenance of proper voltage without interruption. Distribution lines are designed to maintain the voltage at the customer's premises within acceptable limits. A single value of voltage cannot be maintained on any given circuit for all locations and hours of the day, because voltage varies depending on both the distance of the service switch from the supply transformer and the load taken, which varies from morning to night.

Voltage is one of the most important factors in the design of equipment, appliances, and devices, and satisfactory performance cannot be attained with wide variations in the supply voltage. Most appliances and devices operate satisfactorily over some range of voltage, so that a reasonable tolerance is permissible. The total range between maximum and minimum voltages on a system during the whole course of its operation is called the voltage spread. The voltage at the customer's premises is generally highest at off-peak periods, usually around midnight, and lowest during peak loads.

The permissible voltage variation, often specified by government regulations, is plus or minus 5 to 6 percent of the standard value in the urban and suburban areas, with $\pm$ 10 percent variation permitted in rural areas.

As system loads grow, it becomes increasingly difficult to keep voltage variations within a tolerable range. Increasing the voltage level of the system, at either the primary or the secondary distribution level, or both, is an effective method of control, as doubling the voltage permits the permissible load to be quadrupled. Doubling the voltage also increases the shock hazard to consumers, but effective countermeasures, particularly in Europe, minimize such risks. A feasible solution might be to supply homes with a higher voltage for large, built-in appliances and provide a separate transformer for 120 volts for conventional lighting and convenience outlets.

Automatic voltage regulation

It would be uneconomical to design a distribution system in which the voltage drop from no load to full load and from first customer to last customer falls within acceptable limits of voltage regulation. Automatic voltage regulation, therefore, is always provided. The three methods used are: bus regulation in the substation; individual feeder regulations in the substation; and regulation along the line by regulators mounted on poles. Bus regulation can be used only when all the feeders leaving a station are of about the same length and have the same character of load, as is generally the case in any given residential area. Such regulators merely raise or lower the voltage in response to the total load on the substation. Where considerable differences exist between individual feeders, either in length or loading, individual feeder regulators are used.

The problem of regulation on rural lines differs from that in urban areas. The lines being long, the voltage is usually higher and the load per mile is small. The regulators are often mounted on poles along the primary feeder line.

Continuity of service. The other problem is that of continuity of service; that is, how to reduce the number and duration of interruptions of supply (outages) and also to reduce the number of customers affected by the interruptions.

Distribution lines, especially overhead lines, are subject to the effects of lightning storms and high winds and also to such accidents as trucks colliding with distribution poles. Areas affected by such accidents are minimized by the design of the system layout, by use of effective protective measures, and by use of mechanized repair groups to restore service as quickly as possible.

The problem of distribution protection is complicated because in case of a fault it is desirable to remove as little as possible of the faulted circuit in order to minimize service interruptions; yet the protection cost must bear some economic relation to the revenue obtainable from the circuit being protected. This condition means that protective schemes involving circuit breakers and protective relays cannot for the most part be justified.

Primary feeders supply power to distribution transformers. The faults that can occur, therefore, are failure of transformers or the feeders themselves. The transformers are connected to the feeder through a fuse known as the primary cutout. If a feeder supplying many transformers is of considerable length, it is undesirable to drop the whole feeder for a fault some distance away. Between the circuit breaker at the substation and the last distribution transformer, therefore, the feeder is sectionalized by reclosing devices at several locations. The operation of all the fuses and reclosers must be so coordinated that only the fuse or recloser nearest the fault on the supply side locks out if the fault is sustained. When the fault is of a transient nature, after one or more reclosing cycles, the total circuit of the primary feeder resumes operation.

The devices available for this type of protection are single-phase reclosing fuses, single-phase sectionalizers, single-phase automatic reclosers, and three-phase automatic reclosers. Their use is economical because the great majority of faults can be cleared through opening the circuit for only a fraction of a second.

Reclosers are essentially circuit breakers with relatively small rupturing ability, equipped with a mechanism to open and then reclose the circuit automatically. They are usually made to operate four times before locking out. Sectionalizers are similar to reclosers except that they have no rupturing capacity, being designed to open the circuit at zero voltage, that is, after the circuit has been opened by a recloser or circuit breaker.

Advantages of the loop system

The layout of distribution circuits also has a close relation to the reliability of service. In general, with a loop system the affected part in case of a sustained fault is far smaller than with a radial system. Duplicate service is seldom available to residences, but it is often provided to such customers as hospitals, municipal services, and major industrial plants, particularly if material being processed may suffer severe loss during even a short power interruption. Two or more circuits and duplicate apparatus are needed in these cases. Because severe short circuits occasionally involve both circuits of a pair designed to reinforce each other, resulting in a total shutdown, duplication of all facilities will not provide complete assurance of continuous supply.

NETWORKS

Network systems. In metropolitan areas, forms of distribution systems known as network systems have been developed. Begun in New York City in 1922, the secondary network system provides that the outputs of the final transformers, instead of feeding only a few customers, will all be connected directly together over a wide area by large conductors to form a network, and all users will be fed from these secondary lines running along each street. The transformers, instead of being small units, are large, up to 500 kilowatts or more, and are frequently fed directly from the generating stations at generator voltage, often 13,800 volts, by underground cables. The transformers are located in underground vaults, usually

under the streets. The electric system of the secondary circuits is 120/208-volt, three-phase, four-wire Y connection.

Network protectors. Special circuit breakers connecting the transformers to the secondary mains open whenever power feeds from the network to the transformer, even in small amounts. This circumstance may occur when the primary circuits are opened at the generating stations and the transformer idling losses are supplied from the network, or when a transformer defect occurs. Otherwise, whenever power can feed from the transformer to the network, the protector closes. Short circuits in the low-voltage network, which rarely occur because it consists of large conductors, well insulated and placed underground, are cleared by melting of the conductors in the region of the short circuit or by melting of reduced sections known as limiters. Because every main is fed by at least two banks of transformers, one at each end, the first melting of the cable does not cause any interruption in service to customers. Special modifications are used in industrial plants and large office buildings, a system that is economical in regions of high load density.

A modification known as the primary network system extends a similar plan to supply the distribution circuits feeding the final distribution transformers in residential areas. These circuits are usually overhead, and fuses or relays and circuit breakers are relied upon to isolate defective circuits. Each distribution circuit is fed from two sources, however, so that a high degree of reliability is achieved with relative economy. This plan is not as widely followed as the secondary network system.

V. Electric power systems

THE ELECTRIC UTILITY

The electric utility is unique in that it sells a product that must be generated at the instant it is to be used. With existing technology, it is impossible to build up an appreciable stockpile of electric power, so that the utility must be ready to generate enough electricity to satisfy the demand at any time. If the utility has excess generating capacity, it can usually cover any anticipated demand; but an overabundance of excess capacity would mean that too much of the plant would be idle too much of the time, thereby increasing cost. At the same time, few products have a greater need for quality and reliability. Reliability is measured both by the number of interruptions to service and by the number of customers affected. The ideal, to reduce the number of interruptions to zero, would mean exorbitantly high levels of investment; as a matter of practical economics, electric power systems are designed to keep interruptions within tolerable limits. Even to meet this goal requires making appropriate investments in plants and using the latest technology.

Quality of service is measured by the degree of attainment of standard voltage and frequency in the supply of electric power. System design and operation determine how close to the standard the voltage and frequency can be maintained.

Private electric utilities can operate only by the grant of a special privilege or franchise from the national or local government that permits use of streets, subsurfaces, and other public property. The franchise is usually granted only to a single concern in any given area, creating a monopoly. In the past, duplication of service to one area was sometimes permitted to increase competition. In such cases, the streets were cluttered with facilities and equipments were duplicated, leading to an overinvestment. A franchise, therefore, gives the utility certain rights and imposes corresponding duties.

Legal rights of utilities

A utility has legal rights that allow it: (1) to invoke under certain circumstances the law of eminent domain to obtain property—that is, invoke condemnation proceedings to acquire it; (2) to charge and obtain a reasonable rate for the service; and (3) to establish and enforce reasonable rules and regulations under which it provides service. The principal duties are that the utility must: (1) serve all who apply for the service; (2) provide power up to maximum capacity and be prepared to meet future increases in demand; (3) provide service that is adequate and safe; (4) not discriminate unfairly among customers (though it may serve different categories of customers at different rates); (5) not obtain more than a reasonable price for its service.

The physical property of an electric utility usually comprises the following:

1. One or more power stations, which may be thermal stations using coal, oil, gas, or nuclear fuel, a combination of some or all of these, hydroelectric stations, or a combination of both types. Each power station has a switchyard containing transformers, circuit breakers, and control equipment.
2. A transmission line, usually erected on private rights of way.
3. Transmission substations with transformers to reduce the voltage to several subtransmission circuits, and attendant circuit breakers and control equipment.
4. Distribution substations in which the subtransmission voltage is again reduced to supply transformers mounted on poles for general distribution in suburban and rural areas or to supply transformers in vaults for underground distribution. For industrial customers, electricity is delivered at transmission voltage, at subtransmission voltage, or at distribution voltage, depending on the size and location of the factory and the most economical supply in the circumstances.

Divisions of a utility's technical department

To construct and operate these facilities, a utility has three divisions in its technical department: (1) the system planning division (planning, load forecasting, technical research); (2) the system construction division (design and construction of power lines and power stations, engineering standards and specifications); and (3) the system operation division (power generation, power transmission and distribution, load dispatching).

The system planning engineer is responsible for long-term planning. He must anticipate the growth of system load in both time and location because the construction time required for a hydroelectric station is from five to 10 years, for a thermal station three to five years, and for a nuclear power station four to seven years. This planning requires a careful study of civic, residential, commercial, and industrial developments, population growth trends for the district, and plans for roads and bridges. The future availability of fuel resources and cost trends for different fuels—in some cases on an international scale—must also be considered.

The construction engineer is responsible for constructing power stations and lines, but on major construction he usually works in conjunction with an outside engineering firm. He generally lays down the broad outlines of all construction work and also does all the work required for substation and transmission line construction.

The operating engineer is responsible for the day-to-day operation of the system. He supervises personnel in charge of the operation and maintenance of generating stations, transmission lines, and substations, and also a supervisor of operating systems, usually called the load dispatcher, who knows when and where the loads normally come on and go off and must have available the requisite generating capacity. He must also watch the weather because it is closely related to system load and to system disturbances. Sudden darkness causes the lighting load to increase abruptly, and thunderstorms or heavy snows may result in the loss of lines. Even social patterns and events may require monitoring; a popular program on television causes a sudden rise in load.

Economy of operation is important, and because the efficiency of power generation differs for each station at different loadings, the engineer must allocate the generating power to each power station so as to make most efficient use of the system facilities. He must also estimate the power loss in the transmission lines. The most economical load dispatching is a very difficult goal to achieve, and on-line digital computers are often employed to determine the best allocation of load for each power station.

The operating engineer must handle emergencies and be prepared to bring the reserve generating capacity of the

system into operation immediately while arranging for whatever repair work is necessary.

SYSTEM CHARACTERISTICS

In addition to the components of the electric power system named above there is one more, the customer's load, which varies greatly. In residential load, electric lighting is used in the evening, electric cookers are used about mealtime, electric cleaners and irons at any time, and electric refrigerators are switched on and off throughout the day and night by automatic control devices. Radio and television use depends upon the programs. Electric air conditioners and electric room heaters comprise a large residential load, their use depending on the weather.

Effects of weather on residential loads

The commercial load, as for offices and department stores, is high during business hours but low at other times. Lighting and air conditioning account for the greater part of the total load. The industrial load is highest during working hours, which depend on the nature of the factory and the number of shifts. Motor and heating loads usually consume the most electricity.

The load on the electric power system is the varying sum of all the residential, commercial, and industrial loads. An example of a system load curve is given in Figure 5. The demand shown is the hourly integrated average value; instantaneous peak load is 100 percent. Around midnight (M) the load is low, and the facilities are not used very effectively. In order to encourage power consumption during off-peak hours, reduced rates for room heating and water heating are frequently offered. Air conditioning accounts for an increasing percentage of the residential and commercial load and heightens seasonal variations. In some countries, such as the United States and Japan, the summer load exceeds the winter load—as evidenced by brownouts, if not by blackouts, in American urban areas.

A smoothed load curve and the load duration curve are shown in Figure 6. The load duration curve can be derived from the load curve and used for planning purposes. A typical yearly load duration curve for a specific system is shown at the left in Figure 7. This curve may be divided into three parts: base, middle, and peak load. The base load, the lowest load of the system, is present continuously, day and night, throughout the year. Peak load is defined rather arbitrarily as the heavy load portion of heaviest load, accounting for 5 to 15 percent of the hours in the year.

For most economical operation, the base load portion should be supplied by the power stations the operating costs of which are lowest; for example, by run-of-river-

Reprinted from *Principles of Electric Utility Engineering* by C.A. Powel by permission of The M.I.T. Press, Cambridge, Massachusetts. Copyright 1955 by M.I.T.

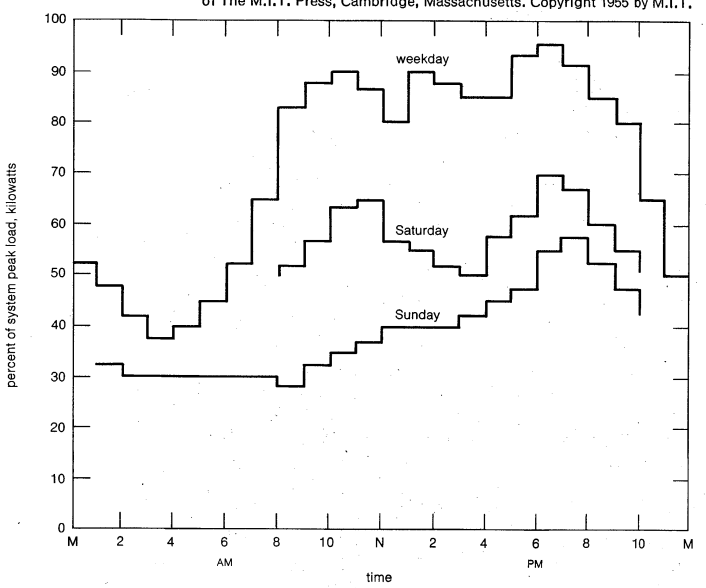


Figure 5: Typical daily load curves for a power station. The load is recorded in hourly steps.

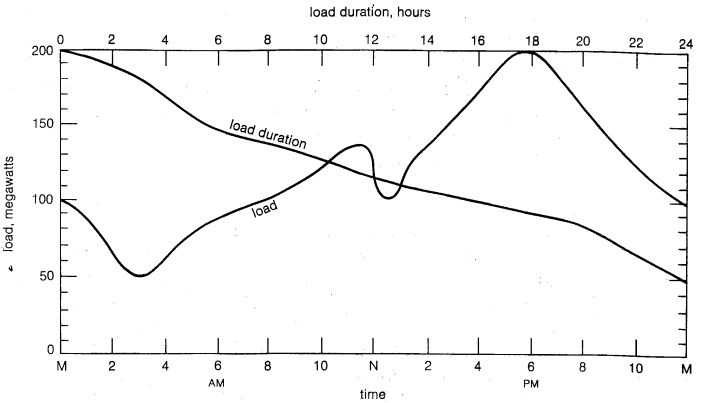


Figure 6: A smoothed version of the load curve of Figure 5 together with a load duration curve, indicating the number of hours each day that a particular load is exceeded.

Reprinted from *Principles of Electric Utility Engineering* by C.A. Powel by permission of The M.I.T. Press, Cambridge, Massachusetts. Copyright 1955 by M.I.T.

flow-type hydroelectric stations, by the most efficient thermal power stations in the system, and, increasingly, by nuclear power stations. The peak load portion should be supplied by power stations the construction costs of which are low. Peak load is relatively high in kilowatts

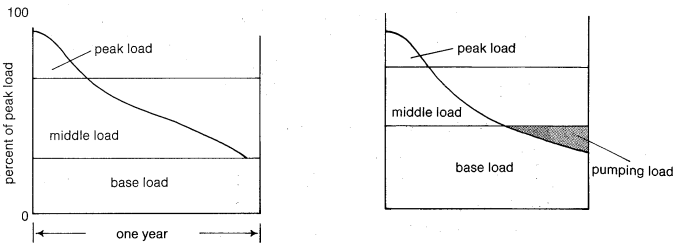


Figure 7: Typical load duration curves for a one-year period, showing how base load can be substantially increased (right) by using power during periods of low demand to operate a pumped storage system (see text).

but relatively short in duration, so that total kilowatt-hours are not great. Peak load may be supplied by pumped-hydro stations, old thermal power stations, and gas turbine stations.

Pumped hydro requires no fuel but requires power for pumping the water to the storage reservoir. Off-peak power can be used for pumping, raising the level of base load (Figure 7, right). The middle load is then supplied by older-type thermal stations. To accommodate starting and stopping of turbines once or twice a day, thermal stations are designed especially for such middle-part use. Effective system planning requires estimates of future changes in the percentages for the three types of load.

INTERCONNECTION

Factors in planning the capacity of a system

To determine how much generating capacity should be built into a power system, a number of factors must be considered. Foremost, of course, is the maximum peak load on the system. Emergency conditions, such as a fault in a generating station or in transmission lines or substations, must be considered, as well as scheduled outages for regular repair and maintenance. To provide the necessary reserve margin above normal peak power requirements, a certain reserve capacity is required, sometimes expressed as 15 percent over peak load and sometimes by the total of the two largest units in the system. Each system must be considered individually. The capacity of the largest generating unit is normally about 7 to 10 percent of the total system capacity.

The investment cost of generating capacity per kilowatt is usually lower as unit size increases, making the use of larger units economically beneficial. In a small system, however, such benefits cannot be realized. Larger systems can be assembled by interconnecting two or more smaller systems, however, thus enabling the use of a larger generator unit. Interconnections must be designed to

give an adequate power flow through interconnecting lines in an emergency. If help from neighbouring systems can be counted on in case of emergency, as can be realized by interconnection, reserve margins can be reduced, resulting in considerable economic benefits.

Interconnection also has economic advantages in day-to-day utility systems operations. Where system load conditions would normally require the operation of a generating station with a relatively low efficiency if the system were operating independently, interconnection can cut fuel costs by making use of more efficient generators from a neighbouring system. Another advantage of interconnection is the flexibility it allows in choosing sites for new generating stations; planning takes into consideration the interconnected systems as a whole, rather than a single independent system. Because of these advantages, electric power systems have grown steadily larger as a result of interconnection between different utilities in a country and even interconnections between countries.

Interconnection introduces many technical problems that must be solved in system operation and planning. Tie-line power flow, for example, in which a tie line transmits to neighbouring electric power systems, must be carefully controlled in the day-to-day operation of the interconnected system. Such power flow is normally predetermined on a contractual basis between the utilities concerned. The complexity of systems with interconnections at many points and at various voltages, however, creates difficult problems, for the solution of which the analog computer has been used since the 1930s. Such computers, known as ac calculating boards, network calculators, or network analyzers, have proved to be effective tools for studying power flow problems in interconnected system operation and planning. In the late 1960s they were being replaced by digital computers, which have greater speed and flexibility. In some instances, on-line computer control is used for the operation of large interconnected systems.

Importance of reliable communications

In addition to the normal voice communication by telephone and radio, a utility requires a large and complex network for telemetering (transmitting measurement information over long distances), circuit protection, and control functions. Extensive use is made of conventional wire circuits, carrier telephony, and microwave networks. In carrier telephony the transmission line acts as a guide for high-frequency waves.

As a power system grows, the fault current during disturbances increases proportionately, requiring that the rupturing capacity of all circuit breakers in the system be increased. A circuit breaker with inadequate rupturing capacity may explode when a system fault occurs. One way to alleviate the problem is to introduce series reactance (a circuit component that impedes the flow of alternating current) into the system to reduce the fault current. Such reactance causes a voltage drop in the system during normal operation, however, and so has drawbacks. Connecting two systems with a dc linking device is also a possibility. No ac fault current flows through the dc linking device; thus fault current is restricted to one system.

Another problem with interconnected systems is the possibility of a large-scale blackout if there is a system fault. Power systems are designed so that a fault can be located almost instantly and the faulted section switched out in a very short time, thus allowing stable operation of the rest of the system, whatever the nature and location of the fault.

Power system design is usually on a single contingency basis, but in system operation two or more contingencies may occur simultaneously, perhaps leading to a widespread interruption of power. Weather disturbances may break down transmission lines, while an error on the part of a load dispatcher, or failure of a protective apparatus (relays or circuit breakers), cannot be excluded. Thus, widespread blackouts, such as that of November 1965 in the northeastern United States and southeastern Canada, can occur. From such large-scale power failures much, at least, can be learned. First, reserve capacity, to be effective in an emergency, should be a spinning reserve; that is, the equipment must be in such a state that the output capacity can be used very soon after the disturbances. The time needed to produce electricity from a thermal station from a "cold" state is several hours. Second, load shedding techniques (*i.e.*, deliberate reductions in load) should be carried out as a necessary evil in case of an emergency. After a serious fault, there may be an imbalance between generation and load. When the load exceeds generating capacity the system frequency drops. Conversely, when generation exceeds load, system frequency goes up. When the system frequency drops by as much as 1.5–2 hertz, steam-turbine generators should be disconnected because of the possibility of mechanical breakdown. This shutdown further increases the difference between generating capacity and load, and system frequency drops even further, eventually resulting in total system breakdown. Consequently, before frequency drops to a critical level, some load must be disconnected. The load dispatcher selects the proper load to be cut off, but if the fault is serious, he may be too late. Thus, it has become the practice to install underfrequency relays at predetermined locations in the system for the purpose of load shedding. Third, in the case of a serious fault, circuit breakers in the system will isolate some of the thermal power stations. It is difficult to start up such stations without outside help. This was one of the main reasons why New York City suffered a prolonged (13-hour) outage in the 1965 blackout. Since then, major thermal power stations have installed gas turbines as auxiliary power sources for start-up. Also the gas turbine, with its quick starting, acts as a spinning reserve in an emergency. Finally, auxiliary power sources should be installed in public buildings, such as hospitals.

The measures mentioned above are countermeasures for alleviating the effect and minimizing the duration of a blackout; it is better to undertake the difficult task of planning, constructing, and operating the power system so that a serious breakdown will not occur. To achieve this objective, preparations must be made for simultaneous faults in the system, or countermeasures must be planned on a multicontingency basis, taking into account possible malfunctions of the system apparatus and human errors. System design work is increasingly so directed.

VI. International aspects

As transmission lines grew to cross national borders and electrical energy became a commodity in international trade, new problems arose and new solutions were developed.

MULTINATIONAL INTERCONNECTIONS

Early international efforts

The first multinational interconnection was introduced in Europe. The Electricity Committee of the Organization for European Economic Co-operation studied the question of international interconnection of supply systems, and in 1951 the Union for the Co-ordination of the Production and Transmission of Electricity (UCPTE) was set up, encouraging international exchange of energy between its members—France, Belgium, The Netherlands, Luxembourg, West Germany, Austria, Switzerland, and Italy. The total energy exchange within this group was about 3,000,000,000 kilowatt-hours in 1950 and about twice that in 1955. By the mid-1950s, the electricity production deficiency in western Europe had been largely made up, and interconnected transmission networks operating at 220 kilovolts (and sections at 380 kilovolts, which was standardized for future transmission) were functioning satisfactorily.

This scheme allowed advantage to be taken of the difference in the annual regimes of thermal and hydroelectric systems and so reduced on an international basis the total installed capacity needed to meet power demands.

The sum of non-simultaneous load transfers was 5,720 megawatts in 1964. Considering the overall transmission capability of the tie lines between the UCPTE countries (9,400 megawatts), the above non-simultaneous loading was 61 percent of the interconnection line capacity. Com-

pared with the load of the group, which amounted to about 72,000 megawatts, total exchange came to 7.94 percent.

The eastern European (Comecon) countries formed an interconnection, first at 220 kilovolts and later at 380 kilovolts, linking the Soviet Union, Poland, Hungary, Czechoslovakia, and Romania. Usually the Soviet Union and Romania are exporting countries and the rest importers.

In the 1960s further international links, between France and the United Kingdom, between Sweden and Denmark, and others began operations.

To encourage discussion of common problems and exchange of technical information, several international organizations were set up. The Conférence Internationale des Grandes Réseaux Électrique à Haute Tension, the World Energy Conference (formerly the World Power Conference), and the Union Internationale des Producteurs et Distributeurs d'Énergie Électrique dealt with transmission, generation, and consumption problems, respectively, while the International Electrotechnical Commission (IEC) was concerned mainly with standardization.

War damage had created a total European deficiency in generating capacity of some 10,000 megawatts by 1945. Reorganization started under the European Recovery Program, assisted by plans made by the Electricity Committee of the Organization for European Economic Cooperation.

The Economic Commission for Asia and the Far East planned an expansion of electric power facilities to increase the installed capacity in Asian countries to more than double the 1950 capacity of some 4,000,000 kilowatts. Postwar reconstruction of electric power systems and development schemes for new generating stations, both thermal and hydroelectric, were assisted financially by the International Bank for Reconstruction and Development.

The high rate of increase in electrical energy demand during the first half of the 20th century led to a corresponding rise in the installed capacity of generating plants. At first, an apparent abundance of coal and much undeveloped waterpower in the industrialized countries allowed production to advance without concern over the exhaustion of resources. The discovery of large reserves of petroleum, also used for electricity generation in some thermal power stations, had the same effect. As the rate of use of these resources in relation to the total proved reserves rose rapidly, however, the position became disquieting. In the decade following World War II the electricity-supply industry was, therefore, greatly concerned with the introduction of new sources of energy. While many sources were studied intensively, it soon became clear that nuclear energy offered the greatest promise as a major future source. The discovery of huge resources of petroleum, however, especially in the Middle East in the late 1950s, changed the energy situation.

Use of oil for power

In the 1960s, petroleum proved to be an abundant and cheap fuel, while domestic coal in many developed countries increased in cost and hydroelectric resources neared exhaustion. Thus many developed countries imported large amounts of Middle East oil for power production. Development of nuclear power was delayed by this cheap oil. Europe in general came to rely on imported oil for some 40 percent of its total energy, and in Italy the rate of dependence reached 70 percent or more in 1969 and nearly the same in Japan. After a decade of cheap oil, the price of oil rose sharply, beginning in the 1970s, because of political changes in many Middle Eastern and African oil-producing countries.

WORLD ELECTRIC POWER

Total electric power production in the world was 4,568,000,000,000 kilowatt-hours and per capita production 1,290 kilowatt-hours in 1969. The situation in various regions is given below.

North America. *United States.* In the early days of the utility industry, private power companies and municipally owned utilities were responsible for the production and distribution of electricity. From the 1930s, federal agencies, public power districts, and rural cooperatives became part of this industry.

In 1960 the number of private utilities was about 480, that of municipal- and state-owned about 2,100, and rural cooperatives about 900. Earlier the number of private utilities had been more than 4,000 and that of municipal- and state-owned more than 3,000.

Although most of the generating capacity was owned by private utilities, their share decreased year by year until about 1961, after which it remained practically constant. Of the total installed capacity in 1936, 91 percent was owned by private companies; this was reduced to 80 percent in 1946, and slightly further since. Federally owned utilities increased from 2 percent in 1936 to 13 percent in 1961.

In 1969 total production of electricity in the United States was 1,551,900,000,000 kilowatt-hours, of which 110,000,000,000 kilowatt-hours was generated by industrial plants and the remainder by utilities. Of the latter, private companies accounted for 76.4 percent, federal facilities 12.8 percent, municipal and state plants 4.8 percent, electric power districts 4.7 percent, and rural cooperatives 1.2 percent. Production per capita was 7,500 kilowatt-hours. Installed generating capacity was 331,900 megawatts, divided much the same as the above percentages. Hydroelectric stations accounted for 16 percent of this capacity and thermal stations, including nuclear and geothermal (tapping steam from underground), for 84 percent.

In the early 1970s there were about 400 privately owned electric utilities in the United States, ranging widely in size, regulated by state commissions or city governments for intrastate or intracity operations. All states had established regulatory bodies by the early 1960s. The state laws require that power companies be permitted to earn a reasonable profit on the part of their investment that is devoted to public use. State agencies determine a reasonable profit level by considering such factors as the cost of borrowing money, the ability to attract money for expansion, the company's financial condition, the effects of inflation, the physical condition of the property, the degree of investor risk, and overall company operations.

Regulation of utilities by state agencies

Regulatory commissions function to protect utility customers from poor service or unfair charges resulting from a misuse of the monopoly status of the utilities and to assure fair treatment for utility investors. Not only are investors a part of the public but good service depends upon the willingness of investors to supply money to the private utilities.

In addition to state and municipal agencies, the Federal Power Commission (FPC) has authority to regulate the rates and conditions of service in which electricity is sold at wholesale in interstate commerce.

Rural electric cooperatives, owned and operated by the people they serve, provided service in the 1960s in nearly every state. More than 96 percent of all farms were electrified, compared with 10 percent when the Rural Electrification Administration (REA) was established in 1935, and cooperatives served 51 percent of these farms.

The FPC issues permits for planning and licenses for construction and operation of nonfederal hydroelectric projects on waters or lands subject to United States jurisdiction. The FPC also issues permits for facilities on the international borders that are to be used to export or import electrical energy to or from Canada or Mexico and reviews and approves proposed rates for the sale of power from certain federal and international projects.

The Federal Water Power Act of 1920 limited the licenses for hydroelectric power sites to 50 years and provided for recapture by the federal government at the end of that period. The act also contains a preference clause stipulating that states, municipalities, or other political subdivisions be granted preference over private applicants in the licensing of proposed hydroelectric projects. The preference clause is used only when the proposals of both public and private applicants are of equal merit.

In the original Reclamation Act of 1902, the Department of the Interior was authorized to begin construction

of water-use facilities to provide for irrigation and municipal water supplies in the arid regions of the West; there was no mention of the production or sale of electric power. In 1906 Congress amended the act so that surplus power generated at the dams of irrigation projects could be leased, giving preference to municipal purposes, for periods not exceeding 10 years.

The Department of the Interior has four subdivisions dealing with the generation and sale of electric energy: the Bureau of Reclamation, the Bonneville Power Administration, the Southwestern Power Administration, and the Southeastern Power Administration. The total amount of energy marketed by the Bureau of Reclamation was approximately 43,000,000,000 kilowatt-hours and that by the Bonneville Power Administration approximately 47,000,000,000 kilowatt-hours in 1969.

The Bureau of Reclamation is concerned primarily with the irrigation of arid and semi-arid lands, and the power it markets is generated at its irrigation projects, at flood-control projects of the Army Corps of Engineers, and at certain reservoir projects of other federal or international agencies. In the late 1960s, the bureau operated about 40 power plants with a total capacity of 7,333 megawatts.

The three remaining Department of the Interior subdivisions deal primarily with power production and marketing. The Bonneville Power Administration was created in 1937 to market power at the Bonneville Dam on the Columbia River in Oregon; later it was given responsibility for distributing power produced at more than a dozen other dams in the Northwest. The Southwestern Power Administration was created in 1943 to market power generated at Corps of Engineers' projects in Arkansas, Missouri, Oklahoma, and Texas, while the Southeastern Power Administration was established in 1950 to sell power generated at 10 Corps of Engineers' projects in the southern and southeastern states.

The objective of all four marketing agencies is to dispose of power from federal projects so as to encourage the most widespread use of electricity at the lowest possible rates to consumers consistent with sound business principles, giving preference to public bodies and rural cooperatives.

The TVA

The Tennessee Valley Authority (TVA), created by an act of Congress on May 18, 1933, developed into the largest single electric power system in the world, with an installed capacity of more than 18,000 megawatts. It is charged with the development of the Tennessee River and its tributaries in the interests of navigation, flood control, and the generation and distribution of electric power. In the late 1960s it was producing annually more than 86,000,000,000 kilowatt-hours, of which more than three-quarters was generated by steam-electric plants and the remainder by hydroelectric plants.

Canada. Development of electricity supply, based on abundant waterpower, is undertaken through provincial power commissions and private undertakings. Some of the commissions are responsible for generation as well as transmission and distribution, while others buy and distribute energy in bulk.

In 1970 the installed capacity of Canadian electric power industries was divided as follows: provincial commissions, over 70 percent; private utilities, 10 percent; and industrial plants, 15 percent. Annual electricity production was more than 203,700,000,000 kilowatt-hours, about three-quarters (77 percent) of which was produced by hydroelectric plants. Total generating capacity was 41,300 megawatts, with hydropower representing 66 percent and thermal power (including nuclear power and gas turbines) 34 percent.

Canada is known for hydroelectric power and extra-high-voltage transmission. The Churchill Falls station in Labrador, one of the largest hydroelectric power stations in the Western world, has a capacity of 5,225 megawatts (11 units of 475 megawatts each). Under consideration in the 1970s was the James Bay project in Quebec, tentatively planned for 17,000 megawatts by 1985. A 500-kilovolt transmission line that carries the hydroelectric power of the Peace River project in British Columbia leads to Vancouver and connects with the Pacific Intertie in the United States. Large amounts of hydroelectric power from the dams in the Manicouagan and Outardes rivers have been transmitted by three-circuit, 735-kilovolt transmission lines, the world's highest voltage transmission lines, to Montreal via Quebec since 1965.

The cost of electricity in Canada is about half that in Europe, but the difference was gradually being reduced in the 1970s. Per capita consumption of electricity in Canada in 1970 was roughly 8,500 kilowatt-hours, second only to Norway.

Europe. Total electric power production in Europe in 1969 was 1,214,100,000,000 kilowatt-hours, about 2,690 kilowatt-hours per capita.

United Kingdom. Municipal and private organizations were governed by various electricity acts from 1900 onward. By 1925, 572 organizations provided electricity and 491 generating stations had a total capacity of 4,422 megawatts.

World War I demonstrated the need for national planning, and the Electricity Act of 1926 established the Central Electricity Board to be responsible for transmission facilities and interconnection of selected generating stations giving bulk supplies to distribution authorities. The construction of a 132-kilovolt national grid, standardized at 50 hertz, was completed by 1938.

After World War II the shortage of generating plants and the need for reorganization became obvious. For several years, load shedding, including blackout (load cut) and brownout (voltage reduction), and the discouragement of energy consumption were resorted to. Not until 1955 did such steps become unnecessary.

Nationalization of the electric industry

In April 1948 the electric industry was nationalized, and about 95 percent of the generation came under government control. In 1957 the industry was reorganized, and the Electricity Council as a planning board; the Central Electricity Generating Board, dealing with electricity generation and transmission; and 12 area boards for electricity distribution were established. Also the North of Scotland Hydro Electricity Board and the South of Scotland Electricity Board were established.

The number of generating stations was reduced and new ones were built, under a succession of annual construction programs that increased the total installed capacity to 24,235 megawatts at the end of 1957. In 1955 a national 10-year program for the construction of 12 nuclear power stations, with a total capacity of 6,000 megawatts, was formulated. Construction started in 1954 on a 275-kilovolt supergrid the total circuit length of which reached 3,932 miles in 1964. In 1963 construction of a new 400-kilovolt supergrid started, with a change from 275-kilovolt lines to 400-kilovolt lines. In 1969 there were 2,340 miles of 275-kilovolt lines and 6,115 miles of 400-kilovolt lines. Total generating capacity had increased to 61,370 megawatts, mostly thermal power (hydropower less than 3 percent). The share generated by industrial plants was about 20 percent in 1955 and decreased to about 10 percent in 1968.

Nuclear power generation in the United Kingdom was the largest of any country in the world in the early 1970s. After a 10-year program, total capacity in eight nuclear power stations was 4,680 megawatts and electricity production 28,500,000,000 kilowatt-hours in 1969. This nuclear power production was about 12 percent of the total power production in the United Kingdom and about one-third of the total nuclear power generation in the world. In the second nuclear power program (from mid-1960s), five nuclear power stations with advanced gas-cooled reactors (total capacity: 6,600 megawatts) were under construction.

Another special feature of the United Kingdom utility industry is the very low load factor of the system. The yearly load factor lies between 60 percent and 70 percent in many countries, but in the United Kingdom it was formerly around 45 percent for long periods and only in the late 1960s rose to 55 percent. To increase off-peak demand, use of electric room heating and electric water heaters is being encouraged by introducing special tariffs (rates).

In Northern Ireland responsibility for electricity supply

is vested in a separate supply board, together with Belfast Corporation, and there are small independent undertakings in the Isle of Man and in the Channel Islands.

France. Since 1935 the national government has made active efforts to lower electric rates. In April 1948 electricity and gas industries were nationalized, and Électricité de France (EDF) was organized, integrating generation and distribution. Generating facilities both from private utilities and industrial plants having a larger capacity than a specified level were transferred to the EDF, except for the Rhône Valley Authority, the National Railways, and the French Coal Corporation.

Increases in French generating capacity

Generating capacity increased steadily: 9,020 megawatts in 1936, 9,540 in 1947, 16,770 in 1956, 21,430 in 1961, and 36,330 in 1969. Total electric energy generated in 1969 was 131,300,000,000 kilowatt-hours, of which 52,700,000,000 was from hydro and 78,600,000,000 from thermal power stations. Per capita generation in 1969 was 2,600 kilowatt-hours.

Export and import of electrical energy are carried out with UCPTE countries and Spain and via Channel cable with Britain. The Rance station on the Rance River Estuary in Britanny is the world's largest tidal power station, with 240 megawatts capacity and a yearly output of nearly 500,000,000 kilowatt-hours.

West Germany. There were, in the early 1970s, approximately 4,000 utilities and nearly 4,100 industrial installations generating some or all of their own electricity in West Germany. Though most of the electric utilities were private enterprises, nearly all were partly financed by public organizations, and purely private industry represented only 4 percent of the total kilowatt-hours. Generating capacity showed a steady increase: 8,320 megawatts in 1936, 9,250 in 1941, 19,180 in 1956, and 50,830 in 1970. In 1970 thermal power contributed 91 percent of generating capacity, the remainder being hydro. Electricity production in 1970 was 93 percent thermal and 7 percent hydro.

Industrial plants contribute a very large share (more than 30 percent) of total power production. Part of this industrial power is fed into the utility network, accounting for as much as 15 percent of total output.

West Germany imports about 7,000,000,000 kilowatt-hours per year, relying heavily on the interconnection by UCPTE. Suitable sites for pumped storage stations are scarce, and some serving West Germany are located in other countries, such as the large Vianden station in Luxembourg. Total capacity of the pumped storage plants used by German utilities is 3,800 megawatts, of which about 2,000 megawatts is outside the country.

Pumped storage is one method of using power in off-peak periods. Room heating and water heating loads using off-peak power are also promoted at special rates; a secondary advantage is air-pollution abatement.

Italy. Before nationalization of the electric power industry in 1963, nearly 75 percent of the total electricity was produced by private utilities. With nationalization, a single government organization was established to generate and sell almost all the electricity in the country. Although hydroelectric capacity has remained essentially constant, its share of total capacity has fallen—82 percent in 1960, 50 percent in 1965, and only 36 percent in 1970—because of the tremendous advances in thermal power generation, from 10,000,000,000 kilowatt-hours in 1960 to 70,000,000,000 kilowatt-hours in 1970, a sevenfold increase. Oil provided the main fuel for this increase, and consumption of oil (mostly imported) in the 1960s increased nearly 10 times.

Italy produces sizable amounts of geothermal power. The first such production there and in the world began in 1905. Capacity in the early 1970s was nearly 400 megawatts and yearly production nearly 3 percent of the total electricity produced. Total installed generating capacity in 1970 was about 33,200 megawatts, and total production was 177,400,000,000 kilowatt-hours, representing nearly 2,200 kilowatt-hours per capita.

Norway. With abundant hydroelectric potential, almost all of Norway's electricity comes from hydroelectric plants. In 1969 about 57,000,000,000 kilowatt-hours of electricity was generated, the world's highest per capita generation, 14,680 kilowatt-hours.

Eastern European countries. Only total power production and per capita generation figures of eastern Europe in 1969 are given here, for brevity's sake. These were: German Democratic Republic, 65,400,000,000 kilowatt-hours and 4,070 kilowatt-hours per capita; Poland, 60,000,000,000 and 1,840; Romania, 31,500,000,000 and 1,580; Czechoslovakia, 43,000,000,000 and 2,990; Hungary, 14,000,000,000 and 1,380.

Soviet Union. After the 1917 Revolution much effort was devoted to increasing the production of electric energy. In 1920 the State Commission for the Electrification of Russia (Goelro) was founded, and in 1927 the First Five-Year Plan was inaugurated. In 1971 the Ninth Five-Year Plan was started.

Before the Revolution, the capacity for electric power generation was very small (1,100 megawatts in 1913). By 1936 it had reached 7,500 megawatts. During World War II, generating capacity increased tremendously, and by 1951 it had reached 22,100 megawatts, second only to the United States. By 1961, capacity reached 74,000 megawatts, nearly ten times more than in 1936.

Total electricity production in 1970 was 740,000,000,000 kilowatt-hours, of which hydro was 17 percent and thermal 83 percent. Total installed capacity in 1970 was 166,000 megawatts, 19 percent hydro and 81 percent thermal.

Hydroelectric power in the Soviet Union

Although the share of hydroelectric generation in the total production is fairly small, Soviet technology is characterized by the construction of giant hydroelectric stations. In the 1940s and 1950s large hydroelectric stations such as the station at Kuibyshev (2,300 megawatts) and the Volgograd station (2,530 megawatts), were constructed; later the Bratsk station (5,000 megawatts, 1967) and the Krasnoyarsk station (6,000 megawatts, 1971) were placed in operation, the latter the world's largest in the early 1970s. The Sayan-Shushenskoye station (6,400 megawatts) was under construction.

The Soviet Union also has large thermal power stations supplying both electricity and heat (hot water and steam) for residential and industrial use. Total capacity of thermal power stations of this type was about 35,000 megawatts and total heat energy supplied was about 5.1×10^{17} calories in 1970.

Because of the Soviet Union's immense land area, extra-high-voltage transmission is a necessity. Moreover, the electric loads are mostly concentrated in European Russia, whereas fuel resources and hydro sites are located mainly in Siberia, and the distance between these two regions is over 1,250 miles (2,000 kilometres).

The Soviet Union had in the early 1970s more than 6,000 miles (10,000 kilometres) of 500-kilovolt lines, a short length of 750-kilovolt lines and 370 miles (600 kilometres) of dc 800-kilovolt lines. For extra-long-distance transmission, 1,150-kilovolt ac transmission and ± 750-kilovolt (1,500-kilovolt) dc transmission (for 1,600 miles) were under extensive study.

Asia. This area, containing almost half of the world's population, was in the 1970s largely underdeveloped in electricity supply except for a few countries, notably Japan, Israel, and North Korea.

The total electric power production in Asia for 1969 is given as 463,000,000,000 kilowatt-hours in the *United Nations Year Book*, but exact figures for China, North Korea, and North Vietnam are not available. Per capita generation, excluding the above three countries, was low, 420 kilowatt-hours in 1969. Because nearly two-thirds of the Asian electric power is produced in Japan, per capita generation in Asia, excluding Japan, was only 170 kilowatt-hours in the same year.

Japan. Several hundred private electric utilities existed in Japan before World War II. During the war, they were incorporated into one Electric Power Generation and Transmission Company and nine regional electric power distribution companies, all privately owned but government-controlled. The company owned about 80 percent of the total capacity of hydro and thermal stations and all of the main transmission facilities. After

the war, however, these government-controlled companies were disbanded, and in 1951 nine regional electric power companies were established, which dealt with the whole field of electric power generation, transmission, and distribution as purely private utilities.

Total installed capacity in 1952 was 10,800 megawatts, of which 6,800 was hydro and 4,000 thermal. In that year, hydro power production was about 2.5 times higher than thermal power production. By 1962–63 the hydropower and thermal power became equal, both in installed capacity and in electricity production. Later, thermal power increased rapidly, and in 1970 the installed capacity was 40,200 megawatts while that of hydropower had increased only to 19,300 megawatts. Per capita generation was 2,450 kilowatt-hours.

Increases in Japanese power production

A special feature of the Japanese electric power industry is the rapid increase in capacity and production. Its increase in the 1950s and 1960s was sevenfold, far surpassing the increase of most other developed countries, placing the Japanese industry in third place in world production since 1966. This rapid increase was made possible by developments in the capacity of thermal power stations fuelled by imported oil.

India. Although electricity was generated in 1903 under the Indian Electricity Act, progress was slow until nationalization of the industry in the 1940s. Both the central and state governments undertook developments, and the annual production of electric energy for public consumption rose from about 2,400,000,000 kilowatt-hours in 1939 to 56,200,000,000 in 1969, though this figure was still only 105 kilowatt-hours per capita. Installed capacity was about 15,500 megawatts, of which 6,130 megawatts was hydro.

Central and South America. Except in large cities electricity consumption remained generally low in the 1970s. The annual electric energy production was 90,300,000,000 kilowatt-hours, and per capita production 350 kilowatt-hours in 1969.

In the past, supplies were developed mainly by private enterprise or municipal authorities, but starting in the 1950s national authorities absorbed an increasing share. Coal resources are very small, but petroleum is produced significantly, especially in Venezuela. Hydroelectric potential (estimated at some 60,000 megawatts, based on ordinary minimum flow) is high, and in the late 1960s development of large hydroelectric plants was undertaken in several countries, such as the Furnas project (final 1,200 megawatts) in Brazil and the Guri project (final 6,000 megawatts) in Venezuela.

Africa. The electric supply situation in this area is quite varied. Relative to the rest of the world, hydroelectric resources are much larger at 187,000 megawatts (at ordinary minimum flow), though only a small fraction of these resources had been developed in the early 1970s. Discoveries of oil and natural gas in the mid-1950s improved the prospects for thermal plants, particularly in Algeria.

The annual electricity production in 1969 for the entire continent was 74,000,000,000 kilowatt-hours and per capita generation 230 kilowatt-hours. Excluding South Africa, this figure was less than 100 kilowatt-hours.

In the Republic of South Africa the Electricity Supply Commission, set up in 1922, is mainly responsible for electrical development. The commission has built large thermal power stations, and electricity, generated from local coal, is cheap. In 1969 the annual production was 44,200,000,000 kilowatt-hours and per capita generation 2,080 kilowatt-hours.

Oceania. *Australia.* The public electricity supply was developed mainly under state electricity commissions with their own networks. Coal, brown coal, and lignite were used for generation, but waterpower, particularly in the eastern section of the Australian Alps and Tasmania, was being developed in the 1970s. The Snowy Mountains hydroelectric scheme, with installed capacity of nearly 2,000 megawatts, produced 2,300,000,000 kilowatt-hours of electric energy in 1969. Total installed electrical generating capacity in Australia in 1969 was 12,260 megawatts, of which hydro's share was 3,600 megawatts. Total electric energy production was 48,700,000,000 kilowatt-hours, and per capita generation was 3,970 kilowatt-hours in 1969.

New Zealand. Regional power boards have long been responsible for source development, and in 1945 a Rural Electrification Reticulation Council was established to coordinate rural supplies. With abundant waterpower, progress in electricity production has been rapid, and in 1969 nearly 80 percent of the total electric energy for public supply was generated by hydro stations. Installed capacity was 3,568 megawatts, of which 777 megawatts was thermal. Per capita generation was 4,660 kilowatt-hours in 1969.

In the North Island little hydroelectric power remains to be developed, and geothermal power is used. The installed capacity of geothermal plants was 192 megawatts and annual production nearly 1,200,000,000 kilowatt-hours in 1969.

Hydroelectric power produced in the South Island was introduced into the North Island through a dc cable across the Cook Strait.

BIBLIOGRAPHY. NATIONAL ELECTRICAL MANUFACTURERS ASSOCIATION, *A Chronological History of Electrical Development from 600 B.C.* (1946), a compilation of historical developments in electrical engineering; CHARLES A. POWEL, *Principles of Electric Utility Engineering* (1955), a work suitable for the layman on all phases of the electric power industry; HAROLD PENDER and WILLIAM A. DEL MAR (eds.), *Electrical Engineers' Handbook*, vol. 1, *Electric Power*, 4th ed. (1953), a handbook covering the entire field of electric power engineering; JOHN G. BROWN (ed.), HYDRO-ELECTRIC ENGINEERING PRACTICE, 3 vol. (1958), comprehensive texts on hydroelectric engineering covering hydrology, civil engineering (dams), mechanical engineering (turbines), and economics in hydroelectric power generation; GEORGE E. REMP, *Fundamentals of Power Plant Engineering* (1949), mainly devoted to the problems of the thermal power plant, from the standpoint of mechanical engineering; ROBERT L. LOFTNESS, *Nuclear Power Plants: Design, Operating Experience, and Economics* (1964), a text dealing mainly with nuclear reactors, including a description of engineering principles and various reactor types and the economics of nuclear power; MENDELL P. WEINBACH, *Electric Power Transmission* (1948), a text dealing with calculational methods of electrical characteristics of ac transmission lines; SWEDEN, VALLENFALLSSTYRELSEN, *The Swedish 380 kv System* (1960), a description of the world's first 380 kv transmission system, with its developments and operational data; SELDEN B. CRARY, *Power System Stability*, 2 vol. (1945–47), text dealing with steady-state and transient stability; RICHARD A. HORE, *Advanced Studies in Electrical Power System Design* (1966), a text dealing with modern analytical methods for the problems in the design and operation of electric power systems; WESTINGHOUSE ELECTRICAL CORPORATION, *Electric Utility Engineering Reference Book*, 3 vol. (1959–65), a reference book covering all aspects of electric power engineering, especially suited for engineers in the utility industry; COLIN ADAMSON and N.G. HINGORANI, *High Voltage Direct Current Power Transmission* (1960), a text dealing with fundamental principles and operational problems in dc transmission, with some examples of actual installations; FRANK E. SANFORD, *Electric Distribution Fundamentals*, 2nd ed. (1947), a text for electrical distribution problems, with historical description, design, and operation of electric distribution systems included; HARRY COTTON, *The Transmission and Distribution of Electrical Energy*, 2nd ed. (1943), a text dealing with electrical and mechanical characteristics of electric power transmission and distribution, with a slight reliance on mathematical treatment; UNION FOR THE COORDINATION OF THE PRODUCTION AND TRANSPORT OF ELECTRIC POWER (UCPTE), *Annual Report*, a series covering the activities of UCPTE committees and giving data on electric power interchange between UCPTE countries; UNITED NATIONS, *Proceedings of the International Conference on the Peaceful Uses of Atomic Energy*, 16 vol. (1955–56) and *Proceedings of the Second United Nation's International Conference on the Peaceful Uses of Atomic Energy*, 33 vol. (1958), compilation of papers and discussions from these two conferences; *World Energy Supplies* (UN, annual), a series of statistics on world energy supplies, including those of electrical power installations and electric energy output on a country-by-country basis; UNITED STATES FEDERAL POWER COMMISSION, *National Power Survey*, 2 vol. (1964), a good, comprehensive survey of the U.S. electric power industry, dealing with its present and future status—provides the layman with a clear explanation of all phases of utility problems.

(To.Ya.)

Electrochemical Reactions

In the most general terms, electrochemical reactions are those chemical changes that are associated with the passage of electrons across interfaces between metal and solution. Under ordinary conditions, the occurrence of a chemical reaction is accompanied by the liberation or absorption of heat and not of any other form of energy; but there are many chemical reactions that when allowed to proceed in contact with two electronic conductors, separated by conducting wires liberate what is called electrical energy and an electric current is generated. Conversely, the energy of an electric current can be used to bring about many chemical reactions that do not occur spontaneously. A process involving the direct conversion of chemical energy when suitably organized constitutes an electrical cell. A process whereby electrical energy is converted directly into chemical energy is one of electrolysis; *i.e.*, an electrolytic process. By virtue of their combined chemical energy, the products of an electrolytic process have a tendency to react spontaneously with one another, reproducing the substances that were reactants and were therefore consumed during the electrolysis. If this reverse reaction is allowed to occur under proper conditions, a large proportion of the electrical energy used in the electrolysis may be regenerated. This possibility is made use of in accumulators or storage cells, sets of which are known as storage batteries. The charging of an accumulator is a process of electrolysis; a chemical change is produced by the electric current passing through it. In the discharge of the cell, the reverse chemical change occurs, the accumulator acting as a cell that produces an electric current.

Finally, the passage of electricity through gases generally causes chemical changes, and this kind of reaction forms a separate branch of electrochemistry, which will not be treated here (see GASEOUS STATE).

General principles. Substances that are reasonably good conductors of electricity may be divided into two groups; the metallic, or electronic, conductors and the electrolytic conductors. The metals and many nonmetallic substances such as graphite, manganese dioxide, and lead sulfide exhibit metallic conductivity; the passage of an electric current through them produces heating and magnetic effects but no chemical changes. Electrolytic conductors of electrolytes comprise most acids, bases, and salts, either in the molten condition or in solution in water or other solvents. Plates or rods composed of a suitable metallic conductor dipping into the fluid electrolyte are employed to conduct the current into and out of the liquid; *i.e.*, to act as electrodes. When a current is passed between electrodes through an electrolyte, not only are heating and magnetic effects produced but also definite chemical changes occur at or in the neighbourhood of the negative electrode, called a cathode. The chemical change may be either the deposition of a metal or the liberation of hydrogen and formation of a basic substance, or some other chemical reduction process; whereas at the positive electrode, or anode, it may be either the dissolution of the anode itself, the liberation of a nonmetal, the production of oxygen and an acidic substance, or some other chemical oxidation process.

Electronic and electrolytic conductors

An electrolyte, prepared either by the melting of a suitable substance or by the dissolving of it in water or other liquid, owes its characteristic properties to the presence in it of electrically charged atoms or groups of atoms produced by the spontaneous splitting up or dissociation of the molecules of the substance. In solutions of the so-called strong electrolytes, most of the original substance, or in some solutions perhaps all of it, has undergone this process of electrolytic dissociation into charged particles, or ions. When an electrical potential difference (*i.e.*, a difference in degree of electrification) is established between electrodes dipping into an electrolyte, positively charged ions move toward the cathode and ions bearing negative charges move toward the anode. The electric current is carried through the electrolyte by this migration of the ions. When an ion reaches the electrode of opposite polarity, its electrical charge is donated to the metal, or, an electric charge is received from the metal. The ion is thereby converted into an ordinary neutral atom or group of atoms. It is this discharge of ions that gives rise to one of the types of chemical changes occurring at electrodes.

History. The study of electrochemistry began in the 18th century, bloomed until the early 20th century, and then ebbed, owing to an excessive use of thermodynamic principles in analyzing the processes that take place at points in the system where the various parts form interfaces. Since about 1950 electrochemistry has undergone a change. The study of processes in solutions has been less stressed, but the study of the transfer of electrons between metals and solution has increased explosively. With this new emphasis electrochemistry is becoming a core science and basis to much of the technology. It promises to be an important part of the foundation of the ecology-oriented society of the future, since electricity as a fuel lacks those products that fall generally under the term pollution.

The first electrochemical reactions studied, in 1796, were those in the cell of silver and zinc plates with blotting paper wetted by aqueous salt solution between them; these cells were constructed by the Italian scientist Alessandro Volta, for whom the term volt was named. This cell was the first primary battery used for the production of electricity.

Michael Faraday formulated the laws of electrochemical stoichiometry, the science that deals with the application of laws of definite proportions and of the conservation of matter and energy to chemical activity. These state that a coulomb of electricity, a unit of charge, reacts with fixed quantities of a substance (*e.g.*, with 1.11800 milligram of silver ions) or else, that 1 gram equivalent of any substance reacts with 96,048 ± 2 coulombs. This latter number represents a fundamental quantity known as one faraday of electricity. The relationship between the chemical affinity of the reactants in the cell and the voltage of the cell when it is operating was precisely defined by the U.S. chemist Josiah Willard Gibbs in 1875 while the relation of this affinity to the potential of the electrochemical cell was initially formulated by the German physical chemist Walter Hermann Nernst in 1889.

Faraday's laws

The period 1910 to 1950 was that of decline in electrochemistry until it became limited mainly to the study of solutions. There was almost no progress in the understanding of electrochemical reactions outside of equilibrium conditions and reversibility, and knowledge of these was applied invalidly to reactions occurring at a net rate; *i.e.*, reactions not in equilibrium and not totally reversible. From about 1950 the study of electrified interfaces, with special reference to the study of the transfer of electrons (called electrodics), gained in importance and became the main aspect of electrochemistry. From about 1960, electrodics began to develop as an interdisciplinary area in the search for solutions to problems such as the source of energy in space flights from fuel cells, the stability of metals in moist environments, the electrochemical aspects of biological functions, extractions from mixtures, and the replacement of fossil fuels such as coal and petroleum and their by-products, by electricity produced or stored electrochemically in transportation.

The rise of electrodics

THE ELECTROCHEMICAL PROCESS

Interactions of matter associated with the passage of an electric current depend upon the characteristics of the negatively charged electron. As the basic particle of electricity, the electron has an affinity for positively charged particles of matter, protons, whether in atoms, groups of atoms, or molecules. This affinity is analogous to the chemical affinity that particles exhibit among themselves. In fact, all chemical reactions result from a shift in the electron structure of atoms, and free electrons can combine with particles of matter (reduction) or be released by them (oxidation). The quantitative relationship between the free electrons of an electric current and the particles of a substance in which they cause a reaction is

defined by the laws of Faraday (see above *History*). The substances that take part in electrochemical reactions called electrolytes, or ionic conductors, have been described above.

Electrons are available in large quantities in a relatively free (mobile) state only in substances called electronic conductors, among which metals are the most important. Thus, an electron conductor must be present as a basic component of any system in which electrochemical reactions are to occur. Furthermore, the availability of electrons in a conductor is limited by energy distribution to such an extent that electrochemical reactions take place only in the immediate vicinity of the electronic conductor's surface; *i.e.*, a few angstroms from the conductor into the solution. These reactions are, therefore, normally considered as occurring at the interface, or common boundary, between an electronic conductor, such as an electrode, and an ionic conductor of electricity, such as an electrolytic solution. Electrochemical reaction will take place, however, only to the extent that electricity can flow through such a system as a whole. To achieve this, it is necessary for the system to form a closed loop, electronically speaking. For schematic representation of these details, see Figure 1.

The closed loop of conductors

Location of electrochemical reactions. Electrochemical reactions are located where the electron conductor meets the ionic conductor; *i.e.*, at the electrode–electrolyte interface. Characteristic of this region, considered to be a surface phase, is the existence of a specific structure of particles and the presence of an electric field of considerable intensity (up to 10,000,000 volts per centimetre) across it; the field is caused by the separation of charges that are present between the two bulk phases in contact. For most purposes the surface phase can be considered as a parallel plate condenser, with one plate on the centre of the ions that have been brought to the electrode, at the distance of their closest approach to it, and with the second plate at the metal surface; between the two plates and acting as a dielectric (*i.e.*, a nonconducting material) are oriented water molecules. This structure is termed the electric double layer and is illustrated in Figure 2.

Thermal motion of the positive ions in the solution makes the condenser plate on the electrolyte side of the interface diffuse; *i.e.*, the ions are distributed in a cloud-like way. This condition justifies the division of the potential change between the bulk of metal and the bulk of electrolyte into two parts: first, that between the metal

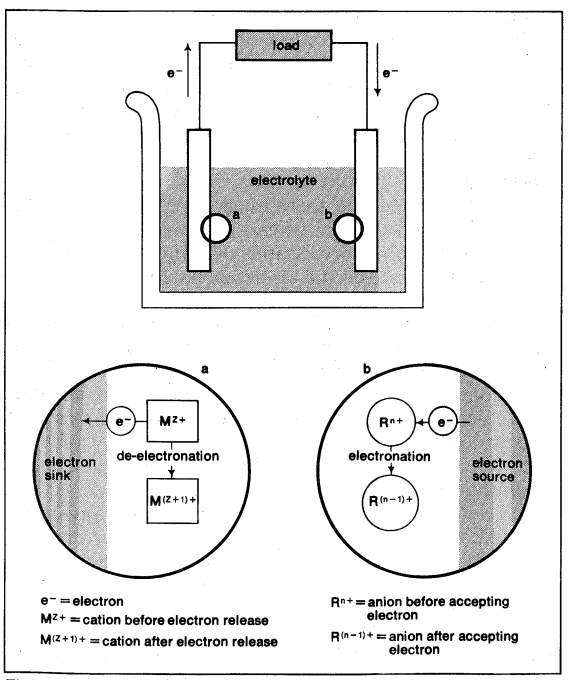


Figure 1: The mechanism of electron movement in an electrochemical cell.

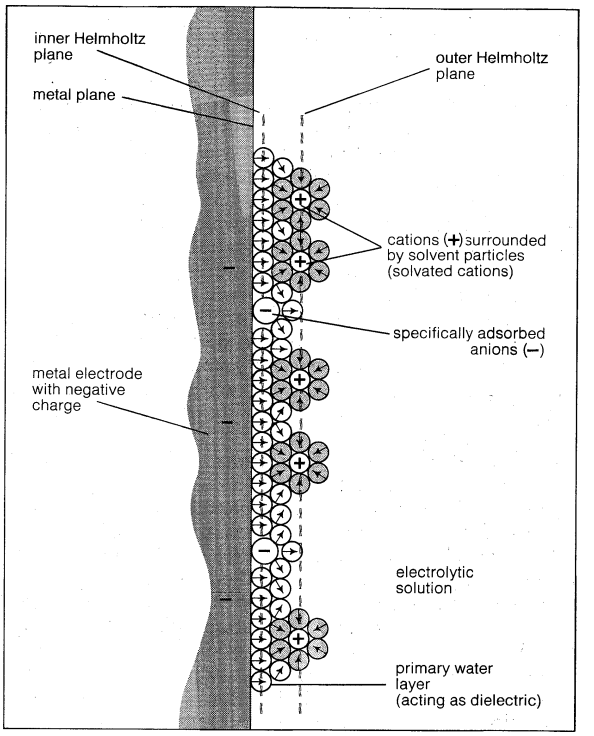


Figure 2: Double-layer structure and change of potential with distance from the electrode surface (see text).

surface and the first ionic layer at the distance of closest approach (called the outer Helmholtz plane, in which the ions are usually surrounded by solvent particles; *i.e.*, are solvated); and second, that between the first ionic layer and the bulk of the solution, the diffuse part of the double layer. The picture is further complicated by the presence of ions in the electrode surface layer in addition to those that are present for electrostatic reasons; *i.e.* by the force of attraction or repulsion between electric charges. Such electrode surface layer ions are said to be specifically adsorbed on the electrode surface. Since this species of ions is attracted by the surface to a distance closer than the "distance of the closest approach" of ions, further subdivision of the inner part of the electric double layer is justified. Hence, the inner Helmholtz plane is introduced as the plane formed by the centres of specifically adsorbed ions. Adsorption of neutral molecules on the surface can also change the properties of the electric double layer. This change occurs as a consequence of replacing the water molecules, and thus changes that part of the potential (electrical) difference across the double layer that is caused by the adsorbed dipoles (water molecules that have a polarity—*i.e.*, they behave like minute magnets—because of their hydrogen-oxygen structure, making one end of the molecule positive and the other end negative).

The absolute value of electrical potential difference, symbolized in calculation by the Greek letters delta and psi, $\Delta\psi$, between the bulk of a metal electrode and the bulk of an electrolyte cannot be measured. Instead, the voltage of a special cell, composed of the specific electrode being studied and of an arbitrarily selected reference electrode, is normally measured; the voltage is referred to as the relative electrode potential, E. Of special interest is that state of the electrode at which there is no net charge (in this case, no unbalanced, or extra positive, charge) at the metal side of the double layer. The relative potential at which this state is achieved is characteristic of each metal. This point is termed the potential of zero charge. At that potential, the field across the double layer is due to orientation of water molecules and other dipoles at the surface only.

Relative electrode potential

Most of the knowledge of the detailed structure of the interface between a metal and an electrolyte arises from experimentation with mercury, the only metal that is

liquid at ordinary temperatures; the double layer structure turns out to have surface tensions that must be measured, and this measurement is difficult with solid metals. By 1970, however, it had been shown that it is possible to measure surface tension changes at the metal-solution interface. Thus, the way to the determination of the double layer structure involving solids was opened.

Substances that are semiconductors can also be employed as electron carriers in electrochemical reactions. Semiconductors are substances which range between serving as insulators at low temperatures and as metallic-type conductors at high temperatures. In the case of semiconductors, however, the electric double layer has a more complex structure in as much as the condenser plate at the electrode side of the double layer also becomes diffuse. Thus, the overall potential difference between bulks of the phases in contact comprises also the potential difference between the bulk of the semiconductor and its surface.

Types of electrochemical reactions. There are several types of electrochemical reactions.

Simple redox reactions. A simple redox reaction is one that involves a change in the electrical charge of a charge carrier, usually a simple or complex ion in the solution, by its taking away, a process called reduction, an electron from the electrode, or its giving an electron to the electrode, a process called oxidation. The same carrier may be present in solution in two states of charge. The higher, more positive charge is called the oxidized state, and the lower, less positive charge is called the reduced state. For example, when ferric and ferrous ions are both present in solution in significant quantity, and when electron exchange with the electrode is sufficiently fast, redox equilibrium is established at the electrode, giving it a well-defined potential, or reversible redox potential.

To summarize, if at one metal-solution interface electrons are coming out of the metal, reducing a component of the solution, there must exist a second metal-solution interface where electrons are going into the metal in the process of oxidation.

The basic electrochemical cell

The two electrodes and the ionic conductor in between (*e.g.*, an aqueous solution of some ionized salt) represent an electrochemical cell. The process occurring in the cell as a whole is a redox process with the reduction of one species spatially separated from the oxidation of another one. As a consequence of Faraday's law, the rates of electrochemical reactions at electrodes (expressed in terms of gram moles per second per square centimetre of electrode surface) are directly proportional to the current density (expressed in amperes per square centimetre); *i.e.*, current flowing through the cell divided by the electrode surface area.

Reactions that produce gases. When hydrogen ions in solution react with electrons ejected from a metal, hydrogen atoms are formed at the surface, where they combine among themselves or with other hydrogen ions and electrons to give gaseous hydrogen molecules. If all the reactions are fast enough, an equilibrium is attained between hydrogen ions and gaseous hydrogen. A metal in contact with solution at which such a situation exists is called the reversible hydrogen electrode, and its electrical potential is arbitrarily taken to be zero; every other electrode can thus be compared with it as it represents the basis for constituting the hydrogen scale of relative electrode potentials. Similarly, negative hydroxyl ions in solution (OH^-) can be made to give up electrons to a metal and, in a series of reactions, the final one is the formation of gaseous oxygen. Chlorine is another gaseous product; it evolves upon electrochemical oxidation of chloride ions in concentrated solutions of neutral and acid salts.

Reactions that deposit and dissolve metals. When a metal ion is reduced and discharged as a neutral atom, or species, it tends to build into the metal lattice of the electrode. Thus, metals can be deposited at electrodes. Conversely, if electrons are taken away from the metal electrode by applying positive potentials to it, the metal ions thus formed can cross the double layer of electric charge at the interface, undergo hydration (combination with water), and enter the solution. The metal electrode thus dissolves. Many metals establish well-defined electric potentials when they are in contact with their own ions in solution.

Oxidation and reduction of organic compounds. A reaction of the oxidation and reduction of organic compounds can also be done at electrodes. Such reactions, however, are mostly irreversible in the literal sense that they lead to products that cannot easily be converted back into the original substance. Exceptions are some oxygen- and nitrogen-containing compounds (quinones, amines, and nitrous-compounds), that can give fairly well-defined reversible potentials.

Mechanism of charge transfer. The causes of the thermodynamically irreversible behaviour of electrode reactions are found in the nature of the elementary act of charge transfer. Like any chemical reaction, this act is inhibited by the existence of an energy barrier between the oxidized and the reduced state. This barrier implies that the reaction could take place only in the special circumstances when, during the course of numerous interactions with other species (atoms, ions, molecules, etc.) surrounding it, a molecule attains an excited state in which it has an abnormal energy content. In most chemical reactions, this energy content must be sufficient for the species to come into what is called the transition state; the transition state characterizes the top of the energy barrier just before a reaction begins. If such a model is applied to electron transfer at an interface, calculation shows that electron exchange reactions at electrodes would be prohibitively slow, a conclusion at variance with the observed phenomena; quantum mechanical laws, however, govern the motion of electrons, and their inclusion changes the calculations to fit reality. Quantum mechanics require that for fast electron exchange to take place, electrons in a particle outside the double layer (*e.g.*, a hydrated ion at the outer Helmholtz plane) must attain certain well-defined quantized energy levels equal to those in which free electrons exist in the metal. Since such states can be attained by the particle at a lower energy-content than that needed for its transfer over the top of the energy barrier, according to the classical view, this fast process of electron exchange between the electrode and a particle in solution is termed electron tunnelling through the energy barrier.

Electron tunnelling

Whereas the rate of chemical processes or, loosely, speed of reaction, can be influenced only by changing the concentrations of reactants or by changing the temperature or both, the rate of electrochemical processes also can be manipulated by changing the electrode potential. Making the electrode more negative increases the number of electrons in the metal ready to *tunnel* to ions and hence the rate of the reduction process increases. Conversely, making the potential more positive decreases this rate and increases the number of particles ready to give away electrons, thus increasing the rate of the oxidation process.

It can be deduced that there must exist a direct proportionality between the rate of reaction and the concentration of the reacting species and also an exponential dependence of the rate on the electrode potential.

At any electrode potential, both reduction of one species and oxidation of the product of reduction are taking place but at different rates; the rate of each reaction is determined by the respective concentration and by the corresponding effects of potential. The rate of an electrochemical reaction can best be described as the electric current density; *i.e.*, a measure of the quantity of electrons moving in a certain volume of space during a specified unit of time.

The relationships can be represented quantitatively by an equation in which the net, or resulting, current (the difference of the rate of electron ejection across the interface to particles in solution, diminished by the rate at which particles in solution inject electrons into the metal) is equated to the difference of the rates of reduction and oxidation and the variables and constants that relate to these reactions. See equation (1), and other equations below in *Calculations* at the end of this article.

Equation (2), known as the Nernst equation, which can be derived from equation (1), gives the value of the electrode potential when the rate of oxidation exactly equals the rate of reduction. Using this value for the electrode potential, equation (3), called the Butler-Volmer equation, can be derived; it represents the most fundamental relationship in electrodic chemistry.

COMPLEX ELECTROCHEMICAL REACTIONS

Electrochemical processes considered so far involve simple reactions of a particle with a single electron to produce a reduced ion (*e.g.*, the ferrous ion of iron with two positive charges, Fe^{++}), or vice versa. Such are the simple ionic redox processes, where the only difference in structure between a reactant and a reaction product may be due to some rearrangement of the neighbouring solvent molecules. When one or more transfers of electrons between the electrode and a species in solution are accompanied by major structural changes (*e.g.*, when hydroxyl ions, OH^-, transform into a molecule of oxygen, O_2, and a water molecule, H_2O, in the process of oxygen evolution at the anode, or positive electrode), the reaction usually consists of a sequence of events, called elementary acts, or unit steps, constituting the reaction mechanism. Intermediate states between the steps usually involve some unstable intermediate species with higher energy content than those of the reactants or of the reaction products.

Complex reaction mechanism

Complex reaction mechanisms can consist of a number of electron transfer steps, with some chemical steps preceding or succeeding the electron transfer steps or taking place in between them. Most organic electrochemical reactions are complex, involving large numbers of electrons in the overall reaction. Usually one step in the reaction encounters the largest energy barrier. The rate of occurrence of this step limits the rate of the overall reaction (*i.e.*, all other steps must occur at the same net rate, although they could provide for a much faster overall change). This step is called the rate-determining step and, for most practical purposes, all intermediate steps before and after it can be considered to be in equilibrium. (It is interesting to note that whenever this is the case, the Butler-Volmer equation [equation 3 below] is applicable, but with specific values of the transfer coefficients α_a and α_c characteristic of the mechanism of the reaction.)

Experimental studies. Measuring the rates of electrochemical reactions (*i.e.*, current densities) as functions of electrode potential under steady-state conditions represents the normal tool of electrodics. Meaningful results could not be obtained, however, until the sensitivity of electrochemical reactions to impurities was realized and high purity techniques were introduced. Even so, the steady-state method often has shortcomings except for relatively slow electrode reactions. In many cases concentration changes at the electrodes prevent using a sufficiently wide current density range for obtaining meaningful Tafel relationships (see Butler-Volmer equation below under *Calculations*). Hence, so-called transient methods have been developed in which one electrochemical factor in the situation is rationally perturbed and the time dependence of others observed. One such method consists of placing a constant current pulse upon an electrode and measuring the variation of the resulting current through the solution. This is called the galvanostatic method for measuring the rate of an electrochemical reaction. Applying a potential pulse while observing the variation of the rate as a function of time constitutes the potentiostatic method. A third method, called the potentiodynamic, or potential sweep, method involves observations of the current as a function of the potential, while the latter is varied at a constant, known rate.

Galvanostatic, potentiostatic, and potentiodynamic methods

The advantages of transient methods over steady-state ones, in which behaviour before the attainment of the steady-state is not part of the observation, are manifold. If observations are made at sufficiently short times, events can be recorded before the onset of concentration changes, and pure activation values can be found. Hence, Tafel relationships can be obtained over a larger current density range (see below *Calculations*) than if one makes measurements over longer times, as is required in the steady-state methods. The structure of the transient states can reveal important information, such as double layer capacitance and surface coverage of the electrode by intermediate species.

Several so-called kinetic parameters, for example, partial derivatives of current density and potential with respect to concentration of chosen reactants, can be extracted from experimental measurements.

Electrochemical measurements have a limited capacity to reveal the state of an electrode surface. Nonelectrochemical methods of studying electrode surfaces, therefore, have been stressed. Optical methods have considerably gained in importance. Ellipsometry (*i.e.*, measuring changes in basic properties of polarized light as it is reflected from an electrode surface) was the first method that made possible a study of monomolecular layers of oxides and adsorbed oxygen as adsorbed organic molecules. Adaptation of such a method to transient use allows change in the surface to be related to the passivity of metals. Another kind of spectroscopy enables infrared spectra of species adsorbed at electrode surface to be taken. Mössbauer spectra may lead to an identification of thin layers on an electrode surface. A good future is seen for further development of refined techniques for the study of electrochemical processes by various combinations of spectroscopic and electrochemical means.

TYPES OF COMPLEX REACTIONS

Electrocatalysis. The problems related to the increase of rates of electrochemical reactions, or, to put it another way, the decrease of overpotential, needed to perform reactions at a given rate are the subject of electrocatalysis. Both increase and decrease are of considerable practical importance since they affect the economics of electrochemical processes. Electrocatalysis is concerned with the electrode as a substrate, or base, for electrochemical reaction and with the effect of its bulk and surface properties on the rate of reaction. Contrary to expectation, the rate of the basic step in electrochemical reaction—electron transfer—is independent of the ease with which electrons are released from the metal. Hence, those simple electrode processes in which no other changes but electron transfer take place have heats of activation virtually independent of the metal substrate.

In a reaction involving the formation of a chemical bond between the electrode substrate and one of the radicals (charged particles) formed on the surface, rates of reaction at a given potential may vary for different substrates by many orders of magnitude. The variation is a function of the strength (or energy) of bonds established between the intermediate species and the surface, and when the substrates are transition metals, regular relations between reaction rate and certain characteristics of the metal's electronic structure are observed. If such bonds are too weak, the catalytic effect of the substrate is small. If they are too strong, the intermediate species is too stable to react further, and the surface of the metal becomes virtually blocked for further reaction. Thus, there is an optimum bond strength that can be seen if the rate of a reaction is plotted against a property of the metal substrate, which is proportional to the ability of the surface to form bonds; *e.g.*, its heat of sublimation.

Electrochemical reactions can also be catalyzed or inhibited by foreign species present in the electrolytic solution. In the most general case, ions can change the rates of reactions by changing the properties of the double layer by specific adsorption effects. Organic molecules inhibit electrochemical reaction by blocking the surface. For some organic ions or dipoles, the blocking effect can begin at a certain well-defined potential. Before attaining it, the electrode reaction gives the usual current density-potential relationship, which is then suddenly interrupted by an abnormal fall of the current, corresponding to adsorption.

Electrocrystallization. Deposition of metals and other substances at electrodes as a consequence of an electrode process exhibits a number of specific features. The electrode process is followed by crystal building, and this re-

sults in a continuous change of the electrode surface. This change, in turn, affects the electrochemical properties of the system—the double-layer capacity and the rate constants of the charge transfer processes. Hence, if electrochemical properties are to be studied, transient methods should be employed that allow measurements to be made before major changes in surface morphology (structure) take place.

Adatoms and adions. Since the two steps, the discharge of ions at the electrode and the incorporation of the discharged ions into the crystal lattice, are separated in time and space, an intermediate species exists at the surface, that of relatively loosely bound and freely moving atoms, called adatoms. Since the electrons tend to join the rest in the bulk of the metal, adatoms appear to have a partial charge, less than that of the elementary positive charge. The adatoms therefore attract solvent molecules, and the species is partially solvated. This reaction justifies adatoms being considered as a kind of adsorbed ion, called an adion, which, however, has already undergone partial discharge.

Because such an intermediate state is possible, ions need not be reduced to the neutral atomic state at the point of incorporation into the crystal lattice. Indeed, energetic considerations reveal this reduction to be improbable. Instead, discharge of ions is favoured on crystal planes. Their motion from the location of discharge to a site on the crystal occurs by surface diffusion. When such surface diffusion is inhibited, it can, in some cases, control the rate of reaction.

Problems of the crystal-building step in electrocrystallization are in many respects identical with those of crystal growth from the gas phase. Crystal building can occur in any one of three possible ways.

Kinks and step growth. Kinks, naturally existing at any metal surface, form a suitable half-lattice position at which an atom is surrounded by one-half of the number of atoms that would surround it in the bulk of the metal; there, adatoms can be successively trapped and thus the crystal lattice is extended along a crystal edge and further on across the surface. This step growth mechanism is shown in Figure 3A. The mechanism, however, has a limited capacity for crystal growth. A step can move as far as the edge of a crystal, and step growth would lead to smoothing of the surface to perfection, but then further growth would cease.

Screw dislocations. Mechanisms associated with screw dislocations, or twinning edges, can provide for a continuous growth of crystals. The screw dislocation mechanism, shown in Figure 3B, is made possible by a specific fault often found in the crystal lattice that may be called a dislocation originating from a shift of one atom in the lattice with respect to a perfect arrangement. This shift may then result in the formation of a monoatomic edge projecting above the electrode surface at which new atoms can be stacked. The stacking produces a turning of the edge around the base atom as a centre, the process producing a spiral growth.

Stacking

Nucleation mechanisms. If growth sites are rare, or if the substrate at which deposition should take place is foreign to the depositing metal, the charge transfer results in an accumulation of adatoms to a concentration considerably larger than that which can exist there at equilibrium with the crystal lattice. This condition is termed supersaturation. In such a situation, agglomeration of adatoms to form crystal nuclei is favoured. Surface energy requirements show that at any degree of supersaturation, nuclei of certain dimensions are stable and can represent sites for further growth. The process is shown in Figure 3C.

The formation of deposits is controlled by the above mechanism as long as the discharge process supplies ample amounts of adatoms as building material over the entire surface. If the rate of deposition is increased, so as to produce near the surface considerable depletion of ions, uneven deposition will start. This is caused by the protrusions of a normally rough metal surface being closer to the bulk of solution than the recessed parts and, hence, getting a somewhat faster supply of the discharging species. Once such a situation is established, it tends to develop further. The faster growing points penetrate into ever richer layers of solution, and this penetration into fertile fields in turn results in an ever faster growth. Thus, a natural consequence of deposition under transport-controlled rate is the amplification of the original surface roughness and the appearance of protruding spikes, called dendrites, as deposits.

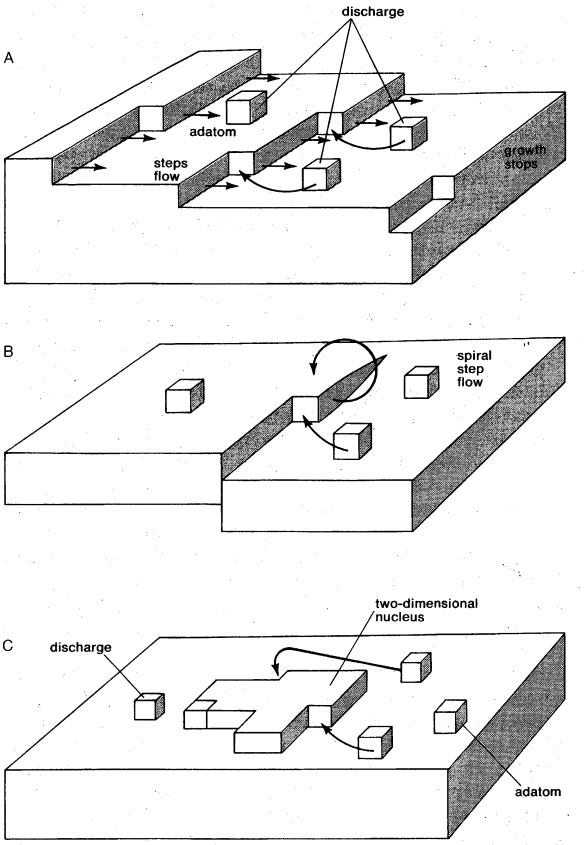


Figure 3: *Mechanisms of crystal growth.*
(A) Step growth, (B) screw dislocation, (C) nucleation.

Additives. A converse process is the smoothing of the original surface irregularities, which may occur when some foreign species, called an additive, is present in solution and adsorbs on the surface and inhibits the process of discharge. If those molecules are incorporated into a growing deposit, a situation may arise in which their supply to recessed parts of the surface becomes slower than to elevated parts. As a result, deposition becomes faster at recessed parts than at elevated ones and leveling of the surface occurs. This process has considerable technical application. Metal dissolution can sometimes be governed by a similar, transport-controlled mechanism, in which case a polishing effect is obtained. Electropolishing has found wide application in practice.

Organic electrode reactions. A very large number of electrochemical reactions involving organic molecules are known. An example is the oxidation of ethylene according to the equation:

$$\underset{\text{(ethylene)}}{C_2H_4} + \underset{\text{(water)}}{4H_2O} \rightarrow \underset{\text{(carbon dioxide)}}{2CO_2} + \underset{\text{(hydrogen ion)}}{12H^+} + \underset{\text{(electrons)}}{12e^-}.$$

Many chemical organic reactions can be made to function electrochemically, a general advantage being that the rate of the reaction is easily controlled by controlling the potential. A change of potential may change the path of the reaction and hence a certain product may be tuned in. In particular, polymerization reactions at electrodes can be stimulated; such a method is now used as a step in the production of nylon.

Electromechanical effects. Several properties of solid bulk metals turn out to be dependent upon electrochemi-

cal reactions. Thus, if the hydrogen atom concentration on an electrode surface exceeds a certain critical value, the hydrogen molecules (H_2) inside intrinsic voids in the metal's interior may exceed a critical gas pressure and cause the voids to open up and spread, thus causing the metal to lose ductility and become brittle. It is the electrochemical reaction, represented by the equation $H^+ + e^- \rightarrow H$ (hydrogen ion plus electron equals hydrogen atom) at the metal–solution interface that controls the surface hydrogen atom (H) concentration and, hence, the hydrogen molecules (H_2) present inside the metal. A limiting electrode potential which causes hydrogen atom concentration exists, above which a metal loses strength. This concept applies also to moist metal surfaces; *i.e.*, most practical metal surfaces, the electrolyte being in effect a surface moisture layer and the ions being produced from the dissolution of atmospheric carbon dioxide. Friction between solids involving moist or wet surfaces is also dependent on electrochemical factors.

Ductile and brittle properties

Multielectrode systems. So far, systems have been considered in which a single electrode process takes place. In principle, at any electrode potential all species present in the system fall into two categories: those that are stable, and those that undergo oxidation or reduction. The stable species are those that at the given electrode potential would not decrease their free energy by giving off or accepting electrons.

At any potential there should occur a codischarge of all unstable species. Thus, the system can be considered as a multielectrode system, consisting of as many electrodes as there are redox couples present. The rate at which different processes occur, however, can be so widely different that usually a single process is by far the dominant one. Systems at which two electrode processes occur at comparable rates are of considerable importance. If two kinds of metal ions are discharged simultaneously, an alloy is formed upon crystallization. The properties of the alloy would in most cases be those determined by the phase diagram (a plot of the temperature of melting versus the composition of the mixed system) for the ratio of quantities of discharged metals given by the rates at which the discharges take place. In many cases, nonequilibrium metal phases are formed giving unusual properties to the alloy.

Corrosion

If a process of metal dissolution (an oxidation process) can occur at a rate comparable to that of some reduction process on the same metal, a corrosion couple is established. Thus, zinc immersed in acid solution tends to establish a potential sufficiently positive for the metal to eject metal ions into the solution (oxidation). At the same time, it is also sufficiently negative for the reduction of hydrogen ions, present in any aqueous solution, to hydrogen gas. Hence, a spontaneous process of hydrogen evolution and of dissolution of the metal will take place. The rate at which this corrosion process occurs is governed by rate laws of the type given by equation (3) (see below *Calculations*). The mixed potential, spontaneously established by the corroding metal, is obtained by equating expressions for the anodic and cathodic currents of the two processes in the corrosion couple.

Corrosion can be prevented in two ways. One is by using an external source to make the potential of the metal sufficiently negative to bring it into the potential region in which the metal is stable, called cathodic protection; and the other is by provoking by some means the formation of a film on the surface that would slow the process. Such films could consist of an oxide or a layer of organic molecules that prevents dissolution and hence is called an inhibitor.

APPLICATIONS

Electrochemical processes are used in many ways and their use is likely to increase because they can replace polluting chemical situations with nonpolluting electrochemical ones. In many fields, however, applications have been profitable for some time. Major categories are listed below.

Metallurgy. All technologically important metals, except iron and steel, are either obtained or refined by electrochemical processes; for example, aluminum, titanium, alkaline earth, and alkali metals are obtained by electrode position from molten salts, and copper is refined by electrolysis in aqueous copper sulfate solutions.

Electroplating. One of the major ways of both decorating objects and improving their resistance to corrosion is by electroplating them. All major metal working industries, particularly the automobile industry, have large electroplating plants.

Chemical industry. Electrolysis of brine to obtain chlorine and caustic soda is an electrochemical process that has become one of the largest volume productions in the chemical industry. Modern processes cover a wide field, from the production of a variety of inorganic compounds to the production of such synthetic fibres as nylon. Intensive research in organic electrochemistry promises major developments in application, particularly with the prospect of greatly reduced electricity costs expected eventually to arise from the development of controlled fusion.

Batteries. Electrochemical storage of electricity is effected in batteries. Such devices are electrochemical cells and consist of two electrodes per unit. As the electricity to be stored is accepted on the plates of the cell, it converts substances on the plates to new substances having a higher energy than the old ones. When it is desired to make the electricity available again, the terminals of the battery are connected to the load and the substances on the battery plates retransform themselves to those originally present, giving off electricity as a product of their electrochemical reactions. The steadily rising production of the lead-acid battery is largely the result of its use for starting the internal-combustion engine, which has had an equally steady rise. Other electrochemical systems are also used as storers. The nickel-iron (Edison cell) and nickel-cadmium battery with alkaline electrolyte are both used in applications where longer lives than those of the lead-acid battery are needed; the silver-zinc battery is used to start airplane engines because of its high power per unit of weight. A variety of new systems is being investigated for covering other needs. One of the greatest challenges to electrochemists and electrochemical engineers is that of producing a battery with sufficient power and energy density to run an automobile the way gasoline (petrol) does. Even if the best hypothetical predictions for removal of polluting chemicals from automobile exhausts is realized, the cleanup will not be sufficient because the expected growth of the automobile population will increase the pollutant rate until by 1985 it will again become serious. Electrochemically-powered automobiles by that date seem to be a necessity.

Fuel cells. The energy of chemical reactions is converted into electrical energy in fuel cells. In these, the fuel (*e.g.*, hydrogen, hydrazine) is fed continuously to one electrode, while oxygen from the air is reacting at the other one. The efficiency of energy conversion in fuel cells is more than twice that attainable by conventional means, for example, by means of internal combustion.

Analytical chemistry. In analytical chemistry, most modern automated instrumental analysis is based on electrode processes; for example potentiometry, used to measure ionization constant.

Biological research. In biology the idea that many biological processes, from blood clotting to the transfer of nerve impulses, are electrochemical in nature continues to spread. The biological conversion of the chemical energy of food to mechanical energy takes place at an efficiency so high that it is difficult to explain without electrochemical mechanisms. Intensive research is developing in various directions in bioelectrochemistry.

CALCULATIONS

The three equations referred to above are stated in this section, and other mathematical considerations are also included. The rate of an electrochemical reaction in terms of oxidation and reduction reactions, the concentration of the reacting species, the electrode potentials and the current densities can all be related quantitatively according to equation (1):

$$i = \vec{i} - \overleftarrow{i} = \vec{k}C_{red}\exp\left(\frac{\alpha_a F}{RT}E\right) - \overleftarrow{k}C_{ox}\exp\left(\frac{\alpha_c F}{RT}E\right)$$

in which i is the net current density, $\vec{i}$ and $\overleftarrow{i}$ are the partial current densities of the oxidation and reduction respectively, C_{red} and C_{ox} are the concentrations of the reducing and oxidizing agents, respectively, $\vec{k}$ and $\overleftarrow{k}$ are the corresponding rate constants, while α_a and α_c are the so-called transfer coefficients; *i.e.*, specific constants giving a proper influence factor to the exponential dependence of rate on the potential, E. In the case of a simple one-electron transfer these factors are termed symmetry factors, for they, in a way, reflect the symmetry of the energy barrier. It can be proved that $\alpha_c + \alpha_a = n$, the number of electrons exchanged in a single act of an electrode reaction.

Nernst Equation. For a particular value of E the two partial current densities must become equal. This value of potential is the reversible electrode potential. From equation (1) one can deduce equation (2):

$$E_{rev} = \frac{RT}{(\alpha_c + \alpha_a)F}\ln\frac{\vec{k}}{\overleftarrow{k}} + \frac{RT}{(\alpha_c + \alpha_a)F}\ln\frac{C_{ox}}{C_{red}}$$

$$= E^\circ + \frac{RT}{nF}\ln\frac{C_{ox}}{C_{red}}.$$

This equation is known as the Nernst Equation; E° is the standard electrode potential (at $C_{ox} = C_{red} = 1$) characteristic of the given redox couple.

The standard electrode potential on the hydrogen scale is related to the thermodynamics of the electrode process. It reflects the standard free energy change of the redox reaction between the electron and the given redox couple, relative to the free energy change that takes place in the hydrogen electrode process.

Butler-Volmer Equation. The reversible electrode potential can be introduced into equation (1) and the potentials taken relative to its value. When so expressed, they are termed overpotentials and can be stated as $\eta = E - E_{rev}$; equation (1) then transforms to equation (3):

$$i = i_0\left\{\frac{(C_{red})_i}{(C_{red})_o}\exp\left(\frac{\alpha_a F\eta}{RT}\right) - \frac{(C_{ox})_i}{(C_{ox})_o}\exp\left(-\frac{\alpha_c F\eta}{RT}\right)\right\}$$

in which i_0 represents the value of either of the (equal) electron-emitting and electron-accepting partial current densities at the reversible potential and is termed the exchange current density. Equation (3) is called the Butler-Volmer Equation and represents one of the most fundamental relationships of electrochemistry.

As overpotentials, either positive or negative, become larger than about 5×10^{-2} volts (V), the second or the first term of equation (3) becomes negligible, respectively. Hence, simple exponential relationships between current (*i.e.*, rate) and overpotential are obtained, or the overpotential can be considered as logarithmically dependent on the current density. This theoretical result is in agreement with the experimental findings of the German physical chemist Julius Tafel (1905), and the usual plots of overpotential versus log current density are known as Tafel lines. The slope of a Tafel plot reveals the value of the transfer coefficient α for the given direction of the electrode reaction.

Difficulties in transport of reaction species. The above conclusions about the overpotential-current density relationship are valid as long as the ratios of concentrations at the electrode surface of the species involved at current density i, C_i and in the absence of current C_o stay close to unity. As the current density is increased, the concentration gradient needed to maintain a corresponding diffusion flux of the species concerned must begin to become appreciable. This condition is possible only if the concentration of the species at the surface starts to differ appreciably from the bulk value; *i.e.*, $(C_i)_i/(C_i)_o \neq 1$. The change in concentration of the discharging species at the electrode surface with time can, in principle, be obtained by using a second order partial differential equation (Fick's law), which, however, has explicit solutions only for a limited number of well-defined boundary conditions.

When significant concentration changes set in, no more exponential dependence of current density on potential can be obtained. It can be derived that, instead, a transition towards a limiting value takes place.

The important case is that in which the concentration of the discharging species at the electrode surface becomes equal to zero. The steady-state (*i.e.*, independent of time) current density obtained in such a case is the highest possible for the given set of conditions (diffusion limiting current density). The value of the concentration gradient in this case is directly proportional to the bulk concentration of the species involved and inversely proportional to the thickness of the diffusion layer (*i.e.*, the layer close to the electrode in which the concentration of the species differs from the species concentration in the bulk). This layer most often has a thickness fixed by hydrodynamic conditions in the solution surrounding the electrode. The definition used most often for the diffusion layer thickness is that of the German physical chemist Walther Hermann Nernst (1864–1941), according to whom this quantity is equal to the distance from the electrode at which the concentration would reach the bulk value if the concentration gradient were constant and equal to that at the electrode surface.

If a current larger than the limiting current is forced upon the electrode, the given electrode process will be able to sustain it only in the initial period in which the layer of solution close to the electrode is not completely exhausted of the discharging ions. As the concentration of ions tends toward zero, the electrode potential will change and another electrode process will start. The time at which the abrupt change of potential toward a new process takes place is termed the transition time. The relationship between transition time, current density, and concentration of the discharging species is given by Sand's equation:

$$\tau^{1/2} = \frac{\pi^{1/2}nFC_oD^{1/2}}{2i}.$$

Since τ is a well defined function of the concentration of the discharging species, the observation of the transition time can also be used as an analytical tool (chronopotentiometry).

BIBLIOGRAPHY. C.L. MANTELL, *Electrochemical Engineering*, 4th ed. (1960), the only text predominantly on the engineering aspects of the field; J.O'M. BOCKRIS and A.K.N. REDDY, *Modern Electrochemistry*, 2 vol. (1970), a textbook slanted towards readers in other fields who want to use electrochemical reactions, but is fairly easy to understand; J.O'M. BOCKRIS and D. DRAZIC, *Electrochemical Science* (1971), a very simple text for freshmen. More advanced works include: K.J. VETTER, *Elektrochemische Kinetik* (1961), the first textbook on the fundamentals and the standard work; P. DELAHAY, *Double Layer and Electrode Kinetics* (1965); and B.E. CONWAY, *Theory and Principles of Electrode Processes* (1965).

(J.O'M.B./A.De.)

Electromagnetic Radiation

Electromagnetic radiation is the propagation of energy through space by means of electric and magnetic fields that vary in time. In the absence of a material medium the velocity is a universal constant, c, the velocity of light. Electromagnetic radiation may be characterized, in accord with the wave theory, by the frequency with which it varies in time, or by its wavelength. Electromagnetic waves exhibit the properties of other forms of wave motion (*q.v.*), including reflection, refraction, diffraction, interference, and, because these waves are transverse (vibrating in a plane at right angles to the direction of propagation), polarization, as described below.

General characteristics. The orderly arrangement of radiation according to wavelength or frequency is called the electromagnetic spectrum, and is represented by the Table shown below, which charts the electromagnetic spectrum. The most familiar frequency range is light, that

narrow band of electromagnetic radiation to which the eye is sensitive, extending from violet on the high-frequency side to red on the low-frequency side, the neighbouring regions of which are consequently called ultraviolet and infrared. Infrared radiation is also known as radiant heat. Gamma radiation was first observed as emission from radioactive substances. The X in X-rays originated after the discovery of such radiation but while its nature was still unknown. Radio waves are also called hertzian waves, after the German physicist Heinrich Hertz, who in 1887 produced the first manmade electromagnetic waves and investigated their properties. Other terms have been applied outside the range of their initial meaning: the word monochromatic (literally "one colour") is applied to any radiation of one frequency, regardless of whether or not it is in the visible part of the spectrum. Microwaves developed for radar (see RADAR) during World War II are small only in comparison with radio waves; in comparison with light waves they are extremely long. The frequencies present in any wave disturbance are determined by the source of the wave, as shown in the following Table.

Essential features of electromagnetic radiation

The essential features of all electromagnetic radiation may be summarized rather briefly. The electric and magnetic fields in an electromagnetic wave act in planes that are perpendicular to each other and to the direction of propagation. According to electromagnetic theory, a changing electric field is accompanied by a magnetic field of a prescribed strength, and a changing magnetic field is accompanied by an electric field. Moreover, these relations maintain themselves as the disturbance spreads in a direction perpendicular to both fields, travelling with a velocity $c = 3 \times 10^{10}$ centimetres per second (approximately 186,000 miles per second) in free space. The energy density in space—that is, the energy per unit volume of an electromagnetic disturbance—and the flux of energy in the direction of propagation are equally divided between the electric and magnetic fields; both are proportional to the square of the field strengths. Momentum is also propagated by electromagnetic waves. Electromagnetic disturbances originate in the acceleration of electric charges, such as those in an oscillatory discharge (*i.e.*, a spark the charges of which alternate in their direction of flow) used by Hertz, and charge accelerations are produced when these disturbances are absorbed.

Radiation has corpuscular (particle-like) properties in addition to those characterizing wave motion. Radiant or electromagnetic energy is quantized in that for a given frequency ν (nu), its energy occurs as an integer times $h\nu$, in which h is another fundamental constant of nature, called Planck's constant after the German physicist Max Planck. A quantum of electromagnetic energy is called the photon. An ordinary particle in motion can be characterized by its momentum—its mass times its velocity. Likewise, the photon has momentum, but it has no rest mass; if a photon comes to rest by being absorbed, it ceases to exist and its energy is converted to some other form of energy. Because the photon energy is proportional to the frequency, the arrangement of radiation according to frequency is also known as its energy spectrum, and in the third column of the Table the photon energy is given. Its unit, the electron volt (eV), is the kinetic energy an electron would acquire in being accelerated through a potential difference of one volt. A photon of yellow light, as an example, has an energy of approximately 2 eV. Visible light and those parts of the spectrum adjacent to it are usually characterized by giving the wavelength, whereas radio and other long waves are more often specified by the frequency of the waves. The observation of extremely high-frequency radiation is made by detecting single individual photons, and its spectrum is described by giving the photon energy in electron volts or some multiple of that unit.

Wavelengths of the Electromagnetic Spectrum

	wavelength (cm)	frequency (sec^{-1})	photon energy (eV)
Cosmic ray photons	3×10^{-22}	10^{32}	4×10^{17}
	3×10^{-20}	10^{30}	4×10^{15}
	3×10^{-18}	10^{28}	4×10^{13}
	3×10^{-16}	10^{26}	4×10^{11}
	3×10^{-14}	10^{24}	4×10^{9}
	3×10^{-12}	10^{22}	4×10^{7}
Gamma rays	3×10^{-10}	10^{20}	4×10^{5}
X-rays	3×10^{-8}	10^{18}	4×10^{3}
Ultraviolet–visible light	3×10^{-6}	10^{16}	40
Infrared	3×10^{-4}	10^{14}	0.4
Microwaves	3×10^{-2}	10^{12}	4×10^{-3}
Radar	3	10^{10}	4×10^{-5}
UHF, VHF, FM shortwave, AM radio	300	10^{8}	4×10^{-7}
	3×10^{4}	10^{6}	4×10^{-9}
Longwave radio	3×10^{6}	10^{4}	4×10^{-11}
	3×10^{8}	10^{2}	4×10^{-13}

Different parts of the electromagnetic spectrum interact with matter quite differently. The Earth's atmosphere is almost transparent to waves in the visible and near infrared range and to radio waves; but it tends to absorb the rest except for extremely penetrating high-energy cosmic-ray photons. For astronomical observations, the atmosphere thus has only two "windows," for which optical and radio telescopes are designed, whereas X-ray astronomy, for example, depends on detectors in rockets and satellites that go outside the atmosphere. Glass can be made almost perfectly transparent to visible light, but it does not transmit the ultraviolet. Other dielectric (nonconducting) substances, such as pitch, are quite opaque to visible light but sufficiently transparent to microwave radar waves to be used for making lenses and prisms. The velocity of electromagnetic waves is lower in material media transparent to them than the free-space velocity by a factor called the index of refraction of the medium. The refractive index of any medium depends on the wavelength: the spectrum of colour can be produced by a glass prism because the short wavelength blue is slowed more in glass than the long wavelength red and thus is bent through a larger angle at each surface of the prism in emerging from glass into air. This phenomenon is called dispersion, and any medium in which the velocity of propagation depends on the wavelength is called a dispersive medium. The only truly nondispersive medium is a vacuum, but for some particular range of the spectrum a given medium may produce very little dispersion. A vacuum is also the only perfectly transparent medium, and all material media absorb strongly some regions of the spectrum.

Dispersive medium

The origins of radiant energy are also widely varied. The frequency range of electromagnetic radiation used for terrestrial communication is largely manmade, although the same frequencies are by no means uncommon in the universe as a whole (see RADIO SOURCES, ASTRONOMICAL). A radiating antenna is part of an electric circuit in which charge is caused to oscillate at a definite frequency, thus giving rise to electromagnetic waves of that frequency. A receiving antenna is part of a circuit tuned to, or in resonance with, the incoming wave. Except for the energy supply, the two circuits, radiating and receiving, are quite similar. Any system that emits radiation can absorb the same radiation. The wavelength characteristic of any emitter or absorber is related to its size, although often not in a simple way. The relatively long radio and television waves are radiated by large antennas, but visible light arises from oscillators of atomic dimensions. Careful analysis of the radiation from such natural emitters as atoms and molecules yields much of the detailed information known about atomic and molecular structure. Gamma rays are produced by accelerations inside atomic nuclei, whereas the most tightly bound electrons, those near the nucleus, may produce X-rays. Outer electrons of an atom give rise to ultraviolet and visible radiation, and molecular vibrations and rotations produce infrared. The energy per photon in line or band spectra corresponds exactly to certain aspects of atomic or molecular structure, and the pattern of frequencies emitted or absorbed serves as a clue to that structure.

This article is outlined as follows:

I. Classical theory of radiation
Historical development
Early ideas about light
Wave theory of light
Relation between electricity and magnetism
The electromagnetic wave concept
The quantum concept
Semiquantitative treatment of radiation
Properties of wave motion
Empirical laws of electricity and magnetism
The radiation field
The spectrum of electromagnetic radiation
II. Quantum theory of radiation
Electromagnetic radiation as particles: the photon
Radiation laws
Photoelectric effect
Compton effect
Raman effect
Wave–particle duality
The effects of electromagnetic radiation
Wavelength dependence
Absorption and scattering processes
Relation of electromagnetic radiation to quantum theories and relativity
III. Mathematical formulation of electromagnetic theory
Maxwell's equations
Transmission of radiation in free space
The wave equation and electromagnetic waves
The relation of radiation fields to sources
Waves in space bounded by conductors
The scattering of electromagnetic radiation
Electromagnetic waves in material media
Antennas

I. Classical theory of radiation

HISTORICAL DEVELOPMENT

Early work by Maxwell and Faraday

It may be said that the history of electromagnetic radiation began in 1865 with the publication of a paper by a British physicist, James Clerk Maxwell, proposing in summary that "we have strong reason to conclude that light itself (including radiant heat, and other radiations if any) is an electromagnetic disturbance in the form of waves. . . ."; or, at the earliest, when Michael Faraday of England published in 1846 the article "Thoughts on Ray Vibrations," following his discovery in 1845 of a connection between magnetism and light. (This was the Faraday effect: in the presence of a magnetic field a substance such as glass becomes optically active; *i.e.*, can rotate the plane of polarization of light.)

Early ideas about light. It is now recognized that light may be defined as visible electromagnetic radiation, and the science of light begins in antiquity. The rectilinear or straight-line propagation of light, a property of all electromagnetic radiation traversing empty space or a physically homogeneous transparent medium, was known to Euclid (a Greek mathematician) *c.* 300 BC, as was the law of reflection: at the boundary between two different media, radiation is reflected with the angle of incidence equal to the angle of reflection. Refraction, the change in direction of propagation of radiation in passing from one transparent medium to another, was recognized in antiquity, but the quantitative law of refraction was first discovered empirically in 1621 by a Dutch mathematician, Willebrord van Roijen Snell.

The law of refraction was actually published by René Descartes of France, who derived it on the basis of his theory that light is a kind of pressure transmitted through the all-pervading ether. According to this theory, light is propagated with infinite velocity, but in passing from a rarefied medium (air) to one more dense (glass or water), it behaves like a corpuscle the velocity component of which parallel to the interface is not affected, but which is speeded up in a direction normal to the boundary surface. The theory is neither correct nor logically consistent, but Descartes's derivation is based on measurable quantities, and is even scientific, compared with views current in the Middle Ages. It was widely accepted, but a contrary view was presented by a French mathematician, Pierre de Fermat, in 1657. In his celebrated principle of least time, he claimed that "Nature always acts by the shortest course." To the Greek geometer Hero of Alexandria (1st century AD) is attributed the theorem that reflected light traverses the shortest distance between two points compatible with meeting the reflecting surface. Fermat succeeded in deriving the law of refraction from the principle of shortest traversal time, but only on the assumption that the speed of light is smaller in a dense medium than in a rare one, contrary to Descartes's derivation. In the mid-17th century it was not yet known whether the passage of light takes any time at all, and Fermat's views were not widely accepted.

The finite velocity of light was first demonstrated by a Danish astronomer, Olaf Rømer, in 1675, from observations of the eclipses of Jupiter's satellites. He noticed that the period of the first satellite is apparently slowed as the Earth and Jupiter move away from each other, so that the eclipses occur later than expected when Jupiter is most remote. This difference is understandable if light requires a finite time to traverse the solar system. From the observations Rømer calculated the time required for light to travel from the Sun to Earth as 11 minutes. It is in fact more nearly eight minutes. In 1728 an English astronomer, James Bradley, announced the aberration of starlight, that apparent annual motion of the stars that is produced by the orbital motion of the Earth. From his observations Bradley computed the time for light to reach Earth from the sun as 8 minutes 12 seconds. Terrestrial measurements of the velocity of light, accurate to within a few percent, were first made in 1849 by the French physicists Armand-Hippolyte-Louis Fizeau and Jean-Bernard-Léon Foucault.

Wave theory of light. The theory that light originates as a vibratory motion and is transmitted as a wave disturbance in a medium was developed by an English scientist, Robert Hooke, and published in his *Micrographia* (1665). He introduced the idea of the wave front, but his wave front was not necessarily perpendicular to the direction of propagation in a homogeneous medium, and he thought the speed of light should be greater in an optically dense transparent substance. Hooke also discovered that light does not travel in perfectly straight lines but is found inside the geometrical shadow of an obstacle placed in its path. In this observation he had been anticipated by an Italian physicist, Francesco Maria Grimaldi, who named the phenomenon diffraction. The wave theory, however, is more justly attributed to a Dutch scientist, Christiaan Huygens, who explained reflection and refraction in terms of Huygens' principle, just as is done in modern elementary textbooks of physics. According to this principle (published 1690), each point on the wave front in the hypothetical ether is regarded as the source of a new disturbance that spreads out spherically, and the new wave front is the envelope of all the individual wavelets that originate from the old wave front.

Hooke's notion of a wave front

An important discovery concerning visible radiation was the observation of double refraction in Iceland spar (calcite), reported by a Danish physicist, Erasmus Bartholin, in 1669. A single ray of light entering the crystal emerges as two rays, the direction of an "extraordinary" ray in the crystal depending on the crystal orientation. Huygens gave a satisfactory account of this phenomenon on the basis of his principle by assuming secondary ellipsoidal wavelets in addition to the spherical wavelets of the ordinary ray. The ellipticity is produced by asymmetry of structure in the crystal. Huygens was not able, however, to explain his own significant discovery that the two rays obtained by double refraction in one crystal behave differently on entering a second crystal: if the crystals are oriented in exactly the same way (or at 180°), there is no further splitting of the rays; any other orientation produces four rays. Isaac Newton of England, considering this phenomenon in 1717, concluded from double refraction that light rays must have "sides"; *i.e.*, properties that depend on directions perpendicular to the direction of propagation of the radiation. Pressure waves, such as sound, are longitudinal and perfectly symmetrical about the direction of propagation. Much earlier, Newton had discovered the composite nature of white light, investigated the colours of thin films, and developed a theory of colour (see LIGHT). It was inconceivable to both

Huygens and Newton that light could be a transverse wave, vibrating at right angles to the direction of propagation, and the difficulties of interpreting the phenomenon of polarization led Newton to reject the wave theory of light.

Revival of the wave theory at the beginning of the 19th century came as a result of discoveries made by the physicists Thomas Young of England and Augustin-Jean Fresnel of France. Young's greatest theoretical contribution was the principle of interference, the description of a property that has come to be a definition of wave in distinction to corpuscle. Young explained his principle by analogy to sound and to surface waves on water and concluded that

> whenever two portions of the same light arrive to the eye by different routes, either exactly or very nearly in the same direction, the light becomes most intense when the difference of the routes is any multiple of a certain length, and least intense in the intermediate state of the interfering portions; and this length is different for light of different colours.

Fresnel carried both theory and experiment much further, explaining diffraction by straight edges, obstacles, and apertures by a combination of Young's principle of interference and Huygens' principle for constructing wave fronts. It may seem strange that the idea of interference did not occur to Huygens, but conceptually his waves were pulses, not periodic, and the association of wavelength with light escaped him.

Transverse vibrations

Acting on a suggestion from Young that light is a transverse wave, Fresnel developed a complete theory of light propagation based on the properties of an assumed elastic medium, the ether. He was able to explain the optics of crystals, both calcite (uniaxial) and the biaxial crystals that were discovered later to be more common. Detailed predictions of this theory were confirmed after his death in 1827, as were the Fresnel formulas, which describe the intensity and polarization of light reflected at the surface of a transparent medium.

Fresnel also investigated the effect of the Earth's motion on the propagation of light and suggested the experiment later carried out by Fizeau to measure the spread of light in a moving material medium, in this instance water. (Fresnel's suggestion of the decisive experiment to distinguish the wave theory from the particle hypothesis by comparing the velocity of light in air and water no longer seemed crucial when performed by Foucault and others in 1850, when evidence for the wave theory had become overwhelming.) It is difficult to do justice to Fresnel's creativity, both experimental and theoretical; almost all of his contributions to the theory of radiation have survived the transition from the assumption of an elastic solid ether to the propagation of electromagnetic fields in free space.

Except for historical interest the details of the elaborate ether theories that occupied many scientists in the 19th century are no longer relevant to the study of electromagnetic radiation. One general question, however, must be applied to all models of a medium designed to transmit light: is the ether affected by the motion of bodies through it? There were many attempts to answer this question experimentally, with inconsistent results. The most ingenious attempt was the experiment devised by the United States physicist Albert Abraham Michelson in 1881 and carried out with the collaboration of Edward Williams Morley, also of the United States, in 1887. Their apparatus was designed to compare the velocity of light in directions parallel and perpendicular to the Earth's motion and thus to measure the speed of the Earth through the ether. The results were negative: no difference could be detected. But by 1887 it was becoming evident that light is an electromagnetic wave, and it was necessary to include a much broader range of phenomena in any theory that might account for its transmission.

Relation between electricity and magnetism. As early as 1760 the celebrated Swiss-born mathematician Leonhard Euler suggested that the same ether that propagates light is responsible for electrical phenomena. In comparison with both mechanics and optics, however, the science of electricity was slow to develop. Magnetism was the one science that made progress in the Middle Ages, following the introduction into the West of the magnetic compass, but electromagnetism played little part in the scientific revolution of the 17th century. It was, however, the only part of physics in which very significant progress was made in the 18th century. By the end of that century the laws of electrostatics—the behaviour of charged particles at rest—were well known, and the stage was set for the development of the elaborate mathematical description first made by a French mathematician, Siméon-Denis Poisson. There was no apparent connection of electricity with magnetism, except that magnetic poles, like electric charges, attract and repel with an inverse-square law force.

Following the discoveries in electrochemistry (the chemical effects of electrical current) by the Italian investigators Luigi Galvani, a physiologist, and Alessandro Volta, a physicist, interest turned to current electricity. A search was made by a Danish physicist, Hans Christian Ørsted, for some connection between electric currents and magnetism, and during the winter 1819–20 he observed the effect of a current on a magnetic needle. Members of the French Academy learned about Ørsted's discovery in September 1820, and several of them began to investigate it further. Of these the most thorough in both experiment and theory was a French physicist, André-Marie Ampère, who may be called the father of electrodynamics. The magnetic effect of a current had been observed earlier (1802) by an Italian jurist, Gian Domenico Romagnosi, but the announcement was published in an obscure newspaper.

Connection between electric currents and magnetism

The list of four fundamental empirical laws of electricity and magnetism was made complete with the discovery of electromagnetic induction by Faraday in 1831. In brief, a change in magnetic flux through a conducting circuit produces a current in the circuit. Almost equal claim to priority could be made for a United States physicist, Joseph Henry and the significant observation that the induced current is in a direction to oppose the change that produces it, now known as Lenz's law, was formulated by a Russian physicist, Heinrich Friedrich Emil Lenz, in 1834. When the laws were put into mathematical form by Maxwell, the law of induction was generalized to include the production of electric force in space, independent of actual conducting circuits, but was otherwise unchanged. On the other hand, Ampère's law describing the magnetic effect of a current required amendment in order to be consistent with the conservation of charge (the total charge must remain constant) in the presence of changing electric fields, and Maxwell introduced the idea of "displacement current" to make the set of equations logically consistent. As a result, he found on combining the equations that he arrived at a wave equation, according to which transverse electric and magnetic disturbances were propagated with a velocity that could be calculated from electrical measurements. These measurements were available to Maxwell, having been made in 1856 by the German physicists Rudolph Hermann Arndt Kohlrausch and Wilhelm Eduard Weber, and his calculation gave him a result that was the same, within the limits of error, as the speed of light in vacuum. It was the coincidence of this value with the velocity of the waves predicted by his theory that convinced Maxwell of the electromagnetic nature of light.

The electromagnetic wave concept. Maxwell's theory predicted that the index of refraction of a transparent medium is equal to the square root of the dielectric constant (a measure of specific inductive capacity) of the medium. This relation was not consistently observed because the dielectric constant was not measured at optical frequencies, and the apparent discrepancy added doubts to the validity of the theory. Maxwell did not include any account of the origin of radiation; and although effects of oscillatory electrical discharges had been observed much earlier by Joseph Henry and others, the theory was not confirmed until after his death with the experiments (1887–88) of Hertz. Hertz produced electromagnetic waves, demonstrated their reflection, refraction, and interference in complete accord with Maxwell's results, and

also took steps toward a theory of radiation, relating waves to their sources. The concept of retarded potentials had been introduced earlier, in 1867, by a physicist, Ludvig Valentin Lorenz, of Denmark, but their relation to the waves predicted by Maxwell was not at first clear. Other predictions of Maxwell's theory, including the pressure of light, were verified in the years that followed.

Practical use of electromagnetic radiation for communication (radio, radar, television) dates from the demonstration of wireless telegraphy by a British physicist, Oliver Joseph Lodge, in 1894, but Lodge himself failed to foresee any practical implications of the experiment. The credit for devising wireless long-distance communication is properly assigned to an Italian engineer, Guglielmo Marconi who in 1901 achieved transatlantic radio communication. This long-range transmission of electromagnetic radiation had been considered impossible because of the Earth's curvature, and the scientific community was initially skeptical. In 1903 Oliver Heaviside of England, and Arthur Edwin Kennelly, an electrical engineer of the United States, postulated a region of the upper atmosphere, called by them the ionosphere, that reflects electromagnetic waves of long wavelength. The ionosphere itself at once became a subject for research.

The concepts of electric and magnetic fields, in the sense of forces filling space, began with Faraday, who visualized them as bundles of lines, and were developed mathematically by a British physicist, Lord Kelvin (William Thomson), and by Maxwell. Maxwell was fully aware of the energy density in the field, but a general theorem relating the energy flux at any point to the electromagnetic fields at that point was announced only in 1884, by an English physicist, John Henry Poynting, and independently by Heaviside. In 1859 Gustav Robert Kirchoff, a physicist of Germany, showed that the frequency distribution of heat radiation in a cavity surrounded by walls (a black body) depends only on the temperature and the frequency, or wavelength. An expression for this dependence was derived by a German physicist, Wilhelm Wein, in 1893 on the basis of thermodynamic principles, but later was found by experiment to be valid only for high frequencies. In 1900 an English physicist, Lord Rayleigh (John William Strutt), published a different formula, based on a theorem of statistical mechanics, that of equipartition of energy among the degrees of freedom of the system. This formula, known as the Rayleigh–Jeans law after Rayleigh and James Hopwood Jeans, states that the intensity of radiation emitted by a black body is proportional to the square of its frequency; but it also predicts infinite intensity for the high-frequency end of the spectrum, although it approaches the correct value for long wavelengths at the other end of the spectrum. The correct law, proposed in 1900 by Planck, initiated the quantum theory of electromagnetic radiation.

The quantum concept. Although Planck originated the quantum hypothesis and thus arrived at the perfectly satisfactory radiation law that bears his name, the quantum of radiant energy, or photon, as it was later named (1926), was the theoretical creation of Albert Einstein in 1905, postulated to describe the photoelectric effect (*q.v.*). The reality of the quantum of electromagnetic radiation, complete with momentum as well as energy, is exhibited still more clearly by the scattering of photons by free electrons (those not bound to nuclei in atoms), the Compton effect discovered in 1923 by Arthur Holly Compton, a physicist of the United States. Einstein, however, had already, in 1916, derived the fundamental equations for the exchange of energy between matter and radiation, introducing the concept of stimulated emission and showing that emission and absorption are directed processes, involving transfer of momentum as well as energy. The result was successfully applied to describe the scattering of light by an assemblage of atoms around 1924 by the physicist Hendrik Anthony Kramers of The Netherlands and Werner Heisenberg of Germany. Stimulated emission means that there will be inelastic scattering, with the production of frequencies that are linear combinations of some natural frequency of the atom with that of the incident radiation. Such radiation, called Raman emission, was discovered in 1928 by a physicist, Chandrasekhara Venkata Raman, of India.

Stimulated emissions

The quantum theory of the pure radiation field was first developed by an English physicist, Paul Adrien Maurice Dirac in 1927; and for treating the interaction of radiation with matter, the electron is represented by the Dirac wave equation. For an account of the quantum theory of radiation, the reader is referred to the final reference listed in the bibliography of this article.

SEMIQUANTITATIVE TREATMENT OF RADIATION

Properties of wave motion. In wave motion (*q.v.*), be it from sound, the undulation of the ocean's surface, or electromagnetic radiation, energy is transmitted through space or through a material medium as some sort of disturbance, without the transport of matter. The disturbance may be sudden, with a sharp onset, or have any other form of beginning. Continuous waves of any kind may be periodic, the form of the disturbance being repeated at regular intervals, and the simplest periodic waves are sinusoidal; *i.e.*, identical in space at any given instant with a sine curve (Figure 1). The frequency ([nu] ν = number of vibrations of the wave at a given point per unit time) of a periodic wave is imposed by the source of the disturbance, but the velocity (v) of the disturbance through space is in general determined by the medium and by the nature of the wave. The wavelength (lambda, λ) of a periodic wave depends on both the source and the medium; the product of the wavelength and the frequency is the wave velocity: $\nu\lambda = v$. Waves are reflected by a boundary between two different media and are refracted (changed in their direction) in passing from one transparent medium into another, where their velocity is different. Waves can be characterized by interference and diffraction, and all electromagnetic radiation exhibits these latter properties.

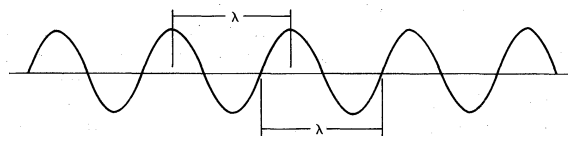


Figure 1: Sinusoidal waves in space, showing wavelength λ.

The distinction between interference and diffraction is not altogether sharp, but the term diffraction is usually applied to the deviation of waves from straight-line propagation, the occurrence of a disturbance in regions shielded from direct view of the source by opaque obstacles. The word interference, on the other hand, is most often used to describe the redistribution of energy in space or time because of the reinforcement and cancellation of parts of the disturbance. Two adjacent periodic sources of the same frequency, for example, reinforce to produce maxima in some parts of space, and cancel to produce minima, or nodes, elsewhere. Another example is provided by two sources with slightly different frequencies that interfere to produce beats, alternate reinforcements and cancellations that are periodic with time.

A surface of constant phase, such as a crest or trough of a periodic wave, is called a wave front. The wave front radiated by a single small source in three-dimensional space is a spherical surface. The energy transported in wave motion is proportional to the square of the wave amplitude (maximum periodic displacement). The energy transport per unit area perpendicular to the direction of wave propagation is the intensity, which for a spherical wave falls off inversely as the square of the distance from the source (and thus the amplitude falls off with the first power). This law follows from the fact that, in the absence of absorption, an undiminished amount of energy flows through any imaginary sphere surrounding the source, the area of a sphere being proportional to the square of its radius. Far away from a source where the radius is large, a resulting wave front will be approximately a plane, with energy transmitted undiminished in

the direction of wave propagation. Sunlight reaching the Earth is essentially plane.

Electromagnetic waves are transverse, and unlike longitudinal waves (as in the case of sound), the amplitude of vibration is perpendicular to the direction of travel; by analogy with mechanical waves, an elastic solid medium would be required to transmit them. Much effort of 19th-century physics was devoted to a study of the properties of this hypothetical medium that offers no perceptible resistance to the motion of planets and other bodies. This was called the ether. The laws of light and electromagnetism were based on the assumption of an ether stationary throughout the universe, with respect to which the laws were presumably valid in their simple form. The ether theories of the 19th century were based on Newtonian mechanics, and on this basis, if the Earth moves relative to the ether, it should be possible to detect its motion. The evidence seemed contradictory. Because of the motion of the Earth about the Sun, distant stars appear to move in orbits of about 41 seconds in angular diameter, a phenomenon that is quite consistent with a stationary ether. On the other hand, a difference between the velocity of light in the direction of the Earth's motion and at right angles to that direction, predicted on the basis of a stationary ether, was not observed in the famous Michelson-Morley experiment. Other experiments sought to distinguish between a stationary ether and an ether swept along by bodies passing through it, but none demonstrated the existence of any ether at all. The problem was resolved only in 1905 with Einstein's special theory of relativity. To trace the connection between light and electromagnetism, however, one must examine the laws of electricity and magnetism.

Distinction between interference and diffraction

Empirical laws of electricity and magnetism. There are four fundamental empirical laws of electricity and magnetism that in combination predict the existence of electromagnetic radiation. Moreover, electrical measurements yield an exact value for the velocity of propagation of this radiation. Using these four laws, the essential features of this synthesis of electromagnetism and light will now be reviewed.

The law usually written first in the list of four is Coulomb's law of electrostatics, which describes the force between two particles having an electric charge. The form of the law is the same as Newton's law of gravitation, with the substitution of charge for mass, except that there are two kinds of charge, positive and negative. Like charges repel and unlike charges attract, with a force that varies directly as the product of the charges and inversely as the square of the distance between them. From this law it is possible to define an electrostatic unit of charge as that which produces unit force on a similar charge at unit distance. The law itself can be stated as an action at a distance, just as the law of gravitation is usually stated, or in terms of an intermediate concept, the electric field. One charge, considered as the source, produces a field in the space surrounding it that can be measured as a force per unit charge by a positive test charge at any given point. The electric field is usually designated by the symbol $\boldsymbol{E}$, a vector quantity (a vector is a quantity for which direction and magnitude are implied; all vectors in the following discussions will be given in italic boldface type; magnitudes in italics only). The concept of field is essential for describing radiation but is merely a convenience in presenting the laws of electrostatics.

A second law deals with magnetism only. In terms of magnetic poles it is another inverse square law, but for fuller understanding it must be stated that isolated poles are nonexistent. Every magnetic pole is part of a dipole; *i.e.*, is accompanied by an opposite pole of equal strength. Experiments may be made with long dipoles, however, such as magnetized needles, and a unit pole is defined as one that produces unit force on a similar pole at unit distance. In terms of poles, a magnetic field may be defined at any point in space as the force per unit pole. Like the electric field, the magnetic field is a vector quantity, its direction being understood as the line along which its force acts.

A third law, known as Ampère's law, relates magnetism to electricity. An electric current, defined as moving charge, is accompanied by a magnetic field, the direction of which is in general at right angles to the motion of the charge. Consequently a current loop (such as a charge moving in a wire bent to a circular shape) is equivalent to a flat magnetic dipole, a "magnetic shell" the faces (opposite sides of the plane of the loop) of which have opposite magnetic polarity. In terms of the magnetic pole strength just defined, this equivalence between magnetic dipoles and a current loop yields a corresponding unit of current, defined as that current flowing in a loop of unit area that would be equivalent to a unit magnetic dipole. The charge conveyed past a point per second of time by this unit current is called the electromagnetic unit of charge and is defined quite independently of the static force it might exert on another charge.

Ampère's law

Analysis shows that this electromagnetic unit of charge does not even have the same dimensions (in terms of mass, length, and time) as the unit defined electrostatically: it is a constant c times larger, in which c has the dimensions of a velocity. The constant may be determined in many ways; the initial method used by Kohlrausch and Weber was to measure the discharge of a capacitor having known charge through the coil of a galvanometer that deflects magnetically according to the magnitude of the charge. The value of the constant is $c = 3 \times 10^{10}$ centimetres per second, to a good approximation. There is nothing, however, in the laws of electricity and magnetism thus far in this account that would suggest any association of the constant c with the propagation of radiation. Radiation depends on effects that vary in time and is not produced by static charges or steady currents.

The fourth fundamental law of electricity and magnetism is known as Faraday's law of electromagnetic induction. This law is the principle governing the behaviour of dynamos and generators, for its most apparent consequence is that an electric current is set up in a closed circuit by a changing magnetic field. The motion of charge in a conductor is the result of an electric field, and the law may be generalized to state that a changing magnetic field is accompanied by an electric field. The electric field so generated is at right angles to the change in magnetic field. The existence of this electric field does not require the presence of a conductor, and the relation holds in empty space, well away from all charges and currents that constitute the sources of magnetic field. (Because, by Ampère's law, magnets are equivalent to configurations of current, one need not include permanent magnets explicitly as sources of magnetic fields.)

Faraday's law concludes the list of fundamental principles of electricity and magnetism, except that the third, Ampère's law, is as an empirical relation not quite complete. According to Ampère, steady currents are accompanied by magnetic fields, but the law takes no account of the accumulation of charge that is possible with currents that vary in time. It is in fact not compatible with the conservation of charge, an axiom in accord with experience that the total charge of an isolated system is constant. This defect was remedied by Maxwell, who added to moving charge what is called the displacement current. The displacement current takes into account the accumulation of charge associated with time variations of the current, and also predicts that a changing electric field is accompanied by a magnetic field in space.

These four fundamental laws, when put into mathematical form, are called Maxwell's equations, and they will be considered explicitly in a later section. Maxwell's equations serve to trace quantitatively the electric and magnetic fields to their sources, the charges and currents. They also predict reciprocal relations between electric and magnetic fields as applied to empty space: that changes in either kind of field are accompanied by changes in the other kind of field, regardless of the specific sources of the fields. These interconnections may be expressed as circulation laws, involving the work done on a unit charge or on a unit pole. Faraday's law is readily written as a circulation law: the work done in

Maxwell's equations

taking a unit electric charge around a closed loop may be called the circulation of E. This work is equal to the rate of change of the total magnetic field through the surface bounded by the loop, the total magnetic field threading the loop. (The term threading comes from Faraday's analogy of the concept of a field to a bundle of lines, the number of lines per unit cross-sectional area representing the strength of the field.) The phrase "equal to" rather than "proportional to" implies the choice of a particular set of units. With the electric field expressed in electrostatic units and the magnetic field in electromagnetic units, a factor c must be included to ensure the correct numerical relationship: the constant c times the circulation of E referred to above is equal to the time rate of change of the magnetic field summed over the surface.

Ampère's law, as amended by Maxwell in the absence of real current sources, may be worded in an analogous way. The work done in taking a (hypothetical) unit pole around a closed loop is the circulation of the magnetic field B and is proportional to the rate of change of the total electric field threading the loop. Again, the factor of proportionality c must be included. If the choice of units is the same as above, Ampère's law states that the constant c times the circulation of B is equal to the rate of change of the total amount of electric field threading the loop. (For a uniformly changing electric field this is just the time rate of change of the electric field multiplied by the area of the loop.)

It can be shown in an elementary way, as appears in Figure 2, that certain mutually perpendicular electric and magnetic field configurations will be propagated through empty space in a direction transverse to the fields at the velocity c, as a result of these reciprocal relations. In this figure, E represents the electric field strength and B represents the magnetic field strength,

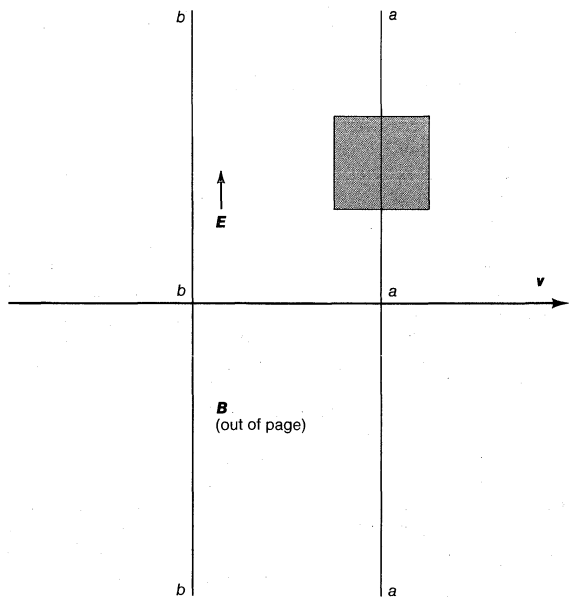


Figure 2: Slab of fields E and B, between planes *aaa* and *bbb*, moving toward right with velocity v as boundary crosses unit square (see text).

both being uniform in the region between two parallel planes *aaa* and *bbb*, with the field E directed toward the top of the page and the field B out of the page, as indicated. If this slab of field moves to the right with speed v, and the circulation laws are applied to stationary unit square loops, partly within and partly outside the field (one of which is shown as a shaded area), the circulation of E around this loop (work done on a unit charge) will be numerically equal to E, because no work would be done on a unit charge outside the field region, and the paths perpendicular to the field contribute nothing. The rate of change of total magnetic field threading this loop would be equal to the rate at which the magnetic field region is encroaching on the loop, namely v for a unit square, times the magnitude of the magnetic field, B. Thus the application of the first circulation law yields an expression in which the product of the constant c and the electric field strength is equal to the product of the velocity and the magnetic field strength, or $cE = vB$. If the square about which the circulation is to be taken is now turned at right angles to the plane of the page, and the circuital law for the field B is applied, there will result an equation analogous to the preceding; namely, $cB = vE$. Both these relations are satisfied for $v = c$ and $B = E$. The fields must have these relations to satisfy the circulation laws. Thus the orthogonal field configuration (electric and magnetic fields perpendicular to each other) will move at constant speed $v = c$ and will maintain itself undistorted; and a disturbance transferred in this way will constitute a pure wave. The conclusion is a consequence of the circuital laws that express the relationships between all magnetic and electric fields varying with time.

This result is entirely equivalent to the more formal mathematical analysis by which the equation for plane electromagnetic waves is derived in the section *Transmission of radiation in free space* (see below). The shape of the wave is arbitrary, and for many purposes it is more convenient to consider a sinusoidal disturbance (see Figure 3) of definite frequency and free-space wave-

Sinusoidal disturbance

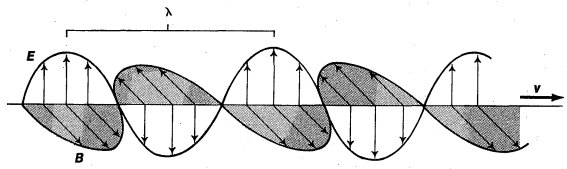


Figure 3: Radiation fields in which vectors E and B are perpendicular to each other and to the direction of propagation (see text).

length than a wave profile with such discontinuities as the one above. Here the electric and magnetic fields are transverse to the direction of propagation, and there is a definite relation between their magnitudes. The field vectors, E, B, and the direction of propagation, taken in the order stated, form a right-handed set of coordinates: if the index finger, second finger, and thumb of the right hand are extended at right angles to each other, they will represent, respectively E, B, and the direction of wave propagation. The reversal of both vectors E and B simultaneously does not change the direction of propagation, but the reversal of either E or B alone corresponds to changing the direction of the wave.

The radiation field. Within radiation fields, as in electric and magnetic fields generally, there is energy, with the energy density proportional to the square of the field strengths in question (E^2 or B^2). The relation between the fields in a progressive wave is such that the energy density is divided equally between the electric and magnetic fields at any point. The flux of energy, the energy crossing unit surface perpendicular to its flow per unit time, is called the Poynting vector and is equal in magnitude to the velocity c times the sum of electric and magnetic field energy densities. The radiation also carries momentum; the density of momentum is equal to the Poynting vector divided by c^2, or the energy density divided by the constant c.

Maxwell predicted the field relations, but he did not trace his predicted radiation to its sources. It is evident that neither static (stationary) charges nor steady currents can produce radiation fields, for the fields associated with such sources diminish rapidly in intensity as the distance from the source increases. Because electromagnetic energy density is proportional to the square of the field strength (electric or magnetic), the radiation field must be proportional to the inverse first power of the distance; for the wave intensity must obey the inverse square law to be actually transported away from the source. (This is because the area of a sphere is proportional to the square of its radius, and equal amounts of energy must be transported through successive spherical surfaces.) Such fields are produced by accelerated charges.

The simplest radiator is an oscillating electric dipole,

two equal and opposite charges vibrating in a straight line about a central point at the same rate but opposite in phase (that is, they are alternately moving toward and away from each other, passing at the central point). The dipole moment is the magnitude of one charge multiplied by its separation from the opposite charge at any instant, and in the oscillation the moment varies from its maximum value in one direction to its maximum value in the opposite direction, because, by convention, the direction of the field is from the positive charge to the negative charge, and the two charges are constantly interchanging positions. A dipole antenna may be represented in this way. The radiation field of an oscillating dipole is polarized, with the electric field vector at any point lying in a plane that contains the axis of the dipole itself; both the direction of propagation and the magnetic field vector are at right angles to this plane; because there are an infinite number of planes passing through the dipole, the magnetic field directions or lines are circles about the axis of the dipole. The radiation intensity depends on the angle between the dipole axis and the direction of propagation, having a maximum in the equatorial plane and a node (minimum) in the direction of the axis of the dipole. If the vibration frequency is fixed, the radiation is of course one wavelength, or monochromatic; but the radiation from accelerated charges generally can be analyzed into a spectrum of dipole radiation, and the strongest radiation from atoms and molecules is ordinarily of this kind.

Dipole radiation

Whereas a purely monochromatic wave is an idealization, any real wave may be represented by a superposition of disturbances having definite frequencies. The oscillating electric dipole is, strictly speaking, also an idealization of actual charge and current distributions, even though the dipole radiation pattern is often a good representation of actual radiation. A more complete description requires that the radiation field be expressed in terms of a series expansion, of which the dominant term is the electric dipole. Only when the configuration of charges is such that the dipole moment vanishes is the radiation not predominantly electric dipolar in character.

THE SPECTRUM OF ELECTROMAGNETIC RADIATION

All electromagnetic waves transport energy from a source to a receiver, as do mechanical waves, but unlike mechanical waves they do not require a material medium.

Electric current waves. Electric current waves, nevertheless, are commonly propagated along transmission lines, such as a pair of parallel wires or a coaxial cable. Transmission lines are in fact necessary to convey high-frequency electromagnetic energy from a generator to a load such as an antenna. The conductors are designed to have as little resistance as possible, and many features of typical transmission lines for high frequencies are those of an ideal line, for which the conductivity is infinite.

Figure 4 shows a cross-sectional view of a pair of long parallel wires fed at one end by a high-frequency alternating current generator. At the position of the cross section there is a difference of potential between the wires associated with their opposite charges, and therefore electric field lines extend from one wire to the other, as indicated. There is also a magnetic field between them associated with their opposite currents, magnetic field lines that crowd between the two wires, shown as dashed lines in the figure. These two fields, electric and magnetic, are perpendicular to each other and to the wires. It is possible to compute the capacitance per unit length of line (the ratio of the linear charge to the potential difference) in terms of the radius of each wire and their distance apart, and also to compute the inductance per unit length (the ratio of the inclosed flux per unit length to the current) in terms of the same dimensions. The application of simple circuit theory leads to a differential equation in the charge per unit length that is independent of the dimensions of the system and the frequency, a wave equation that predicts the charge disturbance will be propagated down the line with a velocity equal to the

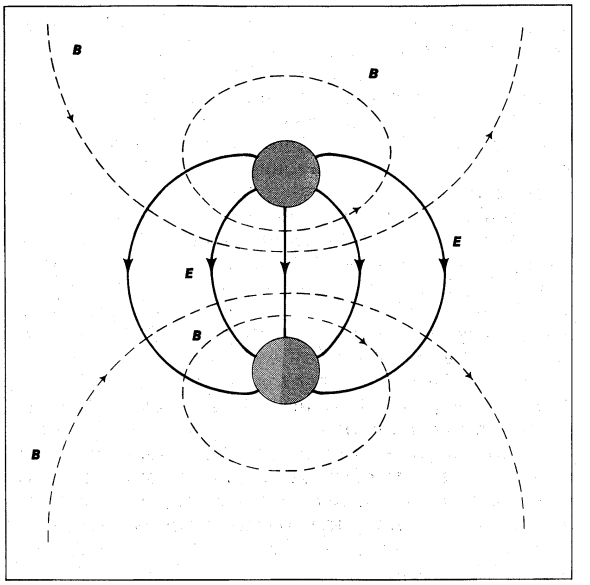


Figure 4: Cross section of a pair of wires (indicated by shading) carrying high-frequency current. Field lines ***E*** and ***B*** and wires are mutually perpendicular (see text).

velocity of light. (For the configuration shown in Figure 4 the wave is transmitted out of the page toward the reader.) Moreover, the distribution of the transverse electric and magnetic fields is exactly the same as if the charges were static and the currents were steady. The actual configuration of fields depends on the geometry of the conductors, but the identity of the fields with those produced under quasi-static conditions and the speed of transmission of the disturbance are quite general. For actual conductors having finite conductivity there must be a component of the electric field along the transmission line, but this component is small for good conductors.

Radio waves. Radio waves are ordinarily transmitted through space without benefit of wires, but they are produced and detected by electrical circuits. These disturbances have relatively low frequencies and consequently long wavelengths, as indicated in the Table. In empty space or a homogeneous medium, they would travel in straight lines and, because of the Earth's curvature, could not be useful in direct communication over long distances. It was this consideration that caused scientists to question Marconi's efforts to develop long-range radio, and Marconi's success, unexpected on theoretical grounds, led to the postulation of the Kennelly-Heaviside layer, or ionosphere. This region of free ions (positively charged particles) and electrons actually consists of more than one layer and fluctuates widely from day to night and from summer to winter.

The discovery of the ionosphere

Despite its fluctuation, the ionosphere has nevertheless been of great practical importance in radio communication. The range of direct communication, by means of what is called the ground wave, varies with wavelength but is something like 100 miles (160 kilometres) for waves 100 metres (330 feet) long, and little more for longer waves. Long-distance transmission depends on sky waves, waves that have been refracted by the ionized layers and are thus returned to Earth. Variations in signal strength are produced by fluctuations in the ionosphere and by interference of the ground and sky waves in regions where both may be received. The ionosphere absorbs appreciably only those waves whose free space wavelength is in the neighbourhood of 200 metres (660 feet). It is this frequency, about 1.5 megacycles, that corresponds to the cyclotron frequency of electrons caused by the Earth's magnetic field at that altitude. In general, the atmosphere is fairly transparent to radio frequencies, giving rise to the possibility of radio astronomy observations with terrestrial telescopes.

Microwaves. Microwaves may be defined as those with wavelengths between 1 millimetre and 30 centime-

tres, corresponding to frequencies between 3×10^{11} and 10^9 hertz (cycles per second). These wavelengths are comparable to the dimensions for which it is feasible to build apparatus, and highly directed beams may be produced by aperture antennas many wavelengths wide. This situation makes possible radar, the application of microwaves analogous to sounding in water navigation. A powerful beam of radiation may be sent out, much as a searchlight sends light, and detection of the reflected radiation gives information about the reflecting object. Because of their short length, microwaves give much better resolution and precision of information than could be obtained with longer wavelength radiation. These waves penetrate clouds of smoke, but they can detect rain because they reflect off water droplets, and so they are widely used in weather forecasting and the mapping of meteorological disturbances. Radar is particularly useful in the navigation of aircraft. Microwave apparatus is also a powerful research tool in the investigation of the properties of matter. Special devices for producing these high frequencies involve resonant cavities, and the energy is transmitted from source to transmitter by wave guides, or coaxial lines consisting of a central metal rod in a hollow conducting pipe.

Infrared waves. Infrared radiation comprises wavelengths from 1 millimetre to the longest wavelength visible light, about 7.5×10^{-5} centimetre. This range, also known as radiant heat, was discovered by the English astronomer Sir William Herschel in 1800. Herschel found, by putting a thermometer at various points in a prismatic spectrum, that ordinary light transmits some heat but that the effect is even more marked beyond the red end of the spectrum. Infrared radiation is readily absorbed by many kinds of matter and is thus effective in warming the substances on which it falls. Because it is more penetrating than visible light, such radiation lends itself to many kinds of practical applications, from physical therapy to detectors of military importance. The energy radiated by hot solids lies largely in this range. Infrared radiation is also characteristic of many natural frequencies of molecules, and by using a spectrograph and suitable photographic plates it is possible to photograph the lines and bands emitted by molecules. The atmosphere is almost completely transparent to a part of the infrared spectrum, in the neighbourhood of 10^{-3}-centimetre wavelength, and infrared photography takes advantage of this phenomenon to overcome the absorption of visible light by atmospheric haze.

Visible light. Visible light is the most familiar part of electromagnetic radiation and may be defined as that portion of the spectrum to which the eye is sensitive. This span is relatively narrow because the frequencies corresponding to violet are only about twice those that are characteristic of red. The observed wavelengths range from about 7×10^{-5} centimetre to 4×10^{-5} centimetre. The theory of light began in antiquity as a theory of vision, but a theory of vision involves many aspects of both physical and biological science. The physical science of light depends largely on instrumental receivers, such as the photographic plate. Visible radiation is produced by luminescence (fluorescence and phosphorescence) and by hot solids; its energy range includes many of the lines and bands characteristic of atoms and molecules, and historically much of our information concerning atomic structure has been derived from an analysis of atomic spectra.

White light includes a continuum of the range of visible wavelengths, which may be described mathematically as a distribution of simple harmonic waves of definite frequency, but it is in practice impossible to produce an infinitely long wave train and it is necessary to consider groups of waves. The phenomenon of interference, so crucial in determining the wave nature of light, can be observed only when the waves are mutually coherent; *i.e.*, when the interfering beams have a definite relation in phase and do not fluctuate independently of each other. If the two beams originate in the same source, they will be coherent unless the difference between their paths from the source to the interference region is too long. (It is of course possible to have partial coherence.) These arguments apply to all types of electromagnetic radiation, but for long waves the phase is more accessible to control than is generally true for optical frequencies, which are of the order of 10^{14} hertz.

Coherent beams

Ultraviolet light. Ultraviolet radiation lies on the short-wavelength side of the visible spectrum. It was reported in 1801 that the spectrum beyond the violet was capable of blackening silver chloride even more effectively than light. About the middle of the 19th century it was found that quartz is much more transparent than glass to ultraviolet rays, and ultraviolet spectroscopy developed rapidly. The detector may be a photographic plate or a fluorescent screen. In fluorescence, invisible radiation is absorbed, and radiation in the visible range is emitted. The maximum sensitivity of photographic emulsions lies in the ultraviolet, but the gelatin absorbs wavelengths shorter than about 2×10^{-5} centimetre, and the far ultraviolet was not at first accessible to investigation. The Earth's atmosphere also absorbs the short-wavelength ultraviolet and thus screens life from this kind of solar radiation.

X-rays. The discovery of X-rays by Wilhelm Conrad Röntgen, a German physicist, late in 1895 is often said to mark the beginning of modern physics. These rays were discovered by virtue of the fluorescence they produce, but their most significant feature was their penetrating power. Many kinds of substances that are quite opaque to visible light and ultraviolet light are transparent to X-radiation. In the flurry of research activity following this discovery, radioactivity was discovered and important new fields of science were opened; but the search for interference effects in X-rays was unsuccessful for nearly 17 years. In 1912 a German physicist, Max Theodor Felix von Laue, suggested that the regular arrangement of atoms in crystals might be equivalent to a ruled grating used to separate the wavelengths of visible light and might be of about the right spacing to make X-ray interference patterns. This prediction was confirmed, and X-ray crystallography was almost immediately developed further by the English physicists William Henry Bragg and his son, William Lawrence, and others. Since that time X-rays and crystal structure have remained closely related: crystal spectrometers are used to analyze X-ray spectra, and X-rays are used to determine the detailed structure of crystals. The definition of the range of radiation called X-rays is arbitrary, but the wavelengths are roughly from 10^{-7} to 10^{-11} centimetre. X-rays originate most commonly from the stopping or slowing of high-energy electrons and are usually detected by photographic plates. In the collisions of energetic electrons with heavy atoms, X-ray line spectra are also produced, definite wavelengths that are characteristic of the atoms, just as are the line spectra in the visible region.

Gamma rays. In addition to X-rays, there are other high-frequency electromagnetic radiations called gamma rays. The name originated in connection with radioactivity, gamma rays being that component of radioactive emission that is not affected by a magnetic field, in distinction to alpha- and beta-ray particles. Gamma rays are more penetrating than the particulate emanations from radioactive substances and more penetrating than longer wavelength electromagnetic radiation. Gamma rays are usually characterized by giving the energy per photon, $h\nu$, in electron volts (eV) or in millions of electron volts (MeV). They interact with matter in three principal ways: the photoelectric effect, the Compton effect (both discussed below), and in their transformation into electron–positron pairs (positrons are positively charged electrons). In order to create an electron pair, the gamma-ray photon must have an energy greater than twice the rest mass energy of an electron (the energy liberated if the combined mass of an electron and positron at rest is converted to energy); *i.e.*, more than 1.022 million electron volts. There is in principle no upper limit to gamma-ray frequencies, and all extremely high-frequency electromagnetic radiation, including the X-rays originating in space, is now classed as gamma radiation. The particle

Gamma radiation

(photon) properties of gamma rays are more pronounced than their wave properties, but their essential character is the same as that of other parts of the electromagnetic spectrum.

The mathematical formulation of the laws of electromagnetism is the set of four Maxwell equations, which may be written in a variety of ways. To appreciate the significance of Maxwell's equations and to understand the theory of electromagnetic radiation, however, simple algebraic equations will not suffice, and therefore vector analysis, a topic of advanced calculus, will be employed throughout the latter part of this article, *III. Mathematical formulation of electromagnetic theory*.

II. Quantum theory of radiation

ELECTROMAGNETIC RADIATION AS PARTICLES: THE PHOTON

Within a decade of the experimental confirmation of Maxwell's theory that light and electromagnetic waves are the same (Hertz, 1887–88), it became evident that the theory of radiation was unsatisfactory. The evidence came from several quite different phenomena, including the optical line spectra characteristic of atomic sources, which the classical theory could not explain. A more flagrant contradiction between experiment and apparently well-reasoned theoretical prediction arose in connection with the radiation from heated solids. It was this contradiction that led to the introduction of the quantum, a concept that has pervaded 20th-century physics.

Radiation laws. All solid substances emit and absorb thermal radiation, and for every particular wavelength each substance has the same radiant and absorbing power. (This relation between emission and absorption also holds for antennas, and its application to gaseous substances has made possible the study of the chemistry of interstellar matter.) Ordinarily the efficiency of absorption depends on the magnitude of the wavelength, but a perfect absorber, or black body, may be defined as one that completely absorbs all wavelengths. A cavity with perfectly absorbing walls, all at the same temperature, will contain radiant energy in thermal equilibrium with the walls, having a distribution of energy as a function of wavelength that depends on the temperature. This spectral distribution may be determined by analyzing the radiation that escapes from the cavity through a small hole. A satisfactory theory of radiation, coupled with the principles of thermodynamics, should yield a correct description of the result; but in fact two different laws were proposed in the late 19th century, neither of which was in complete accord with experiment. Both laws were valid asymptotically, that is, at the extremities of the black body spectrum, one for short waves and low temperatures, the other for long waves but not at all applicable to high frequencies. The correct law for the entire spectrum was supplied by Planck in 1900.

To trace the development of Planck's radiation law it is necessary to review briefly the basis for the incorrect laws that it supplanted. Much of this basis is quite sound and satisfies the detailed predictions that seemed to follow unequivocally from established principles. First one should note that the total energy density of radiation in a cavity is proportional to the fourth power of the absolute temperature; *i.e.*, to T^4. This result was found experimentally by Josef Stefan (1879), and derived theoretically by Ludwig Boltzmann (1884) from considerations of radiation pressure and the second law of thermodynamics; it is called the Stefan-Boltzmann law. Further thermodynamic arguments led to the conclusion by Wien (1893) that the distribution of radiant energy (dW) according to wavelength change ($d\lambda$) must have the form:

$$dW = W_\lambda d\lambda = C_1\lambda^{-5}f(\lambda T)d\lambda = C_1T^5\frac{f(\lambda T)}{(\lambda T)^5}d\lambda,$$

in which $f(\lambda T)$ is a function of the product of wavelength and absolute temperature (λT) only, and C_1 is a constant. The function W_λ must have a maximum, and the equation as to the nature of the distribution leads to a displacement law: the wavelength at which the energy distribution is a maximum (λ_m) multiplied by the absolute temperature (T) is a universal constant found to have the value:

$$\lambda_m T = 0.2898 \text{ centimetre-degree Kelvin.}$$

Wien's law

This law, like the Stefan-Boltzmann law proposed earlier, is fully confirmed by experiment. But the function $f(\lambda T)$ is not so susceptible to general arguments. Wien was able to derive a law (from consideration of a gas in equilibrium with thermal radiation), according to which the energy distribution falls off exponentially for high frequencies. This decrease agrees with observation for the high-frequency range of thermal radiation, but the formula is quite wrong for low frequencies. A more substantial argument, used by Rayleigh, begins by finding the number of modes of electromagnetic radiation (number of standing waves) per unit frequency range in a resonant cavity. In accord with the equipartition of energy theorem, which asserts that thermal energy is on the average shared equally by all possible modes of motion in an amount proportional to the absolute temperature, he assumed that each radiation mode could share equally in the total radiant energy. Each mode is thus equivalent to an elementary oscillator. The number of modes per unit frequency range is proportional to the square of the frequency (ν), and because $\nu = c/\lambda$, $\nu^2 d\nu$ is proportional in magnitude to $\lambda^{-4}d\lambda$, and the distribution should be given by $dW = CT\lambda^{-4}d\lambda$, in which C is a constant. It should be noted that this expression is consistent with the general conclusion of Wien with the function $f(\lambda T)$ simply equal to λT, but it becomes infinite in the limit of short wavelengths, and would predict that all of any finite amount of radiation should run to high frequencies.

The formula derived by Planck in 1900 would deny the equipartition of energy and substitute the assumption that the average energy of each oscillator of frequency ν is some integer times $h\nu$, and that the smallest amount of energy that can be emitted or absorbed is $h\nu$, in which h is a constant to be determined from experiment. The average energy per oscillator in an assembly of oscillators of this kind is $h\nu/(e^{h\nu/kT} - 1)$, and the radiation law becomes:

$$dW = \frac{8\pi ch\lambda^{-5}d\lambda}{e^{hc/\lambda kT} - 1},$$

which is in excellent agreement with the experimental results obtained earlier in 1900. (The constant k, called Boltzmann's constant, is well known in statistical mechanics and like the constant h is determined by experiment.)

The quantum hypothesis, which introduced Planck's constant h, yielded a solution to the paradox of black body radiation, but it was not at first applied to electromagnetic radiation in general. The light quantum was introduced as a theoretical concept by Einstein in 1905.

Photoelectric effect. Its first experimental confirmation was the evidence of its role in the photoelectric effect (*q.v.*). In this effect, discovered by Hertz, light is transformed into mechanical energy: electrons are given off when light of sufficiently high frequency strikes a metal surface, and the maximum kinetic energy of the emitted electrons is the difference between the photon energy ($h\nu$) and a work function (ϕ) that is characteristic of the metal, $E = h\nu - \phi$. This is Einstein's photoelectric equation; it means that light is absorbed in separate individual quanta, each of which may give its entire energy $h\nu$ to a single electron that must spend an energy ϕ to escape from within a metal where it is bound. There is thus a threshold frequency, $\nu_0 = \phi/h$, for the photoelectric effect. At frequencies higher than this threshold, the number of escaping electrons is proportional to the intensity of the incident radiation; but the energy per electron is a linear function of the frequency quite independent of the intensity, in accord with Einstein's equation. All features of the theory have subsequently been verified. The photoelectric effect cannot be explained by the classical theory of the absorption of electromagnetic radiation, according to which the energy necessary for electron emission would have to be accumulated gradually for low-intensity radiation, whereas no time delay is

Einstein's photoelectric equation

observed. The reality of the quantum of energy $h\nu$ and the equivalence of mass and energy leads to recognition of the fact that each quantum has momentum $h\nu/c$, or h/λ. This relation between energy and momentum is the same as in the classical theory of electromagnetic waves.

Compton effect. Compelling evidence of the particle nature of electromagnetic radiation is furnished by the Compton effect, which was announced in 1922. Compton found that in scattered X-rays there are lines of slightly lower frequency than that of the incident ray. The elementary theory of the effect is obtained by requiring both the conservation of momentum and energy in an elastic collision between a quantum of energy $h\nu$, having momentum $h\nu/c$, and an electron initially at rest. Figure 5 is

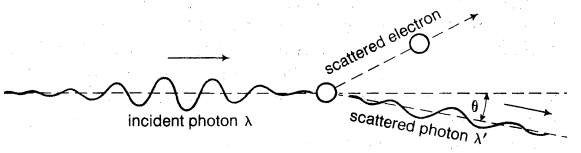


Figure 5: Elastic collision between photon and electron. Photon is scattered through angle θ with reduced frequency (ν) by an electron initially at rest (see text).

a diagram of the collision. It is necessary to use the relativistic formulas for the energy and momentum of the electron, which in the usual experimental arrangement is not observed. The result takes a simpler form in terms of wavelength rather than frequency. In terms of the scattering angle of the photon, the change produced in its wavelength is given by the product of Planck's constant divided by the mass of the electron and the velocity of the incident photon, and a quantity equal to 1 minus the cosine of the scattering angle:

$$\lambda' - \lambda = \frac{h}{mc}(1 - \cos\theta).$$

Here m is the mass of the electron, λ and λ' are the wavelengths of the photon before and after scattering, and θ (theta), the scattering angle, is the angle between the new path of the photon and the path the photon would have taken had it not been disturbed. The constant factor h/mc is called the Compton wavelength and has the value 0.0243×10^{-8} centimetre for electrons. A quantum of radiation with this wavelength would have an energy equal to the rest-mass energy of an electron.

Raman effect. The Compton effect is an example of an elastic collision between the photon and electron, but a photon may also be scattered inelastically, as in the Raman effect. This phenomenon was predicted by an Austrian physicist, Adolf Smekal (1923), and Kramers (1924), and observed in 1928 by Raman. (Equivalent observations were made almost simultaneously by G.S. Landsberg and L.I. Mandelstam in the Soviet Union.) The effect is that scattered light contains, in addition to the incident frequency ν, a series of other frequencies, all related to the original by the formula $\nu \pm \nu'$, in which ν' is a frequency observed in the absorption spectrum of the scattering material, or is equal to the difference of absorption frequencies. The Raman frequencies ν' are usually in the infrared, smaller than that of the incident visible light, so that the effect is the production of a band of lines in the neighbourhood of the incident frequency. If a scatterer is in its ground state (lowest energy state of the atoms or molecule), the scattered lines are all on the low-frequency side; but molecules already excited to higher states can add to the incident frequency, yielding visible light, although the intensity is in general stronger on the infrared side. The Raman effect has been a powerful tool in the study of molecular structure, because it brings evidence of infrared energy differences into the more easily observable visible region.

Wave–particle duality. These effects and others show that electromagnetic radiation is quantized; *i.e.*, consists of photons that behave as particles with energy $h\nu$ and momentum $h\nu/c$. Frequency, however, is a wave concept, having nothing to do with classical corpuscles. None of the phenomena of interference and diffraction that characterize a wave are invalidated by the particle aspect of radiation. Classical radiation theory, as surveyed here, is not actually incorrect, and it is quite sufficient to describe all effects except those that depend on the finite size of Planck's constant h. For long waves, and macroscopic radiators and receivers, quantum effects can ordinarily be entirely neglected, because the observed energy consists of a very large number of quanta. For all wavelengths the interference and diffraction patterns predicted by the classical theory are in accord with observation if the total effect of many quanta is observed. Thus electromagnetic radiation has a dual nature, appearing both as waves and as particles. The distribution of photons in a diffraction photograph, for example, is proportional to the intensity predicted by the wave theory. The incidence of each photon on the photographic plate is well localized; but in the neighbourhood of the diffracting obstacle, the radiation behaves like a continuous wave, impossible to localize within a wavelength at best. It should be noted that whereas a truly monochromatic wave would extend throughout all space, a narrow continuous band of frequencies may reinforce in a space with dimensions the size of a wavelength and cancel elsewhere. The lateral definition of a wave also requires a range of frequencies and phases. A wave packet with a slight spread of frequencies may exist in a definite region of space at a given time, and thus resembles a particle.

This wave–particle duality was extended by a French physicist, Louis-Victor de Broglie (1924), to apply to matter as well as to radiation. For a photon, its momentum (p) was established above, $p = h/\lambda$ (that is, its wavelength is $\lambda = h/p$); de Broglie's hypothesis is that every material particle is also a wave with a wavelength given by the same relation, which can be written $\lambda = h/mv$ inasmuch as a particle's momentum is equal to mass (m) times velocity (v). These waves of wavelength λ are not electromagnetic but are of a kind peculiar to matter. They do have periodicity, as do light waves, and an immediate consequence would be the possibility of diffraction for material particles. The diffraction of electrons was observed in 1927 by Clinton Joseph Davisson and Lester Germer, two physicists in the United States, and independently by George Paget Thomson, a physicist in Britain. The de Broglie hypothesis applies to all particles, and diffraction effects with helium atoms were found as early as 1930. Neutron diffraction is now an important tool for the investigation of crystal structure, supplementing the use of X-rays. An almost immediate theoretical consequence of the de Broglie hypothesis was the development of wave mechanics, one of two alternate formulations of quantum mechanics (*q.v.*).

Diffraction of material particles

In classical (Newtonian) mechanics the position and momentum of a particle may be known to arbitrarily high accuracy. This situation is not true for photons, and the de Broglie relation extends the limitation of precision to all particles. The uncertainty that is inherent in a wave packet makes possible an explanation of interference and diffraction in terms of photons, although the phenomena are conceptually more simple in terms of waves. Figure 6 shows an opaque shield with two openings A and B, called a double slit, and a white card acting as a projection screen. Parallel monochromatic light passing through the double slit will be found to produce on the screen an interference pattern, the intensity of which can be plotted as shown to the right. According to the wave theory, the field amplitudes at the screen cancel at any point, such as x_0, for which the respective distances to the slits A and B differ by half an odd number of wavelengths (*e.g.*, $\frac{1}{2}\lambda$, $\frac{3}{2}\lambda$, $\frac{5}{2}\lambda$, etc.).

The pattern owes its existence to the passage of part of the incident wave front through both slits. If the parallel homogeneous light is thought of in terms of photons, the description must be different, for each photon must go through either one slit or the other. Yet obstruction of one of the slits destroys the interference pattern—both slits must be open all the time. It was first thought that photons passing through the two slits interact with each other to interfere, but the pattern is produced even when

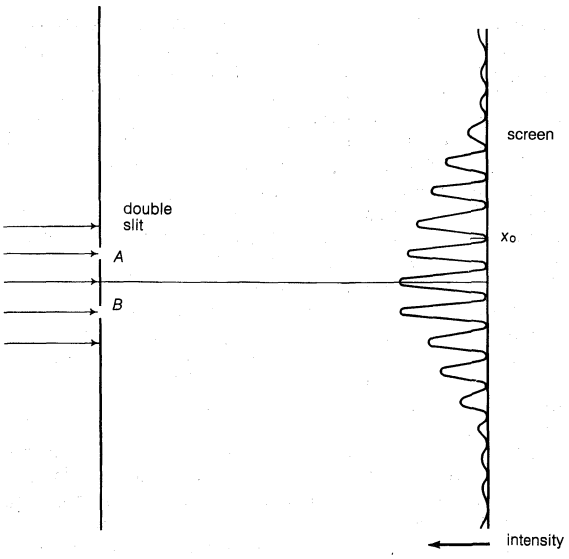


Figure 6: Light incident from the left on two equal slits forming a pattern on a distant screen. Intensity is zero at any point (*e.g.*, x_0) at which the contributions from *A* and *B* are out of phase (see text).

the intensity of the source is so low that not more than one photon is likely to be in the apparatus at any one time. The photon description of interference patterns can be satisfied in terms of probabilities. Probability amplitudes interfere in the same way that other wave amplitudes interfere, and the statistical occurrence of separate photon reception at the double slit yields the same intensity pattern as the classical wave theory. Any experiment that localizes the photon sufficiently to make it pass through a given slit destroys the pattern, just as closing the other slit would produce an alteration in electromagnetic wave intensity at the screen.

In atomic and nuclear systems, radiation occurs in the transition from one quantum state to another. A particle is described by a wave function, the mathematical representation of its equivalent matter wave. Quantized states are characterized by their parity, which is related to the symmetry of their wave functions in space. The parity transformation is inversion (reflection) in the coordinate origin (chosen as the centre of symmetry), defined by replacing the radius vector r with $-r$ whenever it appears. Parity is defined as odd, or $-$, for any function that changes sign under this transformation; and as even, or $+$, for one that does not change sign. Considerations of parity give rise to selection rules for the emission and absorption of electromagnetic radiation. In radiative transitions from one state to another, parity is conserved, so that in transitions between states of different parity the radiation field must be odd, and in transitions between states of the same parity the radiation field must be even. The parity of a particular kind of multipole field may be determined by examining the symmetry of the source, but it turns out that it is the same as the parity of the magnetic radiation field vector: electric multipole fields of order 2^n (in which n is a positive integer) have odd parity if n is odd (electric dipole, octupole) and have even parity if n is even (electric quadrupole). Because the corresponding magnetic multipoles are like the electric multipoles except that the fields are interchanged, it is evident that the opposite rule holds for them: magnetic multipole fields in order 2^n have even parity if n is odd (magnetic dipole, octupole) and have odd parity if n is even.

Selection rules

THE EFFECTS OF ELECTROMAGNETIC RADIATION

Wavelength dependence. All electromagnetic radiation interacts with matter. A beam of radiation is attenuated in passing through matter, so that its intensity decreases as its depth of penetration increases. The decrease in intensity (ΔI) in traversing a small increment of thickness (Δx) is proportional to the intensity itself (I) and to the increment of thickness, that is, $\Delta I = -\alpha I \Delta x$, and if the beam is monochromatic so that α (alpha) is constant, the intensity of the initial beam (I) after a finite penetration x is given by an exponential function $I = I_0 e^{-\alpha x}$. The attenuation may be due to elastic scattering (as discussed in the section of this article dealing with the scattering of radiation), various kinds of inelastic scattering in which the wavelength is altered, or outright absorption. Absorption is strongest at those frequencies that correspond to the resonance frequencies of the matter itself, those that are called ω_0 (omega) in the simple theory of scattering to be presented below. In fact, radiation propagates easily only through materials the atoms and molecules of which have natural frequencies that are quite different from that of the radiation. This resonance absorption may occur at any frequency; for example, the photon energies of visible and ultraviolet light, a few electron volts per photon, are comparable to those involved in chemical reactions, and these photon radiations are responsible for many photochemical reactions. Infrared radiation, or radiant heat, also produces chemical changes, such as burns and the processes that take place in cooking. Electronic ovens make use of longer wavelengths of the high-frequency radio band, usually the microwave range, so that the cooking of food is accomplished evenly by radiation that leaves almost unheated any nonconducting container for the food. (Microwaves do not penetrate metallic containers or wrapping, such as aluminum foil.)

Microwave heating

The relatively high frequencies of X-ray and gamma radiation are likely to change the internal structure of atoms, molecules, and solids (crystals), regardless of whether or not resonance is involved (see RADIATION EFFECTS ON MATTER). This effect is related to their great penetrating power; photons of this energy are not subject to surface scattering and reflection, such as is common with visible light at the boundaries of opaque materials. One should note that classification of high-energy photons according to their origin instead of their energy is becoming a thing of the past. Heretofore, by convention, radiations emitted from the deceleration of electrons or by transitions of electrons to inner atomic shells of atoms were called X-rays, and radiations caused by changes in nuclear energy states were called gamma rays. Part of the complication with earlier nomenclature arose with the recognition of annihilation radiation, photons of energy usually about 500,000 electron volts (0.5 MeV) emitted when a positive and negative electron combine. In this section, the term gamma ray will be used for any high-energy radiation, regardless of origin.

Absorption and scattering processes. Rayleigh scattering (see below) by tightly bound electrons in heavy elements is important for photon energies up to about 1 million electron volts, but Compton scattering is much more important and is the only observed scattering process for higher energies. The reason for this is that the binding energies of electrons in matter are much less than the high energy of hard gamma rays, and so all the electrons essentially behave as if they were free. As was seen earlier in Compton scattering, the scattered radiation has a longer wavelength and is less energetic than the incident ray, so that Compton scattering is accompanied by true absorption of energy by the material, unlike Rayleigh scattering. The attenuation of radiation through matter is then a combination of scattering and absorption in this process. In fact, Compton collisions are the most important absorption mechanism for photon energies in the range from 0.5 MeV to about 10 MeV, for all types of matter, and remain dominant for higher energy photons in materials made up of light elements.

At lower frequencies, and especially for heavy elements, the photoelectric effect is by far the most important absorption process. In this effect the quantum is absorbed as a whole, and an electron is ejected from the atom or molecule, as in Einstein's equation, except that the electrons have various binding energies, and the reaction is much more complicated than originally conceived. The dependence on atomic number is shown by the fact that

the photoelectric effect is the dominant mode of absorption in the element lead (atomic number 82) up to about 0.5 MeV, whereas it dominates the absorption in aluminum (atomic number 13) only for energies less than 0.1 MeV. There is no simple theory of photoelectric absorption because of the complications in the binding of electrons in atoms. Discontinuous jumps occur in the cross section for absorption—for example, at each resonance frequency corresponding to the ionization energy (energy required to completely remove an electron) of electronic subshells.

For gamma rays with photon energies above 1.02 MeV another type of absorption becomes increasingly important. This energy is just twice the rest-mass energy of the electron, and in the field of the nucleus a photon may be converted into an electron pair (an electron and positron; see below); their total energy, rest mass plus kinetic energy, is equal to the photon energy. The nucleus is necessary to the reaction to conserve momentum; because of its large mass, however, it receives practically none of the photon energy. For photon energies greater than four times the electron rest mass, an electron–positron pair may also be produced in the presence of an electron, and for extremely high photon energies the electrons in matter may become as effective as atomic nuclei in pair production. Pair production is the dominant mechanism for photon absorption for energies above about 5 MeV, except for the lightest elements. The positive electrons have no permanent existence: they interact with negative electrons to mutually annihilate each other, with the emission of their total energy as electromagnetic radiation, ordinarily two quanta of roughly equal magnitude, the annihilation radiation mentioned above. More rarely, in the field of a nucleus or another electron, this annihilation may result in a single quantum, a process that is the exact inverse of pair production. This radiation is indistinguishable from other electromagnetic radiation and is subject to the same scattering and absorption processes as is the primary beam of gamma rays.

Role of nucleus in pair production

The energy absorbed from gamma rays is almost entirely converted into the kinetic energy of what are called secondary electrons, and the effects of radiation on matter are produced almost exclusively by secondary electrons. Chemical and structural binding energies are of the order of a few electron volts, and the only permanent changes likely to be produced directly by a single high-energy photon are induced within the nucleus of the atom, a relatively rare event. The electrons that arise from the processes considered, however, dissipate their energy gradually, largely in the production of numerous pairs of ions within the material. These dislocations of external electrons and disruptions of chemical bonds resulting from a single high-energy photon absorption may be numerous. Each initial secondary electron may undergo a thousand collisions before being captured.

The action of radiation on complicated systems is the subject of much study (see RADIATION, BIOLOGICAL EFFECTS OF). Both theory and experiment are much simpler in treating materials consisting of a single kind of atom or of individual identical molecules than in determining what happens in inhomogeneous matter.

RELATION OF ELECTROMAGNETIC RADIATION TO QUANTUM THEORIES AND RELATIVITY

Classical mechanics as developed by Newton and others is relativistic in the restricted sense that it makes rest and uniform motion indistinguishable. This concept is in accord with everyday experience, for it is only the nonuniformities of motion that make a person conscious of travel with respect to the earth's surface, except as he sees objects move with respect to himself. Formally, the classical principle of relativity is stated as the law of inertia, according to which bodies maintain their state of rest or of uniform motion unless acted upon by an external force. The property of matter that measures this inertia is mass. Newton's laws of motion deal with forces and accelerations, departures from uniform motion. Mechanical waves are transmitted by material media that themselves have inertia, and the entire description of mechanical effects is self-consistent. All phenomena can be described classically in terms of a system of three-dimensional space coordinates, at rest or in uniform motion with respect to each other, and of a time scale that is universal, the same everywhere and for all kinds of motion. Any reference system in which the law of inertia holds is called an inertial frame of reference, and there is thus an infinity of inertial frames.

Relativity in the restricted sense did not extend to electromagnetic radiation as predicted by Maxwell and developed by others in the second half of the 19th century. The assumed medium of transmission (*i.e.*, the ether) constituted a unique frame of reference, corresponding to absolute rest. All efforts to detect motion through such a medium failed, however, and in 1905 Einstein proposed a new principle of relativity, with which the propagation of electromagnetic radiation with a universal velocity is compatible, and according to which absolute uniform motion is impossible to define. The revolutionary conceptual consequences of the special relativity theory include the fact that no signal, be it photons or material particles, can travel faster than c, the velocity of light and other electromagnetic radiation in free space; the equivalence of mass and energy; and the relativity of simultaneity—the fact that one's time scale depends on his frame of reference. In Maxwell's equations and the theory of radiation, however, no changes were necessary except in interpretation: the formal expressions of the laws of electrodynamics and the retarded potentials (see below for a definition of retarded potentials in the section *The relation of radiation fields to sources*) and fields remain unchanged, but according to special relativity they are equally valid in any and all inertial frames of reference. On the other hand, it was necessary to modify the laws of mechanics to bring them into conformity with this new relativity principle, and relativity theory extends the appearance of c, as a constant of nature, into all parts of physics. This constant enters the transformation equations that relate the space coordinates and time scale of one inertial frame to those of another, the so-called Lorentz transformation. The laws of electrodynamics are said to be Lorentz covariant, and the laws of mechanics must also be written so as to be Lorentz covariant if any velocities are involved that approach the velocity of light. Thus the theory that arose in response to an apparent failure of electromagnetic theory served instead as a basis for the reformulation of classical mechanics.

Einstein's theory of relativity

The classical theory of electromagnetic radiation, however, is correct only insofar as it is permissible to neglect all effects that arise from the finite size of h, Planck's constant. The introduction of the quantum and the accumulation of evidence for it has already been discussed. In the theory of quantum mechanics for particles, the quantization of radiation is a logically integral part. Quantum mechanics as initially developed by an Austrian physicist, Erwin Schrödinger, and Heisenberg, however, is a nonrelativistic theory, and even the relativistic wave equation formulated by Dirac in 1928 to describe the electron applies strictly only to a single particle. Moreover, the Dirac equation has solutions corresponding to negative energy that cannot be ignored but that could not be taken at face value. The interpretation of these states became quite clear with the discovery of the positron, a particle with all the attributes of the electron except for its positive charge. According to Dirac's "hole theory" all negative energy states are ordinarily filled up with electrons, a uniformity that is impossible to detect and that may be called the electron vacuum. A hole in the distribution of negative energy states constitutes the antielectron, with positive charge and positive energy. Charge is strictly conserved, and a positron appears only as a member of an electron pair. The transition of an electron from a negative energy state to one of positive energy results in the creation of a pair, and the reverse transition constitutes the annihilation of a pair. It is now known that there is an antiparticle for every kind of particle, (*e.g.*, there exists an antiproton having negative charge).

The quantum theory of the pure radiation field, excluding charges and other sources, was also first developed by Dirac and was generalized by Heisenberg and Wolfgang Pauli, an Austrian-born physicist. Complications arise that have no counterpart in classical theory, but it is possible to describe a vacuum radiation field in terms of quantized normal modes in such a way that Maxwell's equations are satisfied in the limit of Planck's constant approaching zero ($h \to 0$), in addition to Lorentz covariance as required by special relativity. These normal modes correspond to photons, which individually carry energy, momentum, and angular momentum (the property of a body associated with a resistance to a change in angular velocity).

In the actual physical world, however, electromagnetic radiation interacts with matter, and particles serve as sources and absorbers for photons. In fact, all physically observable effects of radiation result from such interactions. To describe these effects, coupling terms are introduced as additions to the pure radiation field and that of the charged particle. Many complications arise that have no counterpart in classical theory. As an example, the negative energy states of the Dirac theory are affected by the existence of an electromagnetic field, resulting in what is called the polarization of the vacuum. A constant field produces a constant polarization, which would be unobservable, but straightforward attempts to compute its magnitude have led to the embarrassing result that the effect is infinite. This defect and other divergences of the theory can be remedied by a procedure called renormalization of mass and charge. Inhomogeneous fields produce finite and observable effects that can be calculated. A series expansion is made, essentially in the quantity e^2/hc, the fine structure constant, but the renormalization may be extended to all orders of approximation, so that in principle all results may be obtained quantitatively, as corrections to the predictions of ordinary wave mechanics. The computation of these radiative corrections has been motivated by the experimental discovery of discrepancies between theory and observation on atomic energy levels and the magnetic moment of the electron. To the limit of present-day accuracy, the calculated effects are in complete agreement with experiment.

Vacuum polarization

But the situation cannot be considered satisfactory. In the words of Dirac,

> Physicists have constructed a method for handling the divergent integrals [infinities], the renormalization method, which is successful in many cases in leading to well-defined physical results that can be compared with experiment. However, the method can be considered only as a stopgap, since it does not follow logically from established physical laws.

A more logical theory may involve a revision of some of the fundamental laws of quantum mechanics or a reinterpretation of other basic concepts of physical theory. At present no satisfactory answer to these difficulties is in sight.

The general theory of relativity is a theory of gravitation and is not logically connected with the theory of electromagnetism. Quantum mechanics is also outside the scope of general relativity. General relativity is essentially concerned with the geometry of space-time, and gravitation is viewed as a geometric phenomenon, as time has the attribute of a fourth dimension. In electromagnetic theory and in the special theory of relativity, space and time are interrelated, but they play a purely passive role and are quite independent of the presence of matter and radiation. There have been many attempts to formulate a unified field theory comprising both electrodynamics and gravitation, but none has been successful. The velocity of light plays a role in general relativity, to be sure; and on a local scale the geometry is Euclidean, adequately represented by the inertial frames of the theory of special relativity. In the 1970s rapid advances were being made in general relativity, partly as a result of such significant discoveries in observational astronomy as quasi-stellar objects, pulsars, and exploding galaxies. These investigations of singularities in the theory, the prediction of gravitational collapse, leading to the "black hole" from which electromagnetic radiation cannot escape, are spectacular features of modern astrophysics, for which observational evidence is being sought. These are advances in cosmology and the theory of gravitation, however, and not in the theory of radiation as such. A relation between gravitation and electromagnetism must surely exist, but a full understanding of it seems at present even more remotely obtainable than the resolution of the logical difficulties in quantum electrodynamics.

Unified field theory

III. Mathematical formulation of electromagnetic theory

MAXWELL'S EQUATIONS

Before presenting the mathematical formulation of the laws of electromagnetism, it is necessary to adopt a system of units and dimensions in order to describe electromagnetic phenomena quantitatively, and the choice is sometimes a matter of controversy. The complication arises from the fact that electromagnetic theory, unlike classical mechanics, involves a fundamental constant of nature, c, the velocity of light *in vacuo*, as an essential feature of the theory. (Special relativity extends this feature to all of physics.) This velocity constant c occurs explicitly in some of the equations involving the Gaussian system of units, a system widely used in atomic physics and allied fields. The General Conference on Weights and Measures, which makes international decisions on the definitions of metric units, however, adopted in 1960 the International System of Units (denoted by SI), a system that incorporates technical units such as the ampere, the volt, and the coulomb. In fact, the ampere, a measure of current flow, is one of the basic SI units, and a basic unit of charge, the coulomb, is defined in this system as a secondary unit composed of current times time, or ampere-second. The SI system is particularly suitable for treating applications that involve both circuits and fields. As applied to electromagnetic theory, the SI system of units employs two constants, the dielectric constant of free space (epsilon ε_0) and the permeability of free space (mu μ_0), related by the velocity c, *i.e.*, $\mu_0\varepsilon_0 = 1/c^2$, and so only one of the constants is fundamental to the theory. Regardless of the system used, changing units or even dimensions does not affect the physical content of the equations. The numerical values of the constants involved depend on whether or not the equations are rationalized; the geometric factor 4π appears explicitly in equations implying spherical symmetry, if rationalized constants are employed; and is suppressed, for example, in the continuous source equations. Rationalized SI units will be employed in the sections to follow. It will be seen that all formulas are written so that they may be reduced to their Gaussian equivalents by the substitutions $4\pi\varepsilon_0 \to 1$ and $c\boldsymbol{B}_{\mathrm{SI}} \to \boldsymbol{B}_{\mathrm{Gaussian}}$, after using $\mu_0\varepsilon_0 = 1/c^2$ to eliminate the permeability of free space if it appears.

SI system of units

Written so as to apply to macroscopic charge and current configurations, the first of the fundamental empirical laws of electricity and magnetism is Coulomb's law of electrostatics, expressed in the form called Gauss's theorem (in the following equations, the symbol $\times$ indicates a cross product and the symbol $\cdot$ indicates a scalar product):

$$\int_{\text{closed}} \boldsymbol{E}\cdot d\boldsymbol{S} = \frac{1}{\epsilon_0}\int \rho dV = Q/\epsilon_0 .$$

in which the integral of the normal component of the electric field intensity $\boldsymbol{E}$ over a closed surface (scalar product of $\boldsymbol{E}$ and surface element $d\boldsymbol{S}$) is equal to $1/\varepsilon_0$ times the total charge Q inclosed by the surface; *i.e.*, the integral of the charge density (ρ) over the volume (V) inclosed by the surface. The field strength $\boldsymbol{E}$ is defined operationally as the force per unit positive charge, taken in the limit of a vanishingly small test charge so as not to perturb the measured field. The SI units of the field strength $\boldsymbol{E}$ are units of force per charge; that is, newtons per coulomb.

Because isolated magnetic poles have not been found to exist, the operational definition of the magnetic field

$\boldsymbol{B}$ is made in terms of force on a unit current element: a current I in an elementary line dl will experience a force ($\boldsymbol{F}$) given by the vector cross product of the elementary line and the magnetic field, $\boldsymbol{F} = I\boldsymbol{dl} \times \boldsymbol{B}$. The unit of the magnetic field $\boldsymbol{B}$ is thus a newton per amperemetre. The field $\boldsymbol{B}$ satisfies the relation that the integral of the normal component of the magnetic field over a closed surface is zero, namely, $\int \boldsymbol{B} \cdot d\boldsymbol{S} = 0$. In terms of magnetic poles, this is a statement of the inverse square law plus the fact that there are apparently no free poles.

The circulation laws

The relations between electric and magnetic fields are given by the circulation laws. The closed line integral of $\boldsymbol{E}$ about a loop is equal to the electromotive force given by Faraday's law of induction:

$$\oint \boldsymbol{E} \cdot d\boldsymbol{s} = -\frac{d}{dt}\int \boldsymbol{B} \cdot d\boldsymbol{S}.$$

The surface integral of $\boldsymbol{B}$ over any surface bounded by the closed path of the integral on the left is called the magnetic flux through the surface. Ampere's law may also be written as a circulation law: the line integral of the magnetic field about a closed path is proportional to the total current threading the loop,

$$\oint \boldsymbol{B} \cdot d\boldsymbol{s} = \mu_0 I = \mu_0 \int \boldsymbol{j} \cdot d\boldsymbol{S},$$

in which μ_0 is the permeability of free space, and the total current is equal to the integral of the current density $\boldsymbol{j}$ over a surface bounded by the path of the line integral of $\boldsymbol{B}$. This circulation law is valid for steady currents, but in the presence of time-dependent fields a term must be added to account for the conservation of charge.

By means of purely geometrical theorems, those of Gauss and Stokes, which relate closed surface and line integrals to the vector operators divergence and curl, symbolized by $\nabla\cdot$ and $\nabla\times$, respectively, these macroscopic integral laws may be transformed to differential relations between the fields, point by point. The result is the set of Maxwell's equations. As valid *in vacuo*, these relations may be simply written in terms of the electric and magnetic fields. The charge density (ρ) and the current density ($\boldsymbol{j}$) are assumed to exist in the absence of any continuous material media, and the four Maxwell equations are written

$$\nabla \cdot \boldsymbol{E} \equiv \operatorname{div} \boldsymbol{E} = \rho/\epsilon_0,$$

$$\nabla \cdot \boldsymbol{B} \equiv \operatorname{div} \boldsymbol{B} = 0,$$

$$\nabla \times \boldsymbol{E} \equiv \operatorname{curl} \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \equiv -\dot{\boldsymbol{B}},$$

$$\nabla \times \boldsymbol{B} \equiv \operatorname{curl} \boldsymbol{B} = \mu_0\left(\boldsymbol{j} + \epsilon_0\frac{\partial \boldsymbol{E}}{\partial t}\right) \equiv \mu_0(\boldsymbol{j} + \epsilon_0\dot{\boldsymbol{E}}).$$

(The first and second time derivative of any variable, say x, will be written hereafter as $\dot{x}$ and $\ddot{x}$, respectively.) Here the permeability of free space $\mu_0 = 4\pi \times 10^{-7}$ kilogram-metre per coulomb^2, and the product of the permeability and dielectric constant of free space equals 1 divided by the velocity squared, or $\mu_0\epsilon_0 = 1/c^2$. It is thus evident that the permeability is defined in terms of an arbitrarily chosen electrical unit of current, and the dielectric constant is determined by this definition and the experimental value of the velocity of light, which is almost exactly 3×10^{10} centimetres per second.

These equations correspond to the four empirical laws listed in the section treating radiations semiquantitatively, the third equation here being Faraday's law of electromagnetic induction. The addition to the ordinary current density $\boldsymbol{j}$ in the fourth equation is the vacuum displacement current, and it is necessitated by the continuity equation expressing the conservation of charge,

Displacement current

$$\nabla \cdot \boldsymbol{j} \equiv \operatorname{div} \boldsymbol{j} = -\frac{\partial \rho}{\partial t}.$$

Unlike the true current density $\boldsymbol{j}$, the vacuum displacement current does not correspond to the motion of charge, but its presence makes possible a nonvanishing circulation of $\boldsymbol{B}$ in the absence of charges and currents, analogous to the circulation of $\boldsymbol{E}$ as given by the law of induction. In a dielectric (insulating) medium at rest, a volume electric polarization ($\boldsymbol{P}$; see below) is produced by the presence of an electric field, and the displacement current includes a term that corresponds to the motion of bound charge in a dielectric. The time rate of change of dipole moment per unit volume is dimensionally the same as a current density, and $\dot{\boldsymbol{P}}$ is quite literally a displacement current density. It is convenient to define another field quantity, called the displacement field, by the equation $\boldsymbol{D} = \epsilon_0\boldsymbol{E} + \boldsymbol{P}$, the first derivative of the displacement with time, or in terms of which the displacement current is simply $\dot{\boldsymbol{D}}$. In empty space polarization vanishes, and thus there is no distinction between $\boldsymbol{D}$ and $\epsilon_0\boldsymbol{E}$.

In magnetic materials, analogously, a second field quantity takes account of the volume magnetization, or magnetic dipole moment per unit volume, $\boldsymbol{M}$, the curl of which is the magnetization current. This field is denoted by $\boldsymbol{H}$ and may be defined by the equation

$$\boldsymbol{H} = \boldsymbol{B}/\mu_0 - \boldsymbol{M}.$$

The dielectric and permeability constants ϵ_0 and μ_0 do not appear explicitly in Maxwell's equations if the displacement field and the magnetic field $\boldsymbol{H}$ are introduced. Instead, they are simply

$$\operatorname{div} \boldsymbol{D} = \rho,$$

$$\operatorname{div} \boldsymbol{B} = 0,$$

$$\operatorname{curl} \boldsymbol{E} = -\dot{\boldsymbol{B}},$$

$$\operatorname{curl} \boldsymbol{H} = \boldsymbol{j} + \dot{\boldsymbol{D}}.$$

The equations in this form apply to material media as well as to empty space.

In the absence of all charges and currents the equations for vacuum electrodynamics take on a still simpler form:

$$\operatorname{div} \boldsymbol{E} = 0,$$

$$\operatorname{div} \boldsymbol{B} = 0,$$

$$\operatorname{curl} \boldsymbol{E} = -\dot{\boldsymbol{B}},$$

$$\operatorname{curl} \boldsymbol{B} = \mu_0\epsilon_0\dot{\boldsymbol{E}} = \frac{1}{c^2}\dot{\boldsymbol{E}}.$$

These last four equations lead at once to the prediction of electromagnetic waves that would traverse empty space with velocity c.

TRANSMISSION OF RADIATION IN FREE SPACE

The wave equation and electromagnetic waves. On eliminating the magnetic field by combining the two vector Maxwell equations written above, after taking their curl, one finds that

$$\operatorname{curl}\operatorname{curl} \boldsymbol{E} \equiv \operatorname{grad}\operatorname{div} \boldsymbol{E} - \nabla^2\boldsymbol{E} = -\frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2},$$

or

$$\nabla^2\boldsymbol{E} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} = 0,$$

as $\operatorname{div} \boldsymbol{E} = 0$ in the absence of charge. If the electric field is eliminated from the two vector Maxwell equations in an analogous way, it is seen that the magnetic field must satisfy the same equation

$$\nabla^2\boldsymbol{B} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{B}}{\partial t^2} = 0.$$

Homogeneous wave equations

This and the preceding equation are called homogeneous wave equations.

In Cartesian coordinates the homogeneous wave equation in the vector field $\boldsymbol{E}$ (x, y, z, t) may be written

$$\frac{\partial^2 \boldsymbol{E}}{\partial x^2} + \frac{\partial^2 \boldsymbol{E}}{\partial y^2} + \frac{\partial^2 \boldsymbol{E}}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2}.$$

The meaning of this equation is readily exhibited if one assumes a plane wave in which the electric field is uniform over a plane. The coordinate system may be chosen such that this plane is perpendicular to the z-axis; then the electric vector is the same for all values of x and y for a particular value of z, and the derivatives with respect to these two coordinates vanish. The equation becomes

$$\frac{\partial^2 \boldsymbol{E}(z,t)}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}(z,t)}{\partial t^2}.$$

This second-order relation is satisfied by any vector function of the argument $(z \pm ct)$. The function f $(z \pm ct)$ represents a disturbance that travels in the direction of $\mp z$ with velocity c.

The first-order Maxwell equations must also be satisfied, of course, and because the divergence of the electric field is zero, it follows that the z component of the electric field vanishes except for uniform static fields. All these remarks apply equally to the magnetic field, and thus the wave is transverse, not longitudinal. The electric and magnetic fields, however, are not independent. It may be assumed without loss of generality that the wave is plane polarized, with the electric field vector $\boldsymbol{E}$ lying in the positive x direction in a given plane at a given time t. Then if $\boldsymbol{n}$ is a unit vector in the direction of positive z, normal to the wave front, Maxwell's curl $\boldsymbol{E}$ equation reduces to

$$\boldsymbol{n} \times \frac{\partial \boldsymbol{E}}{\partial z} = -\frac{\partial \boldsymbol{B}}{\partial t},$$

which is in the positive y direction. Thus the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ are perpendicular to each other as well as to the direction of propagation. In fact, it follows generally that a plane wave moving in the direction of increasing z, for which the electric field has the form $\boldsymbol{E}(z - ct)$, has an accompanying magnetic field $\boldsymbol{B}(z - ct)$ such that $c\boldsymbol{B} = \boldsymbol{n} \times \boldsymbol{E}$.

The electric and magnetic fields and the direction of propagation, in that order, form a right-handed set of orthogonal vectors. The relative magnitude of the electric and magnetic fields in free space is $E = cB$ in these units. In Gaussian units the magnitudes of $\boldsymbol{E}$ and $\boldsymbol{B}$ are equal $(E = B)$.

Energy density

The energy density corresponding to an electrostatic field in free space is

$$\frac{1}{2}\epsilon_0 E^2 = \frac{1}{2}ED,$$

as may be computed in an elementary way from the energy stored in an idealized capacitor. Similiarly the magnetostatic energy density is $\frac{1}{2}B^2/\mu_0 = \frac{1}{2}HB$, as computed in an elementary way from the energy stored in an idealized solenoid (coil of wire). The two vector Maxwell equations may be combined by vector manipulation in the following way to give an energy integral that is interpreted in terms of the extension of these expressions to time-varying fields and the flux of energy carried by the radiation. Scalar multiplication of the curl $\boldsymbol{E}$ equation by $\boldsymbol{H}$, and of the curl $\boldsymbol{H}$ equation by $\boldsymbol{E}$, after which one is subtracted from the other, yields

$$\boldsymbol{H} \cdot (\text{curl } \boldsymbol{E} + \dot{\boldsymbol{B}}) - \boldsymbol{E} \cdot (\text{curl } \boldsymbol{H} - \boldsymbol{j} - \dot{\boldsymbol{D}}) = 0.$$

According to a general vector identity,

$$\boldsymbol{H} \cdot \text{curl } \boldsymbol{E} - \boldsymbol{E} \cdot \text{curl } \boldsymbol{H} = \text{div } (\boldsymbol{E} \times \boldsymbol{H}),$$

so that the equation becomes

$$\text{div } (\boldsymbol{E} \times \boldsymbol{H}) + (\boldsymbol{E} \cdot \dot{\boldsymbol{D}} + \boldsymbol{H} \cdot \dot{\boldsymbol{B}}) + \boldsymbol{j} \cdot \boldsymbol{E} = 0.$$

In this equation, the third term $\boldsymbol{j} \cdot \boldsymbol{E} = \rho \boldsymbol{v} \cdot \boldsymbol{E}$ is clearly the rate at which the electric field does work per unit volume on any charge density ρ, and from the relations $\boldsymbol{D} = \epsilon_0 \boldsymbol{E}$ and $\boldsymbol{B} = \mu_0 \boldsymbol{H}$ it follows that the second term

$$(\boldsymbol{E} \cdot \dot{\boldsymbol{D}} + \boldsymbol{H} \cdot \dot{\boldsymbol{B}}) = \epsilon_0 \dot{\boldsymbol{E}} \cdot \boldsymbol{E} + \boldsymbol{B} \cdot \dot{\boldsymbol{B}}/\mu_0$$
$$= \frac{\partial}{\partial t}(\boldsymbol{E} \cdot \boldsymbol{D} + \boldsymbol{H} \cdot \boldsymbol{B})/2$$

is the rate of change of electromagnetic energy density if the latter is identified with $(\boldsymbol{E} \cdot \boldsymbol{D} + \boldsymbol{H} \cdot \boldsymbol{B})/2$. The remaining term is the divergence of the vector $(\boldsymbol{E} \times \boldsymbol{H})$, which may be interpreted as the flux of energy per unit area through a surface normal to the direction of propagation. This flux vector is called the Poynting vector, and the equation relating its divergence to the time rate of change of the electromagnetic energy density is called Poynting's theorem. In a charge-free region, Poynting's theorem states the law of conservation of electromagnetic energy: the time rate of change of the field energy density is equal to the negative divergence of the energy flux.

In a plane progressive wave, with $\boldsymbol{E}$ and $c\boldsymbol{B}$ having the same magnitude, it is seen that the energy density is divided equally between the electric and magnetic fields, and the total energy density is equal to $\epsilon_0 E^2$ at any point. The Poynting vector is therefore $\boldsymbol{E} \times \boldsymbol{H} = c\epsilon_0 E^2 \boldsymbol{n}$, in which $\boldsymbol{n}$ is again the unit vector normal to the wave front. Electromagnetic radiation also possesses momentum in the direction of its propagation. A formula for the momentum density was derived by Maxwell on the assumption of stresses in the ether, but the simplest method to obtain the correct expression is to invoke the principles of special relativity. As has been shown, there is an energy density associated with the electromagnetic radiation fields, which may be denoted by W, and this should correspond to a mass density W/c^2 that is transmitted through space with velocity c. Thus there is also a momentum density of magnitude W/c, a vector quantity in the direction of propagation that in free space is equal to $\boldsymbol{E} \times \boldsymbol{H}/c^2$, the Poynting vector divided by c^2.

Radiation pressure

The momentum of electromagnetic radiation gives rise to radiation pressure. The magnitude of the force exerted per unit area of a material body depends on the reflectivity of the surface: the pressure on a perfect absorber normal to the Poynting vector is equal to W, the energy density, and that on a perfect reflector would be twice as great. Isotropic radiation, as in the ideal black body cavity, consists of randomly directed fields, and the radiation pressure is one-third the energy density on an absorbing probe or on the cavity walls. Radiation pressure is an extremely small effect under ordinary terrestrial conditions, but it plays an important role in the internal dynamics of stars.

An infinite plane wave of the kind being discussed would carry no angular momentum, even if the electric vector rotated periodically in the plane perpendicular to its propagation (circularly polarized light). The existence of intrinsic angular momentum in electromagnetic radiation would necessitate longitudinal field components, and it was shown above that radiation fields are transverse. A plane wave, the wave front of which is of finite extent, however, does involve small longitudinal components, and the angular momentum of a circularly polarized beam of light was measured by a United States physicist, Richard Alexander Beth, in 1936. The magnitude of the angular momentum density is related to the energy density and to the frequency of the light, and for circularly polarized light is in fact given by the energy density divided by 2π times the frequency, or $W/2\pi\nu$, in which ν is the frequency. Because angular momentum must be conserved, the effect can be detected by measuring the torque, or twist, on a half-wave plate, a thin slab of crystal that changes left-hand polarized light to right-hand polarized light, and vice versa; that is, it reverses the direction of rotation of the electric vector in circularly polarized light. The measured result was in accord with the theory to an accuracy of 10 percent. The concept of angular momentum is more useful as applied to spherical waves than to plane waves, as will be seen, and photons have intrinsic angular momentum related to their energy by the same factor, $2\pi\nu$, that is given by the wave theory of radiation.

The relation of radiation fields to sources. In order to relate radiation fields to their sources, accelerated charges, or changing currents, it is convenient to introduce potentials. The magnetic and electric fields can in general be derived from a vector potential (A) and a scalar potential (ϕ), which are functions of the space coordinates and the time, by the relations

$$\boldsymbol{B} = \operatorname{curl} \boldsymbol{A}$$

$$\boldsymbol{E} = -\operatorname{grad}\phi - \frac{\partial \boldsymbol{A}}{\partial t}.$$

The potentials are not completely defined by these equations, but the Maxwell equations that do not involve sources are automatically satisfied if the fields are expressed in this way. The other two equations become

$$\nabla^2\phi + \operatorname{div}\frac{\partial \boldsymbol{A}}{\partial t} = -\rho/\epsilon_0,$$

$$\nabla \times (\nabla \times \boldsymbol{A}) + \frac{1}{c^2}\frac{\partial \boldsymbol{A}}{\partial t^2} + \frac{1}{c^2}\operatorname{grad}\frac{\partial \phi}{\partial t} = \mu_0 \boldsymbol{j}.$$

The divergence of $\boldsymbol{A}$ and the time derivative of ϕ are not specified by their definitions in terms of the fields, and may be chosen so that

$$\operatorname{div}\boldsymbol{A} + \frac{1}{c^2}\frac{\partial \phi}{\partial t} = 0.$$

This is called the Lorentz condition on the potentials, after Hendrik Antoon Lorentz, a Dutch physicist. If this condition is satisfied, the potentials are solutions of the second-order inhomogeneous wave equations,

$$\nabla^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = -\rho/\epsilon_0,$$

$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\mu_0 \boldsymbol{j}.$$

The fields excited by the charge and current sources ρ and j are determined by differentiating the potentials, which are particular solutions of these equations.

Retarded potentials

It was first shown by Ludvig Valentin Lorenz, a Danish physicist, that solutions may be written formally as volume integrals over the source distributions evaluated at the retarded time, equal to the time (t) for which the field is described minus the distance travelled divided by the velocity, or $t' = t - r/c$, which would allow transmission of the fields from the source with velocity c; *i.e.*,

$$\phi(x, y, z, t) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(x', y', z', t - r/c)}{r}\,dV',$$

$$\boldsymbol{A}(x, y, z, t) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{j}(x', y', z', t - r/c)}{r}\,dV',$$

in which r is the distance from a source point (primed coordinates) to the field point (plain coordinates). These are the retarded potentials; mathematically the advanced potentials, with the integrals evaluated at $t + r/c$, are equally acceptable, but in classical radiation theory no physical significance can be attached to such solutions.

A simple example will show that the retardation is responsible for the existence of radiation fields, which must fall off as the inverse first power of the distance, and not faster. If a charge e is restricted to a small region of space near the origin of coordinates, then, to a first approximation, the vector potential at a large distance r is simply

$$\boldsymbol{A}(x, y, z, t) = \frac{\mu_0}{4\pi}\frac{e\boldsymbol{v}(t - r/c)}{r}.$$

The relation $\boldsymbol{B} = \operatorname{curl}\boldsymbol{A}$ requires taking the space derivatives, each term of which would behave as $1/r^2$ at best except for the fact that the distance r is also involved in the retarded time. The radiation field is thus proportional to the time derivative of the velocity. In particular, if the motion of the charge is confined to the z axis in this simple approximation, the magnetic field is azimuthal, and given by

$$B_\varphi = \frac{\mu_0}{4\pi c}\frac{q\dot{v}}{r}\sin\theta = \frac{1}{4\pi\epsilon_0}\frac{q\dot{v}\sin\theta}{rc^3}.$$

Here θ is the angle between the z axis and the line r from source to observer. The corresponding electric vector is along the direction of change in angle θ and equal in magnitude to cB in these units.

The exact potentials corresponding to a point charge, in contrast to this approximation are not obtained by substituting the total charge for the volume integral indicated in the retarded potentials. The problem arises because of the finite velocity of field propagation, so that the integral of the retarded charge density over space is not in general equal to the total charge. The correct potentials for an electron of charge e are

$$\phi = \frac{1}{4\pi\epsilon_0}\left[\frac{e}{R - \boldsymbol{R}\cdot\boldsymbol{v}/c}\right],$$

$$\boldsymbol{A} = \frac{\mu_0}{4\pi}\left[\frac{e\boldsymbol{v}}{R - \boldsymbol{R}\cdot\boldsymbol{v}/c}\right],$$

in which $\boldsymbol{R}$ is the radius vector from the electron to the field point, v is the velocity of the electron, and the square bracket is used to denote the fact that all quantities are to be evaluated at the retarded time, not the time of observation. These are called the Liénard-Wiechert potentials. Before the advent of relativity theory they could not be justified for high-velocity electrons, because it was necessary to assume that Maxwell's equations were valid only in the reference frame of the observer, but the Liénard-Wiechert potentials are relativistically correct and also valid for the entire range of classical electrodynamics.

Liénard-Wiechert potentials

Historically, the radiation fields were first traced to their sources by Hertz, who assumed the existence of an infinitesimal oscillating electric dipole at the origin of coordinates. A vector expression for the fields of an arbitrarily oriented dipole may be written in terms of the dipole moment p, and $\hat{r}$, a unit vector in the direction of the radius vector r from the source to the point of observation:

$$\boldsymbol{B} = \frac{\mu_0}{4\pi}\left(\frac{\ddot{\boldsymbol{p}}\times\hat{\boldsymbol{r}}}{c^2 r} + \frac{\dot{\boldsymbol{p}}\times\hat{\boldsymbol{r}}}{cr^2}\right),$$

$$\boldsymbol{E} = \frac{1}{4\pi\epsilon_0}\left[\frac{\hat{\boldsymbol{r}}\times(\hat{\boldsymbol{r}}\times\ddot{\boldsymbol{p}})}{c^2 r} + \frac{2\hat{\boldsymbol{r}}(\hat{\boldsymbol{r}}\cdot\dot{\boldsymbol{p}}) + \hat{\boldsymbol{r}}\times(\hat{\boldsymbol{r}}\times\dot{\boldsymbol{p}})}{cr^2} + \frac{2\hat{\boldsymbol{r}}(\hat{\boldsymbol{r}}\cdot\boldsymbol{p}) + \hat{\boldsymbol{r}}\times(\hat{\boldsymbol{r}}\times\boldsymbol{p})}{r^3}\right]$$

Here $\ddot{p}$ is the second time derivative of the dipole moment, to be evaluated at the retarded time $t - r/c$. The initial terms comprise the radiation fields, and the Poynting vector is in the direction of increasing r,

$$\boldsymbol{E}\times\boldsymbol{H} = \frac{[\ddot{\boldsymbol{p}}^2 - (\hat{\boldsymbol{r}}\cdot\ddot{\boldsymbol{p}})^2]\hat{\boldsymbol{r}}}{16\pi^2\epsilon_0 c^3 r^2}.$$

Whereas the magnetic field has no component along r, the second term in $\boldsymbol{E}$ is partly in the direction of p. It is this term that can account for the radiation of angular momentum from a rotating dipole.

If the dipole is constrained to lie along the polar axis, both radiation fields are proportional to sine of the polar angle (θ), and the Poynting vector is simply

$$\boldsymbol{N} = \boldsymbol{E}\times\boldsymbol{H} = \frac{\ddot{\boldsymbol{p}}^2\hat{\boldsymbol{r}}\sin^2\theta}{16\pi^2\epsilon_0 c^3 r^2}.$$

The total energy radiated per unit time is the integral of the Poynting vector over a sphere of radius r,

$$\frac{-dU}{dt} = 2\pi r^2 \int_0^\pi N \sin\theta \, d\theta = \frac{\ddot{p}^2}{6\pi\epsilon_0 c^3}.$$

If the dipole moment is interpreted as the position of a charge e moving in the neighbourhood of a charge $-e$ fixed at the origin, $\ddot{p} = e\dot{v}$, and the rate of radiation may be written in terms of $\dot{v}$, the acceleration of the moving charge e,

$$\frac{-dU}{dt} = \frac{e^2\dot{v}^2}{6\pi\epsilon_0 c^3},$$

if U is the energy of the source. This last expression is the Larmor formula for radiation. The acceleration is to be evaluated at the retarded time, but it must be noted that this is not an unrestrictedly valid formula for the radiation from an accelerated charge. The general formula for the rate of radiation from a moving charge, as derived from the Liénard-Wiechert potentials, depends on the velocity as well as the acceleration,

$$\frac{-dU}{dt'} = \frac{e^2}{6\pi\epsilon_0 c^3} \frac{(\dot{v}^2 - (v \times \dot{v})^2/c^2)}{(1 - v^2/c^2)^3}.$$

Relativistically, the velocity is interpreted as the relative velocity of the charge e with respect to the observer, but t' is the retarded time, with respect to which the velocity and acceleration of the source are to be evaluated. The expressions for the electric and magnetic fields are complicated except for simple motions, but in all circumstances $B = r \times E/rc$, so that the magnetic field is always perpendicular to the electric field and to the retarded radius vector from the charge to the field point.

In accord with the principles of Fourier analysis, the solutions of the inhomogeneous wave equation may be constructed from components that are harmonic in time, so that the time dependence may be represented by the exponential coefficient $e^{-i\omega t}$. Here $\omega = 2\pi\nu$ is the angular frequency and i is an imaginary number $\sqrt{-1}$. This mathematical device has the advantage that introducing the factor $-i\omega$ is equivalent to taking the time derivative. This factor can be broken down by Euler's formula into the sum of two terms, one the real part and the other the complex part, so called because it contains the imaginary factor i. The actual fields are to be identified with the real part of the complex field expressions, as obtained by use of the Euler formula $e^{-i\omega t} = \cos\omega t - i \sin\omega t$, together with the real part of the corresponding formula for the complex factor in the spatial coordinates. To obtain instantaneous values of the energy density and other quantities that depend on the product of field strengths, it is necessary to write the real fields before multiplying, but it is a very simple matter to find the time average of such products. The time average of the product of the real parts of two vectors, both of which vary with the time as $e^{-i\omega t}$, is one-half the real part (Re) of the product of one vector and the complex conjugate of the other:

$$[(\mathrm{Re}\ \boldsymbol{E}) \times (\mathrm{Re}\ \boldsymbol{H})]_{\text{time average}} = \tfrac{1}{2}\mathrm{Re}\,(\boldsymbol{E} \times \boldsymbol{H}^*) = \tfrac{1}{2}\mathrm{Re}\,(\boldsymbol{E}^* \times \boldsymbol{H}),$$

for example. (The expression $\boldsymbol{H}^*$ denotes the complex conjugate of $\boldsymbol{H}$, obtained by changing the sign of i wherever it appears.) The product of a complex variable and its conjugate is entirely real. The time-averaging rule holds equally for scalar and vector products. With the time dependence given in this way, the plane wave solution of the homogeneous wave equation is written $E = E_0 \exp[i(k \cdot r - \omega t)]$, in which r is the coordinate or spatial vector of the field point and k is the propagation vector, directed along the direction of wave propagation and of magnitude $\omega/c = 2\pi/\lambda$. The space dependence of an outgoing spherical wave is asymptotically given e^{ikr}/r, with k now the magnitude of the propagation vector the direction of which is that of increasing r.

Propagation vector k

For a linear electric dipole of moment $p\, e^{-i\omega t}$, the spatial factors of the radiation fields are

$$E_\theta = \frac{-\omega^2 p \sin\theta}{4\pi\epsilon_0 c^2} \frac{e^{ikr}}{r},$$

$$H_\varphi = \frac{-\omega^2 p \sin\theta}{4\pi c} \frac{e^{ikr}}{r},$$

in which the subscripts indicate that the electric field is in the direction of changing θ, and that the magnetic field is in a plane at right angles to the propagation vector; *i.e.*, it is azimuthal. (In polar coordinates the polar axis is assumed to be in the direction of the dipole moment p.) In terms of the propagation vector k, which is along the radius vector r, the radiation fields may be written in vector form:

$$E = \frac{e^{ikr}}{4\pi\epsilon_0 r}[(p \times k) \times k],$$

$$H = -\frac{e^{ikr}}{4\pi r}(p \times k).$$

The electric field E also has a component along r that falls off as $1/r^2$, but the magnetic field is entirely azimuthal.

Because of the symmetry of the equations in the electric and magnetic fields, it is evident that an oscillating magnetic dipole would give rise to radiation having the same kind of distribution in space as does an oscillating electric dipole, except that the roles of the fields are interchanged. The electric field of a linear magnetic dipole is azimuthal, whereas the magnetic field has a radial component that falls off only as $1/r^2$ and can account for the radiation of angular momentum.

Only for a few special cases is it possible to perform exactly the volume integrals indicated formally in the expressions for the retarded potentials. For any finite source distribution that can be considered to remain in the neighbourhood of the origin of coordinates, however, a series expansion into multipole moments (dipole, quadrupole, octupole) can be made in which the variable involved is effectively $ka = \omega a/c$, a being a length characterizing the dimensions of the source. The complete radiation fields are then a summation of terms over standard distribution patterns in space, multiplied by coefficients that express the source strength of each distribution pattern. As might be expected, the angular distribution functions are just the spherical harmonics or their derivatives (solutions in terms of angular and radial components), and the radial dependence must be chosen to behave like e^{ikr}/r for large r to correspond to outgoing radiation. These are the multipole fields, which may be divided into two types: the electric multipoles, for which the magnetic field has no radial component; and the magnetic multipoles, for which the electric field has no radial component. The source strength depends on the nature of a source distribution; and because magnetic effects arise from currents, the magnetic multipole strength is generally smaller than the corresponding electric multipole by a factor ka, which for the harmonic time dependence assumed is equivalent of v/c. The radiation intensities are related by the square of this factor.

Multipole fields

It is evident from the dependence of the linear dipole fields on the factor $\sin\theta$ that there is no radiation along the axis of the dipole. Quadrupole radiation fields have two nodal cones where there is no radiation. Physically, electric quadrupole radiation would arise from a pulsating charge distribution of such symmetry that the dipole moment remains zero. The successive multipole moments ordinarily become important only if lower order moments of the source distribution vanish. Actually the integrals over the source need not be moments in the elementary sense; and if two or more moments contribute significantly, the source strength corresponding to each spherical harmonic should be represented by an integral over a more faithful representation of its distribution. In practice this integration is more often necessary in macroscopic antenna systems (the dimensions of which are not necessarily small in comparison with the wavelength of the emitted radiation) than in microscopic systems. In atomic or nuclear radiation the lowest order multipole possible is usually the only one observed.

WAVES IN SPACE BOUNDED BY CONDUCTORS

Boundary conditions. The patterns of radiant electromagnetic energy are affected by material media, whether the materials are transparent to the radiation or are opaque. The transparency of a substance to radiation depends on the wavelength of the radiation as well as on the nature of the substance, but metallic conductors reflect all electromagnetic waves except extremely high frequencies. Metallic surfaces may be made optically flat so as to reflect light in a regular way, but the most useful application of waves in regions bounded by conductors is in the microwave region of the spectrum. These applications include not only radar and other devices for communication, but are widely used as research tools in the investigation of the properties of matter.

Because of the boundary conditions imposed on the electromagnetic fields at the surface of a conductor, standing waves may be set up in a region within conducting boundaries, analogous to the one-dimensional standing waves of a vibrating string in which the wave form is stationary. A substance is called a conductor if charges are free to move within it, and in a perfect conductor there would be no resistance at all to the motion of charges. As a result of the application of external fields there are charges and the possibility of charge motion on the surface, but no fields penetrate the interior of the conducting medium. This property of real conductors is well-known in electrostatics, and its extension to fields that vary with time depends on the relation of conductivity to frequency: for good conductors such as copper and microwave frequencies (order of magnitude 10^{10} hertz), the geometrical representation of fields in cavities or wave guides corresponds to those obtained on the assumption of ideally conducting boundaries. To compute losses in cavities and the attenuation of transmitted waves, however, it is necessary to take the finite conductivity into account.

The boundary conditions on the electric and magnetic fields at the interface between two media of different electric and magnetic properties are inherent in Maxwell's equations for the fields. In the limit of high conductivity the fields are discontinuous at the surface of a conductor, but there are finite surface charges and surface currents that accompany the external fields. At the boundary the tangential component of the electric field vanishes, but there may be an electric field $\boldsymbol{E}$ normal to the surface and a magnetic field $\boldsymbol{B}$ tangential to the surface. The frequencies and field configurations that satisfy these conditions correspond to the normal modes of cavity resonators, regions of space bounded by conducting walls.

The analysis of cylindrical cavities may be immediately extended to wave guides, devices used to transmit electromagnetic energy at microwave frequencies. A wave guide, which usually has a rectangular or circular cross section, is simply a long metallic cylinder with open ends. Radiation may be introduced into wave guides by means of small loops or probes (see ANTENNAS AND WAVE GUIDES).

THE SCATTERING OF ELECTROMAGNETIC RADIATION

Matter is made up of charged particles, which by definition respond to electromagnetic fields. An electromagnetic wave periodic in time will produce periodic motion in the particles and thus produce radiation. In the process as a whole, energy is absorbed from the incident wave and re-emitted, thus scattering the electromagnetic wave. The details of the scattering depend on the nature of the scatterer and the wavelength of the radiation. First to be considered will be the scattering produced by electrons, which are forced to vibrate by an incident electromagnetic wave.

If the velocity of the electron is well below the velocity of light, the effect of the magnetic vector may be neglected. This approximation is valid for optical frequencies and even for soft (low-energy) X-rays. In the classical theory of scattering it is assumed that the electron is harmonically bound to a fixed position, so that it has a natural frequency that may be large or small compared with the frequency of the incident radiation. The forces acting on the electron comprise the binding force and a damping force corresponding to its reaction in emitting radiation, in addition to the external electric force field. If ω_0 represents the binding frequency and γ characterizes the damping, the equation of motion for the electron in the presence of an incident plane polarized electric wave field defined as $\boldsymbol{E} = \boldsymbol{E}_0 e^{-i\omega t}$, directed along x, is

$$\ddot{x} - \gamma\dot{x} + \omega_0^2 x = \frac{e}{m} E_0 e^{-i\omega t}.$$

The scattered wave has the same frequency as that of the incident wave, and a solution of this second-order differential equation, the time dependence of which is given by $e^{-i\omega t}$, corresponds to an acceleration,

$$\ddot{x} = \frac{e}{m}\left(\frac{-\omega^2}{\omega_0^2 - \omega^2 - i\omega\gamma}\right) E.$$

To find the corresponding radiation fields one simply substitutes $\ddot{x}$ in the formulas for radiation from an accelerated charge. The magnitude of the scattered electric field at a distance r from the electron is

$$E(r) = \frac{e\ddot{x}\sin\alpha}{4\pi\epsilon_0 rc^2},$$

in which α is the angle between the acceleration (or the electric vector of the incident field) and the direction of propagation of the scattered radiation; the same function accounts for the fact that no radiation is scattered in the direction of the incident electric field.

It is convenient to consider separately the various approximations to the complete scattering. For a free or weakly bound electron in which its damping force is negligible and its binding frequency is vanishingly small compared to the frequency of the incident, radiation ($\omega \gg \omega_0$), then $\ddot{x} = \frac{e}{m}E$ from the above equation, and the total radiation rate is given by substitution in the Larmor formula,

$$\frac{e^2\ddot{x}^2}{4\pi\epsilon_0 c^3} = \frac{e^4 E^2}{4\pi\epsilon_0 m^2 c^3}.$$

If this rate is divided by the intensity of the primary radiation, $I_0 = c\epsilon_0 E^2$, one obtains the effective scattering cross section per electron, the area from which the energy of the incident beam is removed by the scatterer. This is

$$\sigma_0 = \frac{8\pi}{3}\left(\frac{e^2}{4\pi\epsilon_0 mc^2}\right)^2 = \frac{8\pi}{3} r_0^2,$$

in which $r_0 = e^2/4\pi\epsilon_0 mc^2$, and is the classical electron radius. This result was first derived by Sir Joseph John Thomson of England, and classical free-electron scattering is known as Thomson scattering. It it noted that this cross section is a constant, the numerical value of which is $\sigma_0 = 0.665 \times 10^{-24}\ \text{cm}^2$. For electrons that are indeed weakly bound, the Thomson scattering formula represents the scattering of frequencies up to the X-ray region, but deviations become important for higher energy photons. The quantum mechanical calculation of the scattering cross section for free electrons gives the Klein-Nishina formula, which predicts that the scattered radiation is largely concentrated in the forward direction for extremely high frequencies. Moreover, the frequency of the scattered radiation is less than that of the incident radiation by an amount that depends on the scattering angle, as seen in connection with the Compton effect.

Thomson scattering

The general classical formula for the scattering cross section is

$$\sigma = \sigma_0 \frac{\omega^4}{(\omega_0^2 - \omega^2)^2 + (\gamma\omega)^2},$$

in which σ_0 is the Thomson cross section given above.

The cross section is a maximum for $\omega = \omega_0$, the natural frequency of the bound electron, and the resulting resonance scattering may be very large, because γ is small. For strong binding, on the other hand, such that $\omega_0 \gg \omega$, $\sigma = \sigma_0 (\frac{\omega}{\omega_0})^4$; that is, the cross section, and hence the scattering, varies as the inverse fourth power of the incident wavelength. The inverse fourth power law was first derived by Lord Rayleigh to account for the blueness of the sky and the reddening effect of the atmosphere at sunrise or sunset. Rayleigh assumed that the air molecules scattered randomly, so that intensities add, and with average empirical values of the natural frequencies of the atmospheric constituents, the result is in accord with observation. The formula for scattering per unit solid angle, with the incident beam unpolarized, is also in good agreement with the observed partial polarization of sky light.

For wavelengths long in comparison with the dimensions of the scatterer, the scattering cross section is much smaller than the geometrical cross section. The size of particles in atmospheric haze is of the same order as the wavelength of visible light, and infrared photography of a distant landscape is much clearer than that made with ordinary light. Particles with dimensions larger than the wavelength, however, scatter energy with a cross section independent of the frequency, and infrared light does not penetrate fog. The scattering from larger particles, such as water droplets and smoke particles, is called the Tyndall effect and is important in connection with colloidal solutions as well as in the transmission of light through atmospheric haze.

Tyndall effect

ELECTROMAGNETIC WAVES IN MATERIAL MEDIA

Matter is composed of charged particles that are affected by the application of an external electric field. Dielectrics differ from conductors in that they have no free charges that can move through the material, but there is a displacement of positive and negative charges relative to each other under the influence of an applied field. This charge separation, ordinarily small compared with atomic dimensions, is called polarization. Any configuration of induced bound charge is capable of producing electric fields, just as if they were free charges; but because the existence of any induced charge depends on the field itself, it is convenient to introduce another field quantity the sources of which are the free charges only. In terms of **P**, the dipole moment per unit volume, also called simply the polarization, this displacement field $\boldsymbol{D}$ is defined by the relation

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}.$$

In many dielectrics the polarization of the medium is proportional to the strength of the electric field, so that the displacement field is also proportional to it; that is,

$$\boldsymbol{D} = K\epsilon_0 \boldsymbol{E},$$

in which K is the dielectric constant of the medium, sometimes called the permittivity. This relation is not a principle of electromagnetism but a description of a limited class of matter, and it is called a constitutive equation. An analogous constitutive equation relates $\boldsymbol{B}$ to $\boldsymbol{H}$, but because in dielectrics and almost all conductors an applied magnetic field produces very little change in the charge configuration of the bound charge, one may assume that $\boldsymbol{B} = \mu_0 \boldsymbol{H}$, just as in empty space.

In the absence of free charges and currents, a combination of Maxwell's equations to eliminate the magnetic field leads to

$$\nabla^2 \boldsymbol{E} - \frac{K}{c^2} \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = 0.$$

Exactly the same equation is obtained for the magnetic field $\boldsymbol{B}$. This is the ordinary wave equation, but the wave velocity (c') is now $c' = c/\sqrt{K}$. From the definition of the optical index of refraction, one obtains the Maxwell relation $n = c/c' = \sqrt{K}$, according to which the dielectric constant (K) should be at least approximately equal to the square of the index of refraction n. This relation is confirmed for long waves, and in some substances for the visible spectrum, but for many dielectrics it fails badly. Water is a notable example: the static dielectric constant is about 81, whereas the index of refraction is very nearly 4/3. The reason for these failures is that the polarization depends on the frequency; so that the transmission of electromagnetic waves of high frequency in dielectrics must be treated as a dynamic process, just as in conductors.

One should note that in a dielectric the vectors $\boldsymbol{E}$ and $\boldsymbol{B}$ are in phase and that the electric and magnetic energy densities are equal, just as in free space, although the instantaneous energy density is now $\epsilon_0 K E^2$. The average value of the Poynting vector is the wave velocity c' times the average total energy density.

All media absorb electromagnetic radiation to some extent. The simplest way to account for absorption is to introduce the conductivity σ and to assume that the current density $\boldsymbol{j}$ is proportional to the electric field in accord with Ohm's law: $\boldsymbol{j} = \sigma \boldsymbol{E}$. If the material is nonmagnetic, and if $\boldsymbol{D} = K\epsilon_0 \boldsymbol{E}$, the Maxwell equation corresponding to Ampère's law becomes

$$\nabla \times \boldsymbol{B} = \mu_0 \sigma \boldsymbol{E} + \frac{K}{c^2} \frac{\partial \boldsymbol{E}}{\partial t}.$$

Elimination of B between this equation and the Faraday induction equation then leads to the equation of telegraphy,

Equation of telegraphy

$$\nabla^2 \boldsymbol{E} = \frac{K}{c^2} \frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \mu_0 \sigma \frac{\partial \boldsymbol{E}}{\partial t},$$

and $\boldsymbol{B}$ obeys a similar equation. Because of the dependence of the dielectric constant and the conductivity on the frequency, these equations are useful only for the Fourier components of the field, time variation of which may again be represented by the factor $e^{-i\omega t}$. Then

$$\nabla^2 \boldsymbol{E} + \left(K + \frac{i\sigma}{\epsilon_0 \omega}\right) \frac{\omega^2}{c^2} \boldsymbol{E} = 0,$$

in which the effective permittivity $\eta = K + i\sigma/\epsilon_0\omega$ takes the place of the dielectric constant in the wave equation. Because η is complex (has an imaginary part) there will be both absorption and a phase displacement between the electric and magnetic field vectors. There is correspondingly a complex index of refraction, which may be written $\sqrt{\eta} = n + i\kappa$.

The field variation of a wave traversing the medium in the x direction now depends on the exponential factor

$$\exp\left[-i\left(\omega t + \frac{nx}{c}\right)\right] \exp\left(-\omega \kappa x/c\right),$$

and κ is called the extinction coefficient.

The relative importance of the two terms in the differential equation for the electric field may be expressed in terms of a relaxation time $\tau = K\epsilon_0/\sigma_0$, in which σ_0 is the direct current conductivity. If this relaxation time is long compared with the period $2\pi/\omega$ of the radiation, the wave is propagated without much loss. For pure metals, however, the relaxation time is of the order of 10^{-14} second. Hence, for frequencies lower than those of the optical spectrum, an electromagnetic wave is essentially not propagated within a highly conducting medium. The extent to which the fields do penetrate the conductor depends on the frequency, and the "skin depth" is found by solving the equation with the assumption that the conductivity is extremely large compared to the product of the dielectric constant of free space and the frequency of the field; *i.e.*, $\sigma \gg \epsilon_0 \omega$. The square of the wave number k is then $k^2 = i\omega\sigma/\epsilon_0 c^2$; so that $k = (1 + i)/\delta$, in which $\delta = c\sqrt{2\epsilon_0/\sigma\omega}$ is the "skin depth," the depth to which the fields at normal incidence would decrease to $1/e$ (the logarithmic base $e = 2.7183$) of their values. Within the metal the electric and magnetic fields are no longer in phase, and the energy density is no longer shared equally by the fields but is largely magnetic.

For still higher frequencies a metallic conductor behaves like an ionized gas, in that the frictional forces corresponding to Ohm's law may be much less than the dynamical effect of the electric field on the free electrons. If there are N electrons per unit volume, each with mass m and charge e, the current density j is Nev, and the equation of motion $mv = eE$ may be transformed into

$$\frac{dj}{dt} = \frac{Ne^2E}{m}.$$

Formally, one may put $j = \sigma E$, and with the time dependence given by $e^{-i\omega t}$, it is found that $\sigma = iNe^2/m\omega$. In general it is necessary to add the displacement current $D = iK\epsilon_0 E$; so that

$$j_{\text{total}} = i\epsilon_0\omega(K - \frac{Ne^2}{m\omega^2\epsilon_0})E.$$

Clearly, there is a value of the frequency for which the total current vanishes, namely,

$$\omega_p = \sqrt{\frac{Ne^2}{m\epsilon_0 K}},$$

which is called the plasma frequency. Because the dielectric constant K is not far from unity, this frequency depends only on the density of free electrons and is given by $\nu_p = 8.98N^{1/2}$ cycles per second. For the ionosphere, ν_p is about three megacycles per second, whereas for copper it is 2.5×10^{15} cycles per second. At frequencies higher than the plasma frequency the displacement current is dominant, and therefore the conductor behaves as an insulator for such frequencies.

Plasma frequency

ANTENNAS

Electric dipole radiation is produced by any charge distribution that has a dipole moment that varies harmonically with the time. For application to real antennas, the formulas for radiation fields from an infinitesimal dipole given earlier must be integrated over the finite dimension of the radiating system. Short linear antennas, having length much smaller than the wavelength of the emitted radiation, require only the substitution of the current for the time variation of the charge. If the current I is written as $I_0e^{-i\omega t}$, and l is the length of the dipole, then $\ddot{p} = -i\omega lI_0e^{-i\omega t}$, and the time average of the square of $\ddot{p}$ is $\frac{1}{2}I_0^2l^2\omega^2$. The angular distribution of radiation from a short linear antenna is the same as for an oscillating linear dipole, given by the factor $\sin^2\theta$, with the angle θ measured from the dipole axis, so that no power is radiated in the direction of the axis. The time average of total radiated power ($\bar{P}$) is

$$\bar{P} = \frac{\pi I_0^2}{3\epsilon_0 c}\left(\frac{l}{\lambda}\right)^2,$$

in which $\lambda = 2\pi c/\omega$ is the free-space wavelength of the radiation. The power supplied to maintain this radiation is often expressed in terms of the radiation resistance, defined by analogy with the coefficient of the root-mean-square current, which gives the rate of joule (electrical) heating in a conductor, $\frac{1}{2}RI_0^2$. The radiation resistance is

$$\frac{\pi}{6\epsilon_0 c}\left(\frac{l}{\lambda}\right)^2 = \frac{\mu_0 c\pi}{6}\left(\frac{l}{\lambda}\right)^2 = 789\left(\frac{l}{\lambda}\right)^2 \text{ ohms}$$

for a short antenna in which the current is essentially constant over the length. This radiation resistance represents the power lost as radiation in an antenna supplied with a current $I_0e^{-i\omega t}$.

A somewhat more realistic antenna has dimensions that are not as small in comparison with the wavelength of its radiation. The half-wave antenna has a length $l = \lambda/2$, and is supplied with oscillating power at its centre. The current at any distance x from the centre is given by $I = I_0 \cos(2\pi x/\lambda) \cos \omega t$. Each element Idx of the antenna radiates as a short dipole, and the total fields are found by integrating over the length of the dipole. The angular distribution of the radiation from a half-wave antenna is similar to that from an infinitesimal dipole, but the intensity is somewhat more concentrated in the equatorial plane. The field intensities, and thus the Poynting vector, are independent of the frequency, and the total radiated power is simply proportional to the square of the effective current in the antenna. The radiation resistance of a half-wave antenna is 73.1 ohms. The assumption of the sinusoidal current in the antenna is not entirely justified, because the current distribution would be affected by the radiation losses; but the radiation field configurations are not significantly distorted by these deviations from the idealized half-wave distribution.

Antenna radiation pattern

The radiation patterns from linear antennas more than a wavelength long are more severely affected by the losses involved in the radiation. Linear antennas the lengths of which are an integral number of half wavelengths give rise to simple idealized patterns, symmetric about the axis of antenna, but with a number of lobes equal to $2l/\lambda$, the number of half wavelengths in the antenna. Ideally, the most intense lobe is one that makes the smallest angle with the antenna, but the decrease in current density along the wire produced by losses tends to reduce the intensity of these strong lobes relative to the weaker lobes near the equatorial plane. To radiate energy predominantly in some particular direction, it is more advantageous to use antenna arrays rather than long antennas.

The radiation patterns produced by arrays of dipoles are sensitive both to the separation of the dipoles in terms of the radiated wavelength and to the relative phase with which the antennas are driven, as well as to the length of each dipole. Simple arrays consist of half-wave antennas equally spaced in a direction perpendicular to their length. If the antennas are driven in phase, the radiation is emitted mainly at right angles to the orientation of the array—*i.e.*, broadside radiation. Antennas spaced a half wavelength apart and driven in alternating phases give fields that add constructively in the direction of the array, giving end-fire radiation.

Broadside and end-fire radiators

In engineering practice, antenna arrays have been designed to give a wide variety of radiation patterns. The arrays may be two- or three-dimensional; and by increasing the number of elements and decreasing their spacing, the limiting case of a continuous distribution is obtained. Particularly in the microwave region the energy is radiated by horns or by parabalidal reflectors. The directional characteristics of the radiation may be obtained by integrating over the fields in the radiating aperture because according to Huygens' principle each point on a wave front behaves like a secondary source. The whole process is very like that of describing diffraction effects in optics, except that in optics it is often possible to substitute a scalar field and neglect the orientation of the electric and magnetic vectors.

It is unnecessary to make a detailed separate analysis of receiving antennas to understand their operation. The transmitting antenna sets up a field that the receiving antenna must transform into a signal by means of a suitable detector. The receiving pattern of an antenna is identical with its transmitting patterns, although the receiver must be oriented properly for maximum reception of any particular signal. Maximum detected signal depends only on the direction from which the incident wave comes, but the effectiveness of antenna reception as a function of angular variables is the same as its gain as a transmitter in terms of angular variables with respect to its orientation.

In practice, neither transmitting nor receiving antennas are in free space, and the effects of both ground and the ionosphere must be taken into account. For certain frequencies the atmosphere itself affects the radiation. Depending on the frequency, the ground may be treated as a conductor or as a dielectric with a complex index of refraction.

BIBLIOGRAPHY. For an account of the historical development, see SIR EDMUND WHITTAKER, *A History of the Theories of Aether and Electricity from the Age of Descartes to the Close of the Nineteenth Century*, 2 vol. (1951–53, reprinted

1960). Extensive treatments of visible radiation are given by MAX BORN and EMIL WOLF, *Principles of Optics*, 4th rev. ed. (1970); and R.W. DITCHBURN, *Light*, 2nd ed. (1963). Classical radiation and electron theory are treated in RICHARD BECKER and F. SAUTER, *Electromagnetic Fields and Interactions* (1964); D.J. JACKSON, *Classical Electrodynamics* (1962); JERRY B. MARION, *Classical Electromagnetic Radiation* (1965); and W.K.H. PANOFSKY and MELBA PHILLIPS, *Classical Electricity and Magnetism*, 2nd ed. (1962). See H.H. SKILLING, *Fundamentals of Electric Waves*, 2nd ed. (1948), for a simple account of wave guides. A thorough introduction to the quantum theory is given in W. HEITLER, *The Quantum Theory of Radiation*, 3rd ed. (1964).

(M.Ph.)

Electron

The electron is the lightest stable elementary particle of matter known. Carrying a single negative charge, considered to be the basic charge of electricity, it is a constituent of ordinary matter and occurs free in space. For fundamental properties of the electron see below.

The antiparticle of the electron, the positron, has precisely identical properties except that it carries a positive charge. If it meets an electron both are annihilated in a burst of energy. Because electrons abound everywhere, positrons are rare on earth, being produced only in high energy processes (*e.g.*, by cosmic rays) and live only for brief intervals of time before annihilation by electrons.

Under ordinary conditions, electrons are bound to the massive, positively charged nuclei of atoms by the attraction between opposite electric charges, one of the four forces that affect elementary particles. (Particles with the same charge, such as electrons, repel one another.) In neutral (not electrically charged) atoms the number of electrons equals the number of positive charges on the nucleus of the atom, but any atom may have more or fewer electrons than positive charges and thus be negatively or positively charged as a whole; such charged atoms are called ions.

Within each atom and ion, the electrons are not at rest but seem to move in complicated orbits around the nucleus that holds them against their tendency to repel one another and fly apart; and thus orbiting they build a cloud of negative charge that occupies almost the whole volume of the atom, the positive nucleus being a tiny, heavy mass at the heart of this relatively weightless negative cloud. The size of an atom, and its response to other atoms, particles, light, and other electromagnetic radiation, such as X-rays, is thus determined by the arrangement of the orbiting electrons. Atoms combine to form larger structures, such as molecules. Since the discovery of the electron, the concept that atoms will combine with each other by transferring or sharing electrons has dominated chemistry in terms of the modern valence theory, based on atomic structure, and of the chemical bonding theory, based on the sharing of electrons by adjacent atoms. The attraction between electrons attached to one atomic nucleus and the nuclei of nearby atoms (called Van der Waals force) explains much of the mechanism of the formation of liquids and solids. In nonconducting solids, the electrons are permanently positioned among the nuclei. In semiconductors, metals, and electrolytes, the electrons can move through the whole sample; and if they are made to move in one direction by any of several methods, they create an electrical current.

Free electrons

Electrons may also occur free, not associated with atoms, in a gas or vacuum, and in such cases they can then be described, in the language of classical mechanics, as particles of matter. On the other hand, electrons in atoms and molecules, and in solids, have wavelike properties that require quantum mechanics for explanation. The quantum theory of the interaction of electrons with radiation and with other particles is now considered complete, although solutions of the equations for many-electron systems are still imperfect. In future applications of knowledge about the electron, considerable technical development may be expected in the control and use of electrons in solid-state and vacuum devices.

PHYSICS OF THE ELECTRON

Electrons in early theories of electricity. Although the electron as a concept was not developed until the end of the 19th century, earlier theories foreshadowed aspects of its nature. In the 18th century electricity came to be thought of as an elastic (*i.e.*, self-repulsive) fluid that, according to Benjamin Franklin (1747), consisted of "particles extremely subtile, since it can permeate common matter, even the densest metals, with such ease and freedom as not to receive any perceptible resistance." The characterization of vitreous and resinous electricity as two different fluids, positive and negative respectively (R. Symmer 1759), was made fashionable by C.A. Coulomb (1785), although this conjecture could not then be verified. By 1812 J.J. Berzelius was speculating that "the electricities" might be "matter, destitute of gravitation, but possessing affinity to gravitating bodies," to explain the disruptive effects of an electric discharge.

The atomicity of electricity was first strongly suggested by Michael Faraday's discovery (1833) that the mass of a chemical element liberated by electrolysis of a solution containing a compound of that element was inversely proportional to its chemical equivalent (*i.e.*, valency), and exactly proportional to the total electric charge that had passed through the solution: "there is a certain absolute quantity of the electric power associated with each atom of matter." It was natural, thereafter, to interpret the bond between atoms in a molecule in terms of the electrostatic force between charged atoms, or ions. Yet W. Weber's electrodynamics (1846), based on inverse-square forces at a distance between electric particles, was not widely favoured. For the next generation, speculation on the nature of electricity was dominated by the electromagnetic theory of James Clerk Maxwell (1864), who grounded his mathematical formulas on the notion of strains in a continuous ether, a theoretical substance then believed to fill the universe. As late as 1881, in his *Treatise on Electricity & Magnetism* (2nd ed.), Maxwell used the hypothesis of a "molecule of electricity" to explain electrolysis but remarked that it is "out of harmony with the rest of this treatise."

Electrons in cathode rays. In the 1880s interest was aroused in cathode rays, one of many mysterious phenomena observed in electrical discharges through gases at low pressures. The ray was an invisible radiation emanating from the negative electrode of a cathode tube and travelling in straight lines to produce a phosphorescent glow where it struck a glass envelope. Although the rays were deflected by magnetic fields, it was controversial whether these rays were vibrations of the ether or streams of charged particles. This question was at length resolved (1894–97) by the experiments of J.J. Thomson, who showed, using an apparatus (Figure 1) that was the prototype of the modern cathode-ray tube, that the rays travelled more slowly than light, transported negative electricity, and could be deflected by an electric field. The rays produced at the cathode C were defined as a beam by the slits A, B, and passed between the electrified plates D, E (or between the poles of a magnet), before striking the screen where the deflection could be measured. From the formula of H.A. Lorentz (1892)

First measurements of the electron

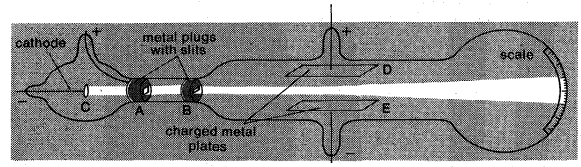


Figure 1: Device for detecting deflection of cathode rays by an electrostatic force acting between two charged plates (see text).

$$\dot{\mathbf{v}} = \frac{e}{m}\boldsymbol{E} + \frac{e}{mc}(\mathbf{v} \times \boldsymbol{H})$$

for the acceleration of a charged particle in the electric field $\boldsymbol{E}$ and magnetic field H, it was then easy to calculate by classical mechanics the velocity v and the mass/charge

ratio of the particle, m/e. This was found to be much smaller than for canal rays; *i.e.*, positive ions. This result meant that, since the unit negative charge exactly equalled the unit positive charge, the mass of the particle carrying the negative charge was considerably less than the mass of the particle carrying the positive charge. By measuring the mobility of ions (*i.e.*, any charged particle) in gases (or estimating the number on which cloud droplets had been nucleated), Thomson and his pupils demonstrated that the charge e was about the same as the charge carried by the hydrogen ion in electrolysis. By 1900 there was overwhelming evidence that electricity consisted of a flow of electric charges, in units of magnitude about 1×10^{-19} coulomb (this was the unit for which G.J. Stoney in 1894 had proposed the name electron) and that the mass of this corpuscle of negative charge was only about 1/2,000 the mass of a hydrogen atom.

Properties of the classical electron. Thomson's electrons were very soon identified as the cause of other natural phenomena. The electrical conductivity of air near a hot poker, observed in the early 18th century, was found to be due to thermionic emission of electrons from a hot metallic surface. Similarly, the photoemission of electrons from a metal plate illuminated with ultraviolet light was shown to be the cause of the photoelectric effect (dissociation of atoms into electrically charged particles resulting from the absorption of electromagnetic radiation). The beta rays emitted by many radioactive substances were found to be energetic electrons. Electrons were observed so universally in nature that any ionizing particle of obvious small mass or negative charge was taken to be an electron, an empirical generalization that was not proved false until the 1930s by the discovery of the elementary particles positrons and mesons in cosmic rays.

Measurement of negative charge

The magnitude of the negative charge e was obviously of basic importance and a scale parameter for the whole of atomic physics. Earlier estimates were superseded by R.A. Millikan (1916), who transformed the cloud droplet method into a precision technique. Through a microscope he timed the fall of a tiny drop of oil, whose size could then be deduced by Stokes' law for the motion of a sphere through a viscous fluid. He then switched on a vertical electric field: any charge on the drop could be calculated from the new rate of fall. From time to time this charge appeared to change abruptly by a constant amount, showing that the drop had captured, or lost, an electron. This direct observation of an atomic event seemed to give the value of e to within 0.1 percent.

Millikan's result was actually too low by about 0.6 percent. The correction came 20 years later by a roundabout route. Comparison of X-ray diffraction from crystals yielded absolute values of the spacings between their atomic planes. Avogadro's number could be calculated more correctly, and new measurements of Faraday's electrochemical constant then gave e to one part in 10^4. By that date similar precision could be obtained in measurements of e/m by a variety of techniques. But the values of e and m quoted below do not depend on any single experiment. These, and other fundamental constants, are now statistically adjusted for consistency with the observed values of the many different physical quantities in whose theoretical formulas they appear.

Electrons in the atom. By 1900, physicists knew that electrons came from inside atoms. What were they doing there? The sharp lines in optical spectra suggested that the electrons could vibrate in various ways with well-defined frequencies. Attempts to explain this action with stable electrostatic models—for example, with a model in which negative electrons are embedded like currants in a bun of positive charge—did not provide the right answers. Then came the alpha-particle-scattering experiments of E. Rutherford (1911), which showed that all the positive charge within an atom must be concentrated in a small, heavy nucleus, leaving all the electrons grouped somehow outside the nucleus. To this model N. Bohr (1913) added ideas from the quantum theory of M. Planck (1900). Each electron is supposed to move rapidly in a circular or elliptical orbit around the nucleus due to electrostatic attraction between it and the nucleus (Figure 2A). But only those orbits are permitted for which the angular momentum is an integral multiple of $\hbar$ (*i.e.* of Planck's constant h divided by 2π); light is emitted or absorbed as the electron acquires or loses energy and jumps from one orbit to another.

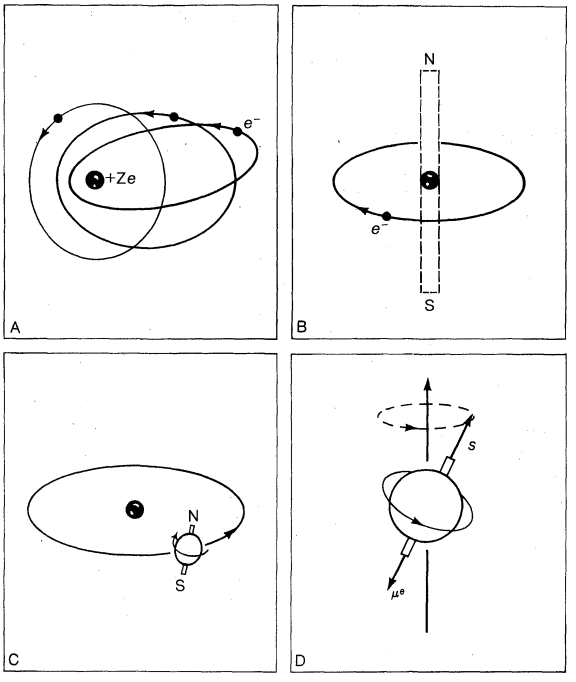


Figure 2: Orbital motion and spin of an electron in an atom, without regard to quantum mechanical concepts (see text).

The Bohr-Rutherford model

The Bohr-Rutherford model is the best classical picture of the electrons in an atom. In particular, it emphasizes their high kinetic energy. Indeed, the electron in the first Bohr orbit is travelling at an appreciable fraction of the velocity of light, c, a dimensionless number called the fine structure constant. It is given by the equation $\alpha = e^2/c\hbar = 1/137.0388(6)$. One of the successes of the model was A. Sommerfeld's prediction (1916) of fine structure in the hydrogen spectrum as a result of relativistic corrections to Newtonian planetary theory.

Electron magnetism and spin. A charged particle moving round and round a closed orbit is equivalent to a loop of electric current that behaves like a small magnet (Figure 2B). Each electron in an atom must therefore give rise to magnetic moment exactly proportional to its orbital angular momentum, which is quantized to the values $l\hbar$ ($l = 0, 1, 2 \ldots$). This contribution is easily calculated as an integral multiple, $l\mu_B$, of the Bohr magneton $\mu_B = e\hbar/2mc$. In an atom containing several electrons with orbits in various orientations, these contributions largely cancel out but could explain the existence of stable magnetized atoms such as iron. In an external magnetic field, H, the axis of each orbit would precess with the Larmor frequency $\omega_L = eH/2mc$. This frequency may combine with the frequency of an optical transition, giving rise to a Zeeman effect: the splitting of spectral lines in the radiation emitted by atoms in a magnetic field. Unfortunately for the simple Bohr theory, the Zeeman effect in many atoms was obstinately anomalous. As an empirical explanation, G.E. Uhlenbeck and S. Goudsmit (1925) conjectured that the electron might have an intrinsic angular momentum, or spin (Figure 2C), with an associated magnetic moment. Each electron is assigned an additional angular momentum, $s\hbar$, whose spin quantum number, s, takes only two values, $s = 1/2$ or $s = -1/2$. But the intrinsic magnetic moment is still of magnitude μ_B. If this is expressed as $\mu_e = g\mu_B s$ with $g = 2$, it is noted that the gyromagnetic ratio (*i.e.*, magnetic moment/an-

gular momentum) is twice as large for electron spin as for orbital motion.

This conjecture is confirmed by a variety of observations. Complicated atomic spectra are successfully calculated by the vectorial addition of orbital and spin moments. In most magnetic materials, the observed gyromagnetic ratio shows that the spin magnetic moment is dominant. A whole field of physics and chemistry deals with electron spin resonance, in which the Larmor precession frequency, $g\omega_L$ for a free electron spin in a magnetic field (Figure 2D), is detected directly by resonant interaction with a microwave field of the same frequency.

Electron radius

It is obviously tempting to picture an electron as a spinning sphere of electric charge whose radius is determined by the dimensional relation $e^2/a = mc^2$ at which the electrostatic self-energy of the charge distribution is comparable with the relativistic energy of the rest mass. This classical electron radius, $a = 2.81785 \times 10^{-15}$m, is an important scale parameter in physics; but the uniqueness of e, the arbitrariness of the quantization rules, and the difficulty of making it properly relativistic, forbid such a purely classical model.

The fundamental constants of the electron are as follows:

Mass: $m = 9.10908(13) \times 10^{-31}$kg
$= 5.48597(3) \times 10^{-4}$atomic mass units
Charge: $e = -1.60210(2) \times 10^{-19}$ coulomb
$= -1$ atomic charge
Spin: $s = \frac{1}{2}$
Magnetic moment μ_e
$= 9.2839(2) \times 10^{-24}$ joule per tesla
$= 1.001159622(27)$ Bohr magnetons

Electrons as waves. Electron theory was revolutionized in 1923 by L. de Broglie's hypothesis that an electron moving with velocity v might be accompanied by a periodic disturbance of wavelength $\lambda = 2\pi\hbar/mv$ (Figure 3A). This analogue of the photon concept (the unit of

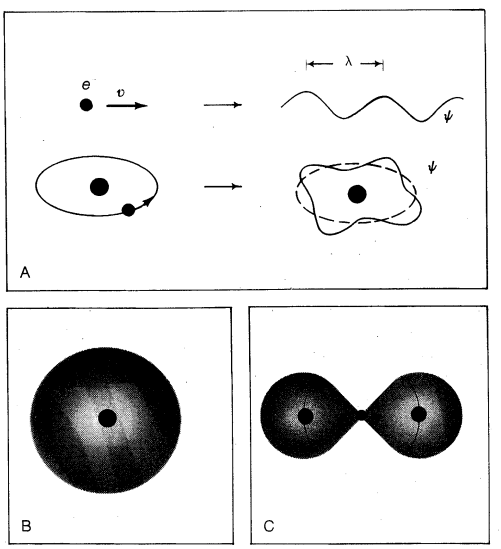


Figure 3: Cloud or probability concept of the electron according to quantum mechanics (see text).

electromagnetic energy that has no mass and no charge) was not, in fact, confirmed by direct observation (diffraction from crystal surfaces) until 1927, but the Bohr quantization rule could now be interpreted as the condition for a standing wave, with a whole number of wavelengths around the circumference of the orbit. In 1926, E. Schrödinger wrote down the partial differential equation for such a wave function (conventionally symbolized by the letter ψ) and showed that it had stationary solutions in the electrostatic field of a nucleus at exactly the energies predicted by the Bohr model. This equation is the basis of all subsequent theory for electrons at the atomic scale.

Of what is ψ a wave? At first it looks as if the electron charge has been spread out into a continuous cloud through the whole volume of the atom (Figure 3B). In fact, the value of $[\psi]^2$ at any point represents the density of probability of finding the electron, still a tiny indivisible particle, in that region. The uncertainty principle (M. Born; W. Heisenberg, 1927) permits only statistical notions of the whereabouts and velocity of the electron as it traverses the Bohr orbit, which thus becomes blurred into an orbital function. But the indefinite stability of such an orbital in time allows it to have a sharply defined energy, as required by spectroscopy.

Many-electron systems. As the number of positive charges on the nucleus increases with increasing atomic weight, the number of electrons increases equally. The arrangement of electrons around the nucleus follows a pattern that can be conceived as a balanced distribution among concentric shells. Each shell can accommodate up to a fixed number of electrons, and this pattern is expressed by the periodic law. The periodic law, therefore, can be interpreted as the filling up of successive closed shells of Bohr orbits, as electrons are added to neutralize increased nuclear charge. Strictly speaking, these wave mechanical electrons, being indistinguishable from one another, cannot be assigned individually to different orbitals in the same atom. But a fundamental postulate, that the many-electron wave function must be totally antisymmetric for the interchange of any two sets of particle coordinates, leads in simple cases to the exclusion principle of W. Pauli (1924): each quantum state can be occupied by only one electron at a time. For example, a given orbital function may not contain more than two electrons, which must then have opposite spins (Figure 3C).

Exclusion principle

Electron assemblies therefore obey Fermi-Dirac statistics. Whereas the molecules in a classical gas at low temperatures would all be crowded into the lowest state, the particles in a dense electron gas are forced to occupy a whole range of states up to the Fermi energy, E_F (Figure 4A). The average kinetic energy per electron remains as

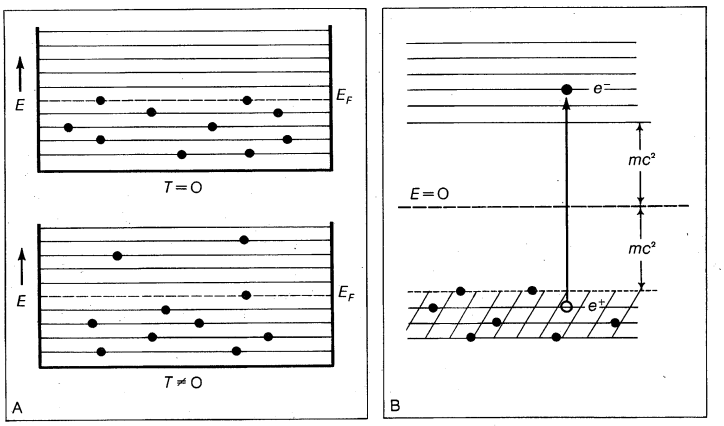


Figure 4: *Energy-level diagrams for the electron and positron.*
(A) Distribution of electrons among energy states: (top) at very low temperatures, (bottom) at higher temperatures.
(B) Formation of a positron by ejection of an electron.

large as if the gas were at an enormous degeneracy temperature, $T_F = E_F/k$ (k = Boltzmann's constant). When the assembly is heated, only a few electrons at the top of the distribution can take up energy by moving into empty higher levels, so that the specific heat is small. This phenomenon is extremely important in metals and in the interior of stars.

Even in a Bohr atom each electron would markedly perturb the orbits of others by mutual electrostatic repulsion. As a first approximation each orbital may be calculated as if the particle were moving in the average self-consistent field (D.R. Hartree, 1928) of its fellows. But many-electron wave functions are really much more complicated than antisymmetric combinations of quasi-independent orbital functions. Significant exchange and correlation factors which describe statistically the way in which electrons avoid one another in accidental encoun-

ters need to be included. The mathematical theory of these terms and of the physical phenomena to which they give rise is still incomplete.

Relativistic electrons and positrons. The unit of energy on the subatomic scale is the electron volt, the energy change of a charge moved through a potential difference of one volt. The relativistic rest mass, mc^2, of an electron is only 0.511×10^6 electron volts (*i.e.*, 0.511 Mev), which is easily exceeded in cosmic rays and in particle accelerators. What happens to the electron then?

As its energy increases, a point particle merely acquires additional mass, but a spinning sphere cannot easily be discussed relativistically. In nonrelativistic quantum theory one needs two wave functions, $\psi\uparrow$ and $\psi\downarrow$, to describe the electron in each of its spin states, together with rules for combining these functions as the spin axis changes direction. How do these functions behave at relativistic velocities? In a classic paper, P.A.M. Dirac (1928) showed mathematically that the wave function must have two additional components: $\bar{\psi}\uparrow$ and $\bar{\psi}\downarrow$. An electron in the state $\bar{\psi}\uparrow$ when accelerated to high energies acquires a proportion of the $\bar{\psi}\uparrow$ function, and so on. But a pure $\bar{\psi}\uparrow$ function seems to describe an electron in a state of negative kinetic energy, a new physical concept. Dirac suggested that these states are not normally observed because in a vacuum these states are all completely filled (rather like the states of an ordinary electron gas at absolute zero temperature). But given sufficient energy (*i.e.*, more than $2\ mc^2$) an electron could be knocked out of a negative energy state to leave a "hole" in the background continuum. Such a hole, being the negation of a particle of negative charge and negative mass, behaves electrically and dynamically exactly like an ordinary particle of positive charge (Figure 4B). A positron, as this hole came to be called, is thus said to be the antiparticle of an electron.

The positron

Without knowing of Dirac's theory, C.D. Anderson (1932), P.M.S. Blackett, and G.P.S. Occhialini (1933) reported the observation of just such particles in cosmic radiation where they could be distinguished from electrons by the opposite curvature of their cloud-chamber tracks in a magnetic field (Figure 5). When gamma rays (*i.e.*, high-energy photons) interact with matter, they can produce electron-positron pairs by exciting electrons out of the negative energy states of the vacuum. Each of these particles, in turn, can emit more photons, thus cascading into a cosmic-ray shower. The positron, however, has a short life in ordinary matter, for when it has been slowed down it meets an electron that drops into the vacant hole with the emission of another gamma ray; *i.e.*, the two particles mutually annihilate one another. Positrons are also emitted from certain artificially radioactive nuclei in the same manner as are beta rays.

Quantum electrodynamics. The motion of a charged body creates an electromagnetic field that reacts back to resist the motion. For the classical electron, a finite radius must be assumed if this self-energy correction is not to become infinite. Somewhat similar mathematical divergences were soon revealed in quantum theory. A moving electron polarizes the vacuum in its immediate vicinity: virtual photons and virtual electron-positron pairs are excited and then almost immediately destroyed, giving rise to infinite corrections, δm and δe, to the assumed mass and charge of the particle.

This difficulty again can be avoided by assigning to the electron a finite radius, but this arbitrary hypothesis cannot be made relativistically self-consistent. A partial solution was found, about 1950, in the renormalization scheme. Since the bare electron unaccompanied by virtual excitations can never be observed, δm and δe are treated as the physical mass and charge, disregarding the fact that the assumed parameters of the hypothetical bare particle would be infinitesimally small. Although obviously unsatisfactory as a fundamental physical theory, this procedure is relativistically respectable and correctly predicts several observed phenomena. For example, the vacuum close to a charged nucleus contains virtual photons that interact with an electron passing that way,

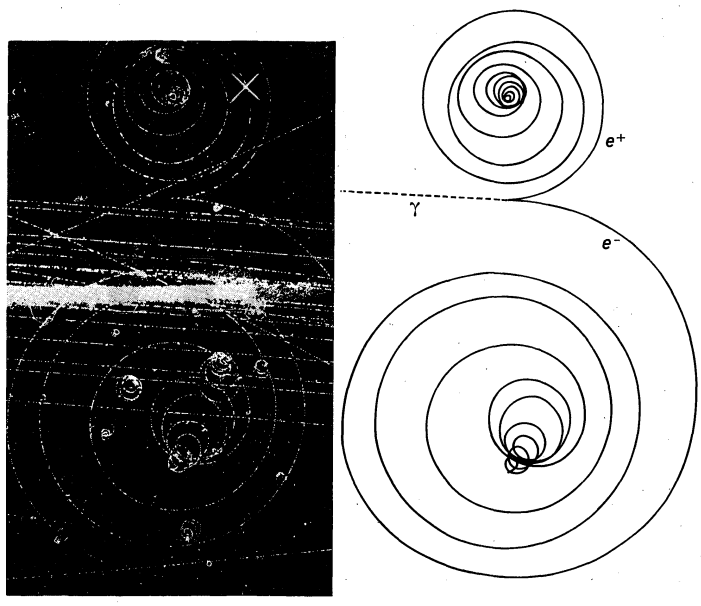


Figure 5: Cloud-chamber study of the spiral paths taken by an electron and a positron resulting from the passage of a gamma-ray photon through a sheet of lead. (The invisible path of the gamma ray is indicated with a dotted line in the diagram of the reaction on the left.) A magnetic field produces paths that are curved in a direction according to the positive or negative charge of the particle.
By courtesy of Brookhaven National Laboratory, Upton, Long Island, N.Y.

slightly changing the electron's energy. (This explains the Lamb shift [W.E. Lamb 1947] of about 1 GHz (gigahertz : 10^9 hertz) in some lines of the spectrum of atomic hydrogen.) The accompanying cloud of virtual excitations also contributes to the anomalous magnetic moment of the electron, whose g-factor is found to be slightly larger than the integral value 2 predicted by the basic Dirac theory.

Electrons in the nucleus. Since electrons are emitted as beta radiation in some radioactive transitions, it was assumed by Rutherford that they were present within the nucleus, neutralizing some of the protons making up the total mass of the nucleus. This is inconsistent with quantum mechanics because, by the uncertainty principle, an electron confined to so small a volume would have much too much kinetic energy to be bound. With the discovery of the neutron and of the fact that a nucleus consisted of a mixture of protons and neutrons, it became clear that beta decay was the spontaneous transformation of one of the neutrons in the nucleus into a proton and an electron, which could then escape from the nucleus. Similarly, a proton could emit a positron and change into a neutron, an event that also produces a more stable nucleus.

Beta decay

Beta decay is peculiar in that the energy of the emitted electron is not always the same but is spread over a wide range (Figure 6A). W. Pauli (1931) and E. Fermi (1934) explained this as the simultaneous production of another particle, the neutrino, which can carry away the rest of the energy but is undetected because it has zero charge and zero rest mass (Figure 6B). A further feature of this process was not suspected until 1956 when T.D. Lee and C.N. Yang showed that parity, one of the symmetry properties of elementary particles is not conserved. Experimentally, this can be observed in beta decay from a magnetically polarized nucleus, where the electrons are emitted preferentially into the hemisphere opposite to the spin vector of the nucleus, whereas if parity were conserved there would be no selectiveness.

The electron as elementary particle. J.J. Thomson's electron was the first to be discovered of a long list of elementary particles. Electrons are classified, along with muons and neutrinos, as leptons, showing that they are not affected by the strong interaction of nuclear forces but suffer the weak interaction that causes beta decay. These particles are all fermions; they obey Fermi-Dirac statistics, have spin $s = 1/2$, and have spinor-wave func-

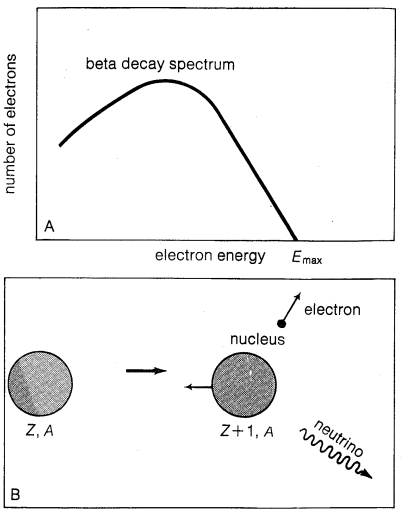


Figure 6: *Beta decay.*
(A) Curve showing how the number of electrons emitted in beta decay vary over a wide range of energies. (B) Schematic diagram showing production of a neutrino.

tions that satisfy the Dirac equation. Each particle has its antiparticle of identical mass but opposite charge. The most puzzling member of this group is the μ-meson or muon, which has all the properties of an electron except that it is much heavier (*i.e.*, $m_\mu = 207\ m$) and lives for only 2×10^{-6} second before decaying, with the emission of a neutrino and an antineutrino, into an ordinary electron. Is the muon, perhaps, merely a highly excited quantum state of the electron? Despite the precision with which the behaviour of these particles can be calculated in electromagnetic fields, up to enormous energies, an understanding of why electricity should occur in packets of exactly this strength, and why the negative particles, in the known universe, should have just this small mass is no nearer than it was for Thomson or Lorentz.

ELECTRONS IN ORDINARY MATTER

Atoms. In a neutral atom of atomic number Z, the Z lowest of all possible orbitals are occupied by electrons. Each orbital is a solution of the Schrödinger wave equation in the self-consistent field dominated by the number of positive charges ($+Ze$) on the nucleus. Each electron has four quantum numbers, denoting the total energy state of the electron; when classified according to their principal and orbital quantum numbers, n and l, the electron states group into shells that, when complete, are spherically symmetric and energetically stable.

Electron shells

The physical and chemical behaviour of an atom is largely, though not altogether, determined by the electrons in orbitals outside the core of closed electron shells. For example, the alkali metals, although widely varying in atomic weight and total number of electrons, have very similar properties because they each have only one electron in the outermost shell. This electron sees the nuclear charge $+Ze$ through the screen of $(Z\text{-}1)$ core electrons (Figure 7); it behaves rather as if it were in the field of a single proton and therefore occupies an orbital similar in energy and dimensions to those in a hydrogen atom, which has a single proton in the nucleus and a single electron. For atoms with two or more such valence electrons, the analysis is somewhat more complicated but easily explains the periodic law of chemistry, the regularities and complexity of atomic spectra, and the familiar fact that all atoms, regardless of atomic number, are roughly of the same size.

Gases. Energy is required to ionize an atom; that is, to remove or add an electron to a neutral atom. An outer electron, with ionization potential of only a few volts, may be ejected by interaction with ultraviolet light; core electrons in heavy atoms need electron impact or X-rays with energies of many kilovolts to be dislodged. In the

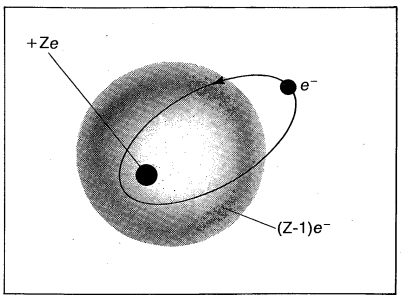


Figure 7: Representation of an atom with one electron outside a core of electron shells (see text).

long run the electron will be recaptured by a positive ion, but this recombination is impeded by dynamical and geometrical factors like those observed when the racked balls are broken in a game of pocket billiards: elastic scattering is easier than capture.

A gas at atmospheric pressure is not normally conductive because of the absence of free ions, but electrons emitted into a rarefied gas (as in a neon, mercury, or sodium vapour lamp) make infrequent collisions with atoms and may thus carry a discharge current. This action is accompanied by an arc or glow from atoms that have been excited into higher quantum states by electron impact. At higher voltages, the gas atoms may themselves be ionized by such impacts, producing more electrons that feed the current. At a critical breakdown voltage, even in a dense gas, the discharge may avalanche into a spark.

A plasma, a neutral gaseous mixture of electrons and positive ions, may thus be produced and maintained by a stabilized electric discharge or by a steady flux of ionizing radiation. Whether artificially produced in an electronic device or naturally occurring on the planetary, solar, or galactic scale, the electrical and magnetic properties of plasmas are due mainly to the electrons, which are vastly more mobile than the positive ions.

Plasma

Electronegative elements (typically the halogens, such as chlorine), which have one vacant orbital in their outermost shells, have an electron affinity of several volts and form stable negative ions. In an ionic solid or salt, such as sodium chloride, the outer electrons of the metal atoms are easily transferred to the halogen atoms (Figure 8A), forming an array of positive and negative ions bound together by their mutual electrostatic attraction (Figure 8B). The electrons all being in stable closed shells, these materials are normally electrical insulators.

But in a transparent alkali halide crystal, an electron at a vacant negative-ion site, or at a substituted metallic ion of higher valency (*e.g.*, a calcium ion [Ca^{++}] for a sodium

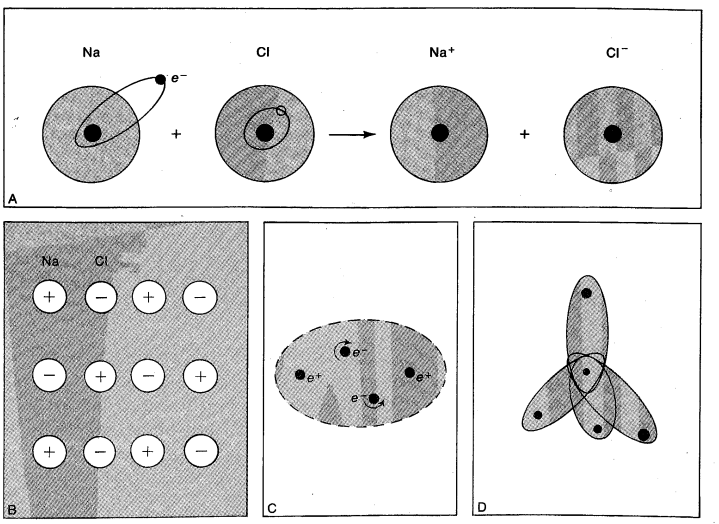


Figure 8: Sharing of electrons between atoms to produce bonding (see text).

ion [Na^+]), is only weakly bound to the local effective positive charge. The system has quantum states analogous to the hydrogen atom levels but much closer spaced in energy, and thus it gives rise to visible optical absorption in a so-called colour centre. A colour centre is readily ionized by a photon or energetic particle, releasing the electron to produce photoconductivity followed by other complex optical effects, such as fluorescence or luminescence, at other such centres.

Molecules. An electron moving through an ionic crystal of a salt is accompanied by a local polarization of the lattice, and this may substantially affect its apparent mass or mobility; such an electron is called a polaron. The behaviour of free electrons in molten salts is not well understood, but when the crystal is dissolved in water, or some other polar liquid, the electrons are taken up into negative ions, such as hydroxyl (OH^-).

Shared electrons

Shared electrons are the binding force in most chemical molecules. Consider two hydrogen atoms brought close together. In terms of reducing their total energy, it pays the electrons to merge their separate atomic orbitals into molecular orbitals that then enclose both nuclei. Two such electrons, by taking opposite spins (Figure 8C), may occupy the same bonding orbital building up a negative charge cloud between the two nuclei and thus binding them together electrostatically. The semiclassical model (G.N. Lewis, 1916) of the chemical bond as a pair of electrons held jointly by the two atoms was given this quantum-mechanical interpretation by W. Heitler and F. London in 1927.

This explanation of the valency rules in covalent compounds is fully justified theoretically and experimentally. Such familiar features as the rigidity of the tetrahedral arrangement of bonds about a carbon atom (*e.g.*, in methane, CH_4) follow immediately from the symmetry conditions on four molecular orbitals shared by a single nucleus (Figure 8D). But the convenient assumption that a pair of electrons is actually localized in each bond is not strictly correct; the true many-electron states are complicated combinations of wave functions extending throughout the molecule and much affected by correlation and exchange. This becomes significant in unsaturated molecules (*e.g.*, the benzene ring) where several, apparently equivalent ways of assigning electrons to bonds become mixed quantum mechanically (resonate) to produce a ground state of much lower energy than any single configuration.

The distribution of electrons in a molecule must take account of the differing electronegativities of the component atoms or subgroups, which may tend to polarize or stabilize the charge cloud or both. In many organic structures the electrons are relatively mobile, being able to

Electron transfer

transfer charge along or around the chains of atoms. Since chemical reactions occur between molecules or radicals with electrons to share or transfer, the rate of reaction depends critically on the electron distribution at the point where the link is made or broken. For example, nitration of an organic molecule occurs preferentially at places where the NO_2^+ radical can easily draw to itself an electron. The importance of electron transfer mechanisms in theoretical chemistry cannot be exaggerated.

Metallic cohesion. The chemical bond concept explains the cohesion of many solids that are no more than indefinitely extended molecules, whether regular crystals as in diamond or disordered networks as in silicate glasses. Since the electrons are all paired in saturated covalent bonds, these materials are basically electrical insulators, although they are easily made semiconducting by thermal excitation.

But an extended three-dimensional framework of directed bonds requires at least four bonds per atom. To make a condensed phase (solid or liquid) of a metal, such as sodium, with only one electron per atom, all the electrons must be set free to be shared by all the atoms. Metallic cohesion may be described as the most economical way of gluing together an assembly of impenetrable positively charged spherical ions with a fairly compressible negatively charged fluid of electrons, using purely electrostatic forces. Since this electron fluid is at a relatively high density, it is statistically degenerate and each particle has high kinetic energy. In the absence of specific quantum-mechanical forces, such as electron affinity for closed atomic shells, the electrons are not bound or localized in space and move rapidly throughout the whole sample. This picture is adequate to explain many of the properties of metals, such as their ability to form alloys over wide ranges of concentration, and many mechanical properties such as ductility and malleability.

Band structure of solids. By 1900, this picture of a metal was widely accepted as an explanation of high electrical conductivity. But a major contradiction, the absence of an electronic contribution to the specific heat, was not resolved until 1928 when A.H. Sommerfeld pointed out that Fermi-Dirac statistics would make this negligible at ordinary temperatures. In the quantum mechanical free electron model, each valence electron is assumed to occupy a distinct, free particle, de Broglie-wave state as if in an empty box bounded by the surface of the specimen (Figure 9A). Because of the Pauli exclu-

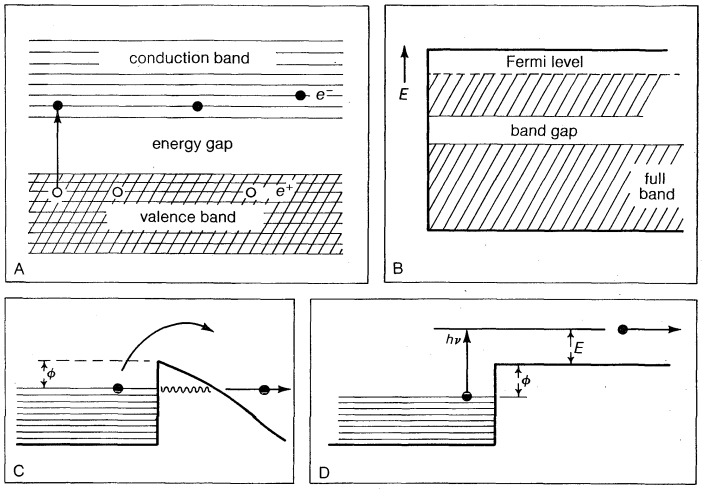


Figure 9: *Energy-level diagrams illustrating the band theory of solids.*
(A) Energy gap in a solid. (B) Transition of electron from valence band to conduction band. (C) Bound electron escaping from a metal. (D) Photoelectric effect.

sion principle, all the states up to the Fermi level are occupied so that only those electrons on the Fermi surface can respond to thermal excitation, or to external electric and magnetic fields. This model gives excellent semi-quantitative explanations of many basic metallic properties (heat conductivity, thermoelectric power, magnetic susceptibility, optical phenomena, etc.) and the answers are little affected by including the electron-electron repulsion in a complete many-electron calculation.

But, surely, the movement of each electron will be strongly resisted by collisions with the metallic ions. F. Bloch (1929) showed that the propagation of a quantum-mechanical electron through a regular crystal is really a problem in diffraction theory. An electron of given energy (*i.e.*, of given de Broglie wavelength) will either be transmitted without loss or it will suffer continual Bragg reflection and be unable to travel at all through the solid. The electron wave functions thus fall into allowed bands, with gaps between them (Figure 9A). The electronic band structure of these Bloch states is the analogue, for a solid crystal, of the atomic- or molecular-orbital structure of an atom or molecule; from it almost all the physical properties involving electrons may be deduced.

Conduction electron

For example, a conduction electron in an almost pure, almost perfect, metallic crystal at low temperatures may have a mean free path of many microns. In a suitably contrived setup with microwaves and magnetic fields, subtle phenomena may be observed showing exactly how the wavelength of the corresponding Bloch function varies with the direction of propagation through the crystal. The Fermi surface mapped in this way is often quite

similar to the spherical surface in the free electron model, implying that the electrons really are able to travel quite easily through the lattice in almost any direction. This can be attributed to the weakness of the pseudopotentials of many metallic ions; they do not have as much effect on the electrons as we should have supposed on classical grounds. Of course impurities and imperfections in the crystalline order, especially the inevitable thermal motions of the atoms at ordinary temperatures, scatter the electrons from one Bloch state to another, reduce the mean free path to a few microns, and produce the temperature-dependent electrical resistivity characteristic of metals. Surprisingly enough this nearly free electron picture survives in metals above the melting point, where all semblance of regular crystalline order has disappeared.

Superconductivity. In many metals the electrical resistivity vanishes completely below a critical temperature T_c of a few °K. This superconducting state, first observed by K. Onnes in 1911, did not seem to depend very much on the purity of the specimen, and the state was later discovered to have other physical properties that could be empirically correlated by assuming that the proportion of superconducting electrons to normal electrons increased as the temperature fell below T_c. But the basic mechanism of the transition remained quite mysterious until 1950, when H. Fröhlich drew attention to the coupling produced between conduction electrons through the electron-phonon interaction.

In nontechnical language, an electron passing close to an ion shifts it slightly by electrostatic interaction, and this changes the field seen by any other electron in the vicinity. The effect is very small but is magnified by a subtle quantum-mechanical coherence factor associated with the condensed state of the Fermi-Dirac gas. The result is to pair the electrons and to produce a small energy gap over the whole Fermi surface of the metal. An electric current, once established, cannot decay because there are no allowed states into which the paired electrons can be scattered. As the temperature approaches T_c, however, ordinary unpaired electrons are excited into states above the energy gap and behave as normal electrons. The essential coherence of the wave functions of paired superconducting electrons is eventually destroyed, either by thermal excitation or by a strong external magnetic field. It is noteworthy, however, that coherence persists when electrons tunnel quantum mechanically through a thin insulating film between two superconducting metals.

Magnetic metals. In a transition element, such as iron, two shells of atomic orbitals are simultaneously incomplete. The spins of the d-electrons in the inner shell tend to line up to give the maximum magnetic moment—*i.e.*, several Bohr magnetons per atom. Salts containing positive ions of these elements therefore show strong magnetic effects. But what happens to the d-electrons in the metallic state where the valence electrons are freed to form an ordinary conduction band? The appearance of permanent magnetization shows that the polarization of the spins within each ion still persists and that there is now a tendency for magnetic moments in neighbouring ions to align parallel with one another. On the other hand, optical evidence supports the general principle that the true electron wave functions must form a narrow d-band of Bloch states spread throughout the crystal. The problem of reconciling the localization of the magnetic moments with the extended character of the states, and of explaining the interactions between spins in neighbouring ions, has by no means been finally solved.

d-electrons

Semiconductors. In some crystalline solids, such as silicon and germanium, the valence electrons occupy a complete band (Figure 9B). In the equivalent chemical language, all the bonds are saturated so the material ought to be an insulator. But the energy gap to the next band may be small enough (*e.g.*, 1 ev) for a few electrons to be set free by thermal excitation. At room temperature an intrinsic semiconductor has an appreciable electrical conductivity that increases rapidly with temperature as more electrons are excited into the conduction band. Excess electrons may also be introduced by deliberately replacing an atom of the crystal by one of higher valency, thus producing n-type material.

But an unoccupied Bloch state near the top of the valence band also allows the movement of electronic charge through the solid. By analogy with the positron, this hole is a p-type carrier since it behaves electrodynamically like an ordinary particle with a positive charge. Since an excess of holes may be produced by doping the crystal with an element of lower valency, crystals containing both n-type and p-type regions are easily manufactured and form the basis of innumerable solid state devices.

An n-type carrier in the conduction band is not really at all like a free electron. Being strongly influenced by the regular crystal lattice, it may appear to have an effective mass very different from m, often depending on the direction in which it is moving. The same applies to the holes, whose dynamical properties may be very complex in detail. Although scattered by impurities and by lattice vibrations, both types may have quite high mobility as carriers of current and do not recombine very rapidly when brought into the same region. The choice of a semiconductor material for practical use obviously depends on such properties.

Surface emission. The conduction electrons in a metal are contained by a surface potential barrier of several volts, the work function, ϕ. They may be helped to escape by several means (Figure 9C). When the temperature is high enough, thermionic emission is observed. The classical formula of O.W. Richardson (1902) gives a rate of emission governed largely by the Boltzmann factor, $\exp(-e\phi/kT)$; since the escaping electrons come from the high energy tail of the Fermi-Dirac distribution, this is also valid quantum mechanically. Schottky emission is also thermally activated but is aided by a strong external electric field (~1V/m [*i.e.*, volts/metre]), which slightly lowers the effective work function. At much higher fields the barrier becomes thin enough to allow cold emission by tunnelling. Light of frequency ν causes photoemission of electrons (Figure 9D) whose maximum energy is given by the Einstein formula (1905) $E = h\nu - e\phi$ with a spectrum of energy depending upon the structure of the energy bands in the solid from which the electrons have been excited. An energetic electron striking a metal surface may also cause the emission of one or more secondary electrons, the basic mechanism of the photomultiplier tube.

Tunnelling

A source for the emission of electrons into a vacuum should obviously have the lowest possible work function. A photocathode surfaced with an alloy of alkali metals containing caesium ($\phi = 1.8$V) is sensitive into the infrared but cannot stand up to high temperatures. The most efficient thermionic source is the oxide-coated cathode, a refractory metal with $\phi \sim 4$V covered with an alkaline earth oxide (*e.g.*, BaO). The mode of action is not thoroughly understood, but seems to depend on excess metal from the oxide forming a very thin film of low work function on the surface.

ELECTRON CURRENTS AND BEAMS

Control. A current of energetic electrons within an evacuated vessel is a characteristic element of many technical devices of electronics and electrical engineering. The electrons are usually produced from a thermionic cathode and controlled by electrical and magnetic fields generated by voltages applied to various electrodes within the vacuum vessel or by currents in external coils. Many such devices contain an electron gun, a succession of apertures through which the electrons are accelerated by electrostatic fields into a well-defined beam of a few kilovolts energy. If much higher energies are required ($> 10^7$ ev), the electrons are made to travel in bunches along a tube or round and round in a magnetic field whilst being continuously accelerated by a synchronous electric field. Each electron then suffers a vast relativistic increase of mass and travels very nearly with the velocity of light.

Since the flow of electrons is always outward from the cathode, the simplest thermionic vacuum diode is a recti-

fier of alternating current. This current may be controlled by the voltage applied to a grid through which it passes on its way to the anode, yielding a triode, the prototype electronic valve device. Further grids are added to control space charge, the buildup of electrons in a region of the vacuum tube giving rise to fields opposing the entry of further electrons, and secondary emission of electrons from the anode. To obtain amplification or oscillation at very high frequencies, the electrons are formed into a beam and travel in bunches through a succession of resonant cavities in a straight line or bent around a magnetic field.

Electron microscopes

A succession of precisely placed cylindrical electrodes or coaxial coils, suitably energized, can focus a beam of electrons just as if it were a beam of light passing through a lens. In the transmission electron microscope the beam is directed through the object, a thin slice of, say, biological material, which stops a proportion of the electrons in the denser areas. Further electrode or coil lenses then produce an image at very high magnification, which may be observed photographically or on a fluorescent screen. Alternatively, in the scanning electron microscope, the beam is focussed into a tiny spot that plays over the surface of the specimen, in synchronism with the beam in a cathode ray tube, where variations of current through the specimen are recorded as a picture of the surface. For X-ray microscopy, the spot is focussed on an anode where it approximates to a point source for X-rays. In the electron microprobe analyzer, the characteristic X-rays produced by the electron spot beam on the surface provides identification of the chemical elements in the specimen at that locality.

Chemical and physical effects. A beam of electrons striking condensed matter gives rise to numerous chemical and physical effects. The phenomenon of electron diffraction is the analogue of X-ray diffraction and provides crystallographic and structural information about the material. The study of surfaces by low energy electron diffraction, reflection of electrons of 10–100 ev, is technically difficult but likely to be rewarding. Fluorescence and secondary electron emission are the mechanisms used for the transformation of variations of electron beam current into visual information.

As discovered by W.K. Röntgen (1895), energetic electrons produce X-rays when they strike a solid surface. At a few tens of kilovolts, the continuous spectrum due to the sudden stopping of the charged particle is predominant, although the fluorescence line spectrum emitted by atoms that have lost deep core electrons is of great importance as an identifiable characteristic of the element. In fact, only about $\frac{1}{1,000}$ of the kinetic energy of the beam goes into X-rays: the rest appears as heat as the electrons are scattered, slowed down, and stopped within a depth of a few hundred atomic layers. Electron beam heating is a clean technique for producing a very high temperature over a very small surface region, and it has innumerable applications in microelectronics and materials technology.

Electron beam heating

Electrons with energies in the MeV range can displace atoms in metals, producing vacancies, interstitials, and other defects. At higher energies again, various nuclear reactions can take place, as well as the copious production of gamma radiation.

But, since the electron does not have any strong interactions with nucleons, the elastic scattering of GeV electrons by, say, a target of liquid hydrogen depends essentially on the electric charge distribution within the proton. The classical notion of an elementary particle as a tiny sphere of electricity still has its place in a quantum-mechanical world.

BIBLIOGRAPHY. R.A. MILLIKAN, *Electrons (+ and −), Protons, Photons, Neutrons, Mesotrons, and Cosmic Rays*, rev. ed. (1947), a description of the properties of the electron and other particles, with a good account of the growth of atomic theory; D.L. ANDERSON, *The Discovery of the Electron* (1964); S. GLASSTONE, *Sourcebook on Atomic Energy*, 3rd ed. (1967), basic information on atomic radioactivity, isotope and nuclear structure; D.H. FRISCH and A.M. THORNDIKE, *Elementary Particles* (1964), a factual description of elementary particles, their basic properties, and the experimental techniques used to study them; LINUS PAULING, *The Nature of the Chemical Bond and the Structure of Molecules and Crystals* (1960), a classic work on bond orbitals, resonance, and electronic structure of atoms; C. KITTEL, *Introduction to Solid State Physics*, 3rd ed. (1966) a discussion of the electronic properties of solids; V. WEISSKOPF, "Recent Developments in the Theory of the Electron," *Reviews of Modern Physics*, 21:305–315 (1949); T.L. SQUIRES, *An Introduction to Electron Spin Resonance* (1964), a discussion of the techniques of electron spin resonance and the practical problems associated with instrumentation.

(J.M.Z.)

Electronic Music and Instruments

In the broadest sense, any music produced or modified by electrical, electromechanical, or electronic means can be called electronic music. Some authorities prefer the term electrophonic music as being inclusive of those methods of sound production that are electrical but not technically electronic (*i.e.*, not dependent on devices utilizing the emission of electrons), but this name has not won acceptance in common parlance.

Definition and nature of electronic music

More narrowly, for a piece of music to be electronic, its composer must have anticipated the electronic processing that is subsequently applied to his musical concept, and the final product must reflect in some way his interaction with the medium he wanted to employ. This is no different from saying that a composer should have an orchestra in mind when he composes a symphony and a piano when he composes a piano sonata. When the definition is thus restricted, it becomes clear that a conventional piece of popular music does not become electronic music when played on an amplified guitar, nor does a Bach fugue become electronic music if played on an electronic organ instead of a pipe organ. Some experimental compositions, often containing chance elements and perhaps of indeterminate scoring, permit but do not necessarily demand electronic realization, but this is a specialized situation.

Electronic music in the narrower sense is produced from a wide variety of sound resources—from ordinary sounds and the sounds of musical instruments picked up by microphones to electronic oscillators (generating basic acoustical wave forms such as sine waves, square waves, and sawtooth waves) and complex computer installations —that are recorded on tape and then edited into a permanent form. Ordinarily, except for one type of performed music that has come to be called "live electronic music" (see below), electronic music is heard played back through loudspeakers either alone or in combination with ordinary musical instruments.

This article covers both early experimentation with electronic sound-producing devices—electronic music in the broad sense—and composers' subsequent exploitation of electronic equipment as a technique of composition. Throughout the discussion it should be clear that electronic music is not a style but rather a technique yielding diverse results in the hands of different composers.

Historically, electronic music is one aspect of the larger development of 20th-century music strongly characterized by a search for new technical resources and modes of expression. This search began very largely after World War II. Before 1945, composers sought to liberate themselves from the main Classical–Romantic tradition of tonal thinking and to reconstruct their thinking along new lines, for the most part either Neoclassical or atonal and 12-tone, in which a composition is built up entirely from a tone-row consisting of all 12 notes of the ordinary chromatic scale.

The pre-World War II period was accompanied by substantial experimentation with electrical and electronic devices. The most important outcome for the composer was the development of a number of electronic musical instruments (such as the Hammond organ and the theremin) that provided new timbres he could score into his compositions and that laid the technical foundations for the future development of electronic music proper—*i.e.*, composed with electronic techniques—from about 1948 onward.

Electronic music is represented not only by a wide variety of 20th-century works, most of them serious concert pieces, but also by a substantial literature of theatre, film, and television scores and, most recently, multimedia works that use all kinds of audiovisual techniques—and more besides. Electronic music for theatre and films seems to many an especially appropriate replacement for a disembodied, nonexistent orchestra heard from a tape or a sound track. Electronic popular music has also won adherents. This mostly consists of arrangements, often poor, of standard popular music for electronic synthesizers; the tentative use of electronic alterations by some of the more ambitious and experimental rock groups; and the preparation of recordings by studio techniques that are quite innovating and interesting.

HISTORY AND STYLISTIC DEVELOPMENT OF ELECTRONIC MUSIC

Beginnings. During the 19th century, attempts were made to produce and record sounds mechanically or electromechanically. For example, a German scientist, Hermann Helmholtz, traced the wave forms of regular sounds primarily to check results of his acoustical researches. An important event was the invention of the phonograph by Thomas Edison and Emile Berliner, independently, in 1877. This invention not only marked the beginning of the recording industry but also showed that all the acoustical content of musical sounds could be captured (in principle, if not in actuality at that time) and be faithfully retained for future use.

Cahill's pioneer work

The first major effort to generate musical sounds electrically was carried out over many years by an American, Thaddeus Cahill, who built a formidable assembly of rotary generators and telephone receivers to convert electrical signals into sound. Cahill called his remarkable invention the telharmonium, which he started to build around 1895 and continued to improve for years thereafter. The instrument, which was introduced in Holyoke, Massachusetts, in 1906, failed because it was complex, impractical, and could not produce sounds of any magnitude because amplifiers and loudspeakers had not yet been invented. Nevertheless, Cahill's concepts were basically sound. He was a visionary who lived two generations ahead of his time, and his instrument was the ancestor of present-day electronic-music synthesizers.

An Italian Futurist painter named Luigi Russolo was another ancestor of synthesized music. As early as 1913 Russolo proposed that all music be destroyed and that new instruments reflecting current technology be built to perform a music expressive of the industrialized present. Russolo subsequently actually did build a number of mechanically activated *intonarumori* (noise instruments) that grated, hissed, scratched, rumbled, and shrieked. Russolo's instruments and most of his music apparently vanished during World War II.

Between World War I and World War II, developments occurred that led more directly to modern electronic music, although most of them were technically, rather than musically, important. First was the development of audio-frequency technology. By the early 1920s basic circuits for sine-, square-, and sawtooth-wave generators had been invented, as had amplifiers, filter circuits, and, most importantly, loudspeakers. (Sine waves are signals consisting of "pure tones"—*i.e.*, without overtones; sawtooth waves comprise fundamental tones and all related overtones; square waves consist only of the odd-numbered partials, or component tones, of the natural harmonic series.) Also, mechanical acoustical recording was replaced by electrical recording in the recording industry in the late 1920s.

Invention of the electronic organ

Secondly, there was the development of electromechanical and electronic musical instruments designed to replace existing musical instruments—specifically, the invention of electronic organs. This was a remarkable achievement and one that absorbed the attention of many ingenious inventors and circuit designers. It should be stressed, however, that it was always the objective of these organ builders to simulate and replace pipe organs and harmoniums, not to provide novel instruments that would stimulate the imaginations of avant-garde composers.

Most electromechanical and electronic organs employ subtractive synthesis, as do pipe organs. Signals rich in harmonic partials (such as sawtooth waves) are selected by the performer at the keyboard and combined and shaped acoustically by filter circuits that simulate the formant-frequency, or resonant-frequency, spectra—*i.e.*, the acoustical components—of conventional organ stops. The formant depends on the filter circuit and does not relate to the frequency of a tone being produced. A low tone shaped by a given formant (a given stop) is normally rich in harmonics, while a high tone normally is poor in them. Psychologically, one expects this from all musical instruments, not only organs but also orchestral instruments.

Some organs operate on the opposing principle of additive synthesis, whereby individually generated sine waves are added together in varying proportions to yield a complex wave form. The most successful of these is the Hammond organ, patented by Laurens Hammond in 1934. The Hammond organ has odd qualities that have been overlooked by composers, mainly because it is so seldom heard to best advantage. The richness of its harmonic content does not diminish as the player goes up the keyboard. The German composer Karlheinz Stockhausen (in *Momente*, 1961–62), Arne Nordheim, the Norwegian composer (in *Colorazione*, 1968), and a few others have scored specifically for this instrument.

New instruments in the 1920s

Thirdly, there has been the development of novel electronic music instruments designed to supply timbres not provided by ordinary musical instruments. During the 1920s there was a burst of interest in building an extraordinary variety of such instruments, ranging from the practical and interesting to the absurd. The most successful of these new instruments are relatively few in number, are monophonic (*i.e.*, can play only one melodic line at a time), and survive because of their practicality and because some important music has been scored for them. These are the theremin, invented in 1920 by a Russian scientist, Leon Theremin; the Ondes Martenot, first built in 1928 by a French musician and scientist, Maurice Martenot; and the trautonium, designed by a German, Friedrich Trautwein, in 1930.

The theremin is a beat-frequency audio oscillator (sine-wave generator) that has two condensers placed not inside the circuit chassis but, rather, outside, as antennas. Because these antennas respond to the presence of nearby objects, the pitch and amplitude of the output signal of the theremin can be controlled by the manner in which a performer moves his hands in its vicinity. A skilled performer can produce all sorts of effects, including scales, glissandi, and flutters. A number of compositions have been written for this instrument since the 1920s.

The Ondes Martenot consists of a touch-sensitive keyboard and a slide-wire glissando generator that are both controlled by the performer's right hand and some stops controlled by the left hand. These, in turn, activate a sawtooth-wave generator that delivers a signal to one or more output transducers. The instrument has been used extensively by several modern French composers, particularly Olivier Messiaen.

The trautonium, like the Ondes Martenot, uses a sawtooth-wave generator as its signal source and a keyboard of novel design that permits not just ordinary tuning but unusual scales as well. Most of the music composed for this instrument is of German origin, an example being a *Concertino for Trautonium and Strings* (1931) by the German composer Paul Hindemith. In about 1950 a polyphonic version (capable of playing several voices, or parts, simultaneously) of this instrument was built by Oskar Sala, a former student of Trautwein and Hindemith, for preparing sound tracks in a Berlin film studio.

Tape music. With tape music the history of electronic music in the narrower sense of its definition begins. This history seems split into two main periods: an early (by now classical) period lasting from the commercial introduction of the tape recorder immediately following

Invention of the tape recorder

World War II until about 1960; followed by a second period, in which the tape medium has been dealt with less intensely for its own intrinsic properties, but rather in terms of the manner in which it fits into the musical and artistic environment. The invention of the tape recorder suddenly gave composers an exciting new musical instrument to explore, test, and use for new musical experiences. For more than a decade, fascination with the thing itself was bound to be the dominant motivation for composing electronic tape music. Also, musically, the 1950s, in contrast to the 1960s, were relatively introverted years: in all kinds of music, the focus of interest was technique and style, especially with the avant-garde. A composer whose concern centred not on these relatively neutral factors but on, for example, subjective expression or social activism was dismissed by the avant-garde as being out of touch.

In time, the medium became fairly well understood, the techniques for handling it became increasingly standardized, and a substantial repertory of characteristic and historically important compositions came into being. The burning issues were whether tape would replace live musicians; whether the composer was at last freed from the humiliations he so often endures to get his music into the concert hall; and whether a new medium of expression had been created, quite different from and independent of instrumental music, analogous, say, to photography as opposed to traditional painting.

It became increasingly evident, however, that there was no reason to think that the electronic-tape medium would eliminate instrumental performance by live musicians. Tape was increasingly regarded as something that could be—but did not need to be—treated as a unique medium. Thus the notion that the tape recorder could function as one instrument in an ensemble grew more and more popular, particularly in the concert hall. This conception obviated the visual monotony of an evening in an auditorium with nothing to look at but a loudspeaker. To this has been added a further stage of evolution, namely, live electronic music, in which the tape recorder and its tape is eliminated or greatly restricted in function, and transformations of the sounds of musical instruments are effected at the concert with electronic equipment. Not infrequently, this kind of performance environment also involves scores in which aleatory (chance, or random), improvisatory, or quasi-improvisatory musical guidelines for the manipulation of such equipment are supplied by a composer who prefers to let what happens just happen. Actually, it is open to question whether live electronic music is really an advance or a reversion to a more primitive state of the art, in the sense that it is the enhancement of the timbres of familiar instruments, rather than music conceived totally in terms of electronic media per se.

Establishment of electronic music studios

The first period of development is certainly one into which Europeans put the most consistent and sustained work. Tape music quickly gained recognition and financial support, and, before long, a number of well-equipped electronic music studios were established, primarily in government-supported broadcast facilities. Some important work was also done in the United States, but this was much more fragmentary, and not until after 1958 did Americans begin to catch up, either technically or artistically.

In 1948 two French composers, Pierre Schaeffer and Pierre Henry, and their associates at Radiodiffusion et Télévision Française in Paris began to produce tape collages (analogous to collages in the visual arts), which they called *musique concrète*. All the materials they processed on tape were recorded sounds—sound effects, musical fragments, vocalizings, and other sounds and noises produced by man, his environment, and his artifacts. Such sounds were considered "concrete," hence the term *musique concrète*. To this Paris group certainly belongs the credit both for originating the concept of tape music as such and for demonstrating how effective certain types of tape manipulation can be in transforming sounds. These transformations included speed alteration, variable speed control, playing tapes backward, and signal feedback loops. The engineers in this group built several ingenious special tape-manipulating machines. Schaeffer, however, was not interested in and in fact opposed the use of electronic oscillators as sound sources, claiming that these were not "concrete" sound sources, not "real," and hence artificial and anti-humanistic.

Two of the most successful and well-known *musique concrète* compositions of this early period are Schaeffer and Henry's *Symphonie pour un homme seul* (*Symphony for One Man Only*, 1949–50) and Henry's *Voile d'Orphée* (*The Veil of Orpheus*, 1953), extracted from a larger ballet score, *Orphée* (1953), written for the Belgian dancer Maurice Béjart. These and similar works created a sensation when first presented to the public. *Symphonie pour un homme seul*, a descriptive suite about man and his activities, is an extended composition in 11 movements. *Le Voile d'Orphée*, the finale of the ballet, is concerned with the descent of Orpheus into Hades.

The Paris studio has continued to flourish. The bias against purely electronic sounds gradually subsided, and compositions have been produced at a steady rate by a group of composers, including the Greek composer Yannis Xenakis.

The second event of significance was the formation of an electronic music studio at Cologne by Herbert Eimert, a composer working for the Nordwestdeutscher Rundfunk (now Westdeutscher Rundfunk, or West German Radio), who was advised in turn by Werner Meyer-Eppler, an acoustician from the University of Bonn. Eimert was soon joined by Stockhausen, who composed the first really important tape composition from this studio, the now famous *Gesang der Jünglinge* (*Song of Youth*, 1956). The Cologne studio soon became a focal point of the reemergence of Germany as a dominant force in new music.

Stockhausen's first tape composition

At Cologne, emphasis was immediately placed on electronically generated sounds rather than concrete sounds and on electronic sound modifications such as filtering and modulating rather than only on tape manipulation. Eimert and Stockhausen also published a journal, *Die Reihe* ("The Row"), in which appeared articles emphasizing the "purity" of electronic sounds and the necessity of coupling electronic music to serial composing (using ordered groups of pitches, rhythms, and other musical elements as compositional bases) that made no more sense than the Paris group's insistence on using only nonelectronic, non-serial material. This activity was part of the campaign of the 1950s that also brought about the collapse of Neoclassicism (a style that drew equally on 20th-century musical idioms and earlier, formal types), the emergence of the Austrian composer Anton von Webern as the father figure of the new music, and the development of total serialism, pointillism (a style making use of individual tones placed in a very sparse texture), intellectualism, and an emphasis on technique above all. The examples set by these two studios were soon widely imitated in Europe, with the sponsoring agencies being usually, as in these two original instances, government-owned broadcast organizations. This trend continued in the 1960s, with many more studios, from modest to elaborate, being set up in almost every major urban centre in Europe. As time passed, the techniques and equipment in the newer studios became more standardized and reliable, and the rather peculiar issue of concrete versus electronic sounds ceased to concern anyone.

In the United States, the production of electronic music, until 1958, was much more sporadic. The only continuing effort of this sort was the project of two U.S. composers at Columbia University, Otto Luening and Vladimir Ussachevsky, to create a professional tape studio and to compose music illustrating the musical possibilities of the tape medium. Luening and Ussachevsky often collaborated on joint compositions. They gained particular attention for the composition of several concerto-like works for tape recorder and orchestra. In 1959 Luening and Ussachevsky joined with another U.S. composer, Milton Babbitt, to organize, on a much larger scale, the Columbia-Princeton Electronic Music Center, in

which an impressive number of composers of professional repute have worked.

Varèse's *Déserts*

Other tape compositions in the early 1950s in the United States were largely those of individual composers working as best they could under improvised circumstances. One major composer who did so was Edgard Varèse, from France, who completed *Déserts* for tape and instrumental ensemble in 1954, and *Poème électronique*, for the Philips Pavilion at the 1958 Brussels World's Fair. Another was John Cage, an American, who completed the tape piece *Williams Mix* in 1952. Both Varèse and Cage had anticipated the electronic medium; Cage's *Imaginary Landscape No. 1* (1939) for RCA test records and percussion can well be regarded as a forerunner of current live electronic music.

Starting with the establishment of the Experimental Music Studio at the University of Illinois in 1958 by the U.S. composer Lejaren Hiller and the University of Toronto studio in 1959 by the Canadian composer Myron Schaeffer, the formation of facilities for both production and teaching began to move forward. The number of studios in university music departments grew rapidly, and numerous private studios also were established.

The number and variety of individual components of equipment may vary in a well-designed "classic" studio, but basically the equipment may be divided into five categories: sound sources such as sine-wave, square-wave, and sawtooth-wave generators, source for white noise (containing all possible sound frequencies), microphones for concrete sounds, and miscellaneous complex wave function generators; routing and control circuitry (patch panels and switching boards for coupling components together, adding circuits and mixers, amplifiers, output connections); signal modifiers (amplitude and ring modulators, frequency shifters, artificial reverberators, variable band-pass and third-octave or notching filters, variable-speed tape recorders, time compression–expansion devices, and other, more specialized, equipment); monitors and quality-control equipment (frequency counter, spectrum analyzer, VU metres that monitor recording levels from soft to loud, oscilloscope, power amplifiers with loudspeakers and headsets, and workshop facilities), and recording and playback equipment, including at least three high-quality tape recorders.

With this kind of equipment, the composer tape records sounds, both electronic and microphoned; modifies them singly or in montages by operations such as modulation, reverberation, and filtering; and finally re-records them in increasingly complex patterns. Inevitably, a major part of the composer's effort is involved in tape editing, unless he is satisfied with the crudest string of effects merely linked together in sequence. Like any other kind of music, the aesthetic merit of an electronic music composition seems to depend not only on musical ideas as such but also on the way in which they relate to one another and how they are used to build up a musical structure.

Integration of live instruments and tapes

The integration of the tape recorder with performance ensembles has become a rather popular form of chamber music, if not of symphonic music. Varèse's *Déserts* is an early example of this. It is scored for a group of 15 musicians and a two-channel tape and consists of four instrumental episodes interrupted by three tape interludes. In other works, the tape recorder is "performed" together with the remaining instruments rather than merely in contrast to them. The problems of coordination, however, immediately become overriding, for it is difficult for a group of performers to follow a tape exactly. Obviously, the tape dominates the situation, remorselessly moving along no matter what happens in the rest of the group. Balance of the tape loudspeakers with the group also has to be precisely adjusted, so as not to drown out the sound produced by the players and so that they can hear their cues.

Thousands of electronic-tape compositions were in existence by the early 1970s, many of ephemeral interest. It is relatively rare for a composer to have established his reputation solely as a composer of tape music. Pierre Henry perhaps is an example, but in general the important names in instrumental music of the 1950s and 1960s are the significant contributors in electronic music, too.

Stockhausen remained in the forefront of electronic-music composers with several important pieces following *Gesang der Jünglinge*. These included *Kontakte* (*Contacts*, 1959–60), for tape, piano, and percussion, and *Telemusik* (1966), a tightly organized piece for tape alone. Luciano Berio and Bruno Maderna, both Italians, worked for a while at the Radiotelevisone Italiana (at that time called Radio Audizioni Italia or RAI) studio in Milan. Besides *Différences* (1958–60), a composition for tape and chamber group, Berio's tape pieces include *Thema-Omaggio a Joyce* (*Homage to Joyce*, 1958) and *Visage* (1961), which exploit the unusual voice of American singer Cathy Berberian.

In the United States, the Columbia-Princeton Electronic Music Center has had the greatest output, a long list of composers besides Luening and Ussachevsky having used its facilities. Tape music from the University of Illinois studio includes Salvatore Martirano's *L's GA* (1967), a savage political satire for tape, films, helium bomb, and gas-masked politico. The University of Toronto studio, in spite of its technical excellence, has not been well represented on disks. One Canadian piece that is very amusing, however, is Hugh LeCaine's *Dripsody* (1955), all the sounds of which are derived from the splash of a single drop of water. John Cage's two principal tape pieces are *Williams Mix* (1952) and *Fontana Mix* (1958), both of which are complex tape collages.

Music synthesizers. Composing tape music by the classic method is neither easy nor free of technical pitfalls. A complex piece must be assembled from hundreds or even thousands of fragments of tape. Splicing all these sounds together consumes a vast amount of time and can also lead to an accumulation of errors and a deterioration of the sounds. Consequently, substantial efforts have been expended to reduce this work load and at the same time improve quality. Music synthesizers are the current product of these efforts. They cannot, however, be regarded as more than an intermediate technological development because of later computer technology to be reviewed below.

In contrast to Cahill's period, by the 1950s the means were finally at hand to construct full-scale music synthesizers, specifically, the RCA Electronic Music Synthesizers, designed by Harry Olson and Herbert Belar, research scientists working at the RCA Laboratories at Princeton, New Jersey. The first of two such machines was introduced in 1955. A second, improved model was turned over to the Columbia-Princeton Electronic Music Center in 1959.

The basic advance of the RCA synthesizer was an information input mechanism, a device for punching sets of instructions into a wide roll of punched paper tape. The composer had to know what the various coded punches would do and had to assemble his score sequentially from the beginning to the end of this tape; but he could, at any time during the programming process, interrupt this activity to listen to what he had punched, to make corrections, and to edit the material before making a final paper tape that then constituted the "master score" of the composition. The tape could be run at any desired speed; slowing did not lower the pitch of his sounds.

Contents of a synthesizer

The synthesizer contains a comprehensive assembly of signal generators, modulators, filters, and other components familiar from the electronic-music studio plus a few special circuits of its own, such as a glide generator. All those components are activated by the coded input prepared by the composer. The punches on the tape are decoded to yield a final product recorded on ordinary audio tape.

Although strong claims have been made that this machine can produce almost any sound of conceivable interest to musicians, it does have substantial limitations. It can produce at most only four sounds (four lines of music) at a time, and it lacks certain important functions, notably frequency modulation. The punched-tape input, and hence the master score, is valuable only so long as

the machine exists. The machine was also very expensive and was built with vacuum tube circuitry (electron tube), which requires substantial maintenance.

The composer whose name became particularly associated with the RCA synthesizer was Milton Babbitt. Among his compositions created with this machine are *Composition for Synthesizer* (1961), *Vision and Prayer* (1961), *Ensembles for Synthesizer* (1963), and *Philomel* (1964). Babbit had developed a precisely defined compositional technique involving total serialization (*i.e.*, of every musical element). When he became aware of the synthesizer, he was anxious to use it, because it gave him the opportunity to realize his music precisely rather than approximately, as had hitherto been the case.

In about 1960 a new circuit, the voltage-controlled oscillator (VCO), attracted the attention of engineers interested in electronic music because the frequency of its output signal is proportional to an independently generated input voltage rather than being internally set. The response is immediate because no mechanical couplings or controls are required. Also, these circuits can be built using transistor or integrated-circuit components, as opposed to vacuum tube circuitry, thus eliminating bulky configurations and high cost. An electrical engineer, Robert Moog, first designed and marketed several types of compact synthesizers of moderate price that supply a quite extended range of possibilities for sound manipulation. In addition to VCO's, which produce sine, square, sawtooth, and triangular waves, a Moog synthesizer contains white-noise generators, attack and decay generators (controlling a sound's onset and fading), voltage-controlled amplifiers, and bandpass filters and sequencers. One major advance in sound manipulation provided by VCO's is frequency modulation. If the input is a periodic function, the output frequency will vary periodically to provide tremolos, trills, and warble tones. Moog's synthesizer must now compete with several other synthesizers of essentially the same design, the Buchla Electronic Music Box, the ARP, and the Putney synthesizer. The last is an inexpensive instrument, apparently designed primarily for instructional use.

Design of the Moog synthesizer

Found in many music departments, these popular synthesizers eliminate much of the drudgery of tape splicing, but at a price. The range of timbres and processes is more limited because they operate by subtractive synthesis and impose transients that affect all partials (component vibrations) of a complex wave identically. A great advantage of a harmonic tone generator built by James Beauchamp, an electrical engineer at the University of Illinois, also from VCO's, is that it uses additive synthesis—*i.e.*, creating sound by combining signals for pure tones (sine waves)—instead of producing them by removing partials from a complex signal. It is so designed that each partial of a sound can have its own entry point, its own rise time, and its own decay time. The improvement in tone quality is enormous, because the ear normally expects nuances such as higher partials decaying faster than lower ones. Salvatore Martirano's *Underworld* (1965) is a good example of music in which the tape was made largely by additive synthesis.

A composer closely associated with synthesizers is Morton Subotnik, who has produced such extended pieces as *Silver Apples of the Moon* (1967). These pieces were created on the Buchla synthesizer, and any one of them demonstrates in relatively unmodified form the types of sounds one may obtain with these instruments.

A word may be said about realizations of instrumental music through synthesizers, notably an early, commercially successful album called *Switched-on Bach*, arrangements made by Walter Carlos on a Moog synthesizer. The record displays technical excellence in the sounds created and makes the electronic synthesis of music more intelligible to the general listening public, which is useful so long as it is realized that the materials on the record are arrangements of familiar music, not original compositions.

The success of *Switched-on Bach*

Computer music. The latest development is the use of electronic digital computers for both music composition and analysis. These can be used in five different ways: (1) routine data processing and information retrieval, including music library applications and abstracting; (2) processing of music notation and music printing; (3) acoustical, theoretical, and musicological research; (4) music composition; and (5) sound synthesis. In all these fields, considerable research and experiment was being carried out in the 1970s, with perhaps musicological research, the third way noted above, being the most widespread and advanced activity. But composition and sound synthesis, which are complementary processes because the first may lead to the second, are of most direct interest to composers. A composer may elect to use a set of compositional programs to produce a composition. He may then stop using a computer and print his results for transcription to instrumental performance. Alternatively, he may transfer his results directly into electronic sounds by means of a second set of programs for sound synthesis. Finally, he may desire only to convert an already composed score into sound. When he does this, he translates his score into a form that can be entered into a computer, say, by data cards, and uses the computer essentially as a data translator.

Use of electronic digital computers

The first point to understand about computer composition is that it is not a style but a technique. In principle, any kind of music, from traditional to completely novel, can be written by these machines. For a composer, however, the main appeal consists not in duplicating known styles of music, but, rather, in seeking new modes of musical expression that are uniquely the result of interaction between man and this new type of instrument.

At present, composers above all need a compiling language comprised of musical or quasi-musical statements and a comprehensive library of basic compositional operations written as closed subroutines—in effect, a user's system analogous to computer languages (such as Fortran) used by mathematicians. Incomplete attempts have been made to fill this need, one of which is Musicomp, worked on intermittently by Lejaren Hiller since 1963. But two major obstacles stand in the way of building up an effective musical computer language. The first is the obvious one of allocation of sufficient time, money, and other resources. The second is defining what goes into the subroutine library; *i.e.*, of stating with precision the smallest units of activity or decision making that enter into the process of musical composition. Unlike mathematics, in which traditional modes of thinking prepared the way for such a definition of subroutines, in music the defining of "modules" of composition leaves even sophisticated thinkers much more at sea.

The earliest example of computer-composed music is the *Illiac Suite for String Quartet* (1957) by two Americans, the composer Hiller and the mathematician Leonard Isaacson. It was a set of four experiments in which the computer was programmed to generate random integers representing various musical elements, such as pitches, rhythms, and dynamics, which were subsequently screened through programmed rules of composition. Several basic styles of composition—from traditional to experimental—were tested in an exploratory way, and the results were transcribed for instrumental performance. Even at this early stage, the composers realized that the best way of investigating computer composing was to proceed from random to orderly; that is to say, by programming what should be rejected.

Earliest example of computer-composed music

Two very different compositions, *ST/10-1,080262* (1962), by Yannis Xenakis, and *HPSCHD* (1968), by John Cage and Lejaren Hiller, are illustrative of two typical approaches to this technique. *ST/10-1,080262* is one of a number of works realized by Xenakis from a Fortran program he wrote in 1961 for an IBM 7090 computer. Several years earlier, Xenakis had composed a work called *Achorripsis* by employing statistical calculations and a Poisson distribution to assign pitches, durations, and playing instructions to the various instruments called for in his score, so that a uniform distribution of all these occurs during the course of the composition. He redid the work with the computer, retitled it, and at the same time produced a number of other, similar compositions. *HPSCHD*, by contrast, is a work of indeterminate

length scored for one to seven harpsichords and one to 51 tape recorders. When first performed in 1969, it was put on as a multimedia work and was allowed to run nearly five hours. For *HPSCHD* the composers wrote three sets of computer programs. The first, for the harpsichord solos, solved Mozart's *Musical Dice Game* (K294d), an early chance composition in which successive bars of the music are selected by rolling dice, and modified it with other compositions chosen with a program based on the Chinese oracle *I Ching* (*Book of Changes*). The second set of programs generated the 51 sound tracks on tape. These contained monophonic lines in microtone tunings based upon speculations by the composers regarding Mozart's melodic writing. The third program generated sheets of instructions to the purchasers of a record of the composition.

Basic techniques for producing sounds with a computer

The three basic techniques for producing sounds with a computer are sign-bit extraction, digital-to-analog conversion, and use of hybrid digital–analog systems. Of these, the first is cheap but limited, the second is the most sophisticated and expensive, and the third is the most recent, permitting synthesis in real time.

Sign-bit extraction has occasionally been used for compositions of serious intent—for example, in *Computer Cantata* (1963), by Hiller and Robert A. Baker, and in *Sonoriferous loops* (1965), by Herbert Brün—but more often it is an amusement indulged in around computer laboratories for playing Christmas carols and things of that sort. A number in the accumulator of a computer is really a sequence of binary digits (bits) that are physically represented by two different voltage levels. Consequently, if a position like the sign bit is caused to swing periodically from 0 to 1 (+ to −, arithmetically) by, for example, repeatedly adding a number to itself, and an electrical tap is taken from this spot, a square wave can be obtained. Because a square wave contains the odd partials of the harmonic series, it can be shaped, within reason, by filters and transients generators to provide a modest range of timbres.

Digital-to-analog conversion is by far the ideal technique for computer sound synthesis. This process was developed in the U.S. by Max Mathews and his colleagues at Bell Telephone Laboratories in the early 1960s. Its latest version is called Music V.

Digital-to-analog conversion (and the reverse process, analog-to-digital conversion, which is used to put sounds into a computer rather than getting them out) depends on the sampling theorem. This states that a wave form should be sampled at a rate twice the bandwidth of the system if the samples are to be free of quantizing noise (a high-pitched whine to the ear). Because the auditory bandwidth is 20–20,000 hertz (Hz), this specifies a sampling rate of 40,000 samples per second though, practically, 30,000 is sufficient, because tape recorders seldom record anything significant over 15,000 Hz. Also, instantaneous amplitudes must be specified to about 12 to 13 bits so that the jumps from one amplitude to the next are low enough for the signal-to-noise ratio to exceed commercial standards (55 to 70 decibels).

Music V is more than simply a hardware system, because it embodies an "orchestration" program that simulates many of the processes employed in the classical electronic-music studio. It specifies unit generators for the standard wave forms, adders, modulators, filters, reverberators, and so on. It is sufficiently generalized that a user can freely define his own generators.

PROBLEMS AND PROSPECTS

Science has brought about a tremendous expansion of musical resources by making available to the composer a spectrum of sounds ranging from pure tones at one extreme to random noises at the other. It has made possible the rhythmic organization of music to a degree of subtlety and complexity hitherto unattainable. It has brought about the acceptance of the definition of music as "organized sound." It has permitted the composer, if he chooses, to have complete control over his own work. It permits him, if he desires, to eliminate the performer as an intermediary between himself and his audience. It has placed the critic in a problematic situation, because his analysis of what he hears must frequently be carried out solely by ear, unaided by any written score.

In the future, certainly, more inexpensive synthesizers will be manufactured, especially many aimed at commercial exploitation and institutional use by music students. On a more sophisticated level, the further development may be expected of hybrid digital–analog systems for major electronic studio installations and, on a more modest level, as live-performance instruments. Better input systems may also be expected for entering musical and instructional information into computers for both composition and sound synthesis. Punched cards and tape and their equivalents will be replaced by optical scanners, light pens, and similar devices that trace a score directly onto a plate to which the computer is sensitive. One may hope that by the 1980s some really generalized and comprehensive programs for music composition with computers will begin to exist in a form that is transferable from one computer to the next. Another development will be more computer applications in the visual arts that can be coordinated to music, especially computer films and control systems for theatrical and intermedia presentations. The use of computer systems should increase in Europe, where little has been done so far (exceptions are the University of Utrecht studio directed by Gottfried Michael Koenig and the new Swedish Radio studio in Stockholm). Also, it is hoped that digitalized computer archives of important electronic music will be established, because audio tape seriously deteriorates with time, taking with it the present repertory.

It is important also that research in music acoustics be encouraged, not only for a better understanding of the objective physical sounds themselves, about which much has yet to be learned, but also for a better understanding of the subjective psychological response to all such sounds. The composer is no longer limited to using sounds that can be played by conventional instruments but only to those that can be perceived by the ear.

It is, of course, extremely hazardous to predict what will happen artistically. Perhaps the 1970s will prove to have been a period of consolidation of the technical discoveries and inventions and abstract systematic compositional systems of the 1950s with the much more uninhibited and flamboyant musical activities of the 1960s.

BIBLIOGRAPHY. The two most thorough histories of electronic music up through the 1950s are ABRAHAM MOLES, *Les musiques expérimentales* (1960); and F.K. PRIEBERG, *Musica ex Machina: Über das Verhältnis von Musik und Technik* (1960). Books that deal primarily with relevent aesthetic problems and artistic viewpoints include JOHN CAGE, *Silence* (1966); LUIGI RUSSOLO, *L'arte dei rumori* (1913; Eng. trans., *The Art of Noises*, 1967); K. STOCKHAUSEN, *Texte zur elektronischen und instrumentalen Musik*, 2 pt. (1963–64); and IANNIS XENAKIS, *Formalized Music* (1971). Electronic musical instruments and their components, many of which are also used for electronic music, are discussed in R.H. DORF, *Electronic Musical Instruments*, 3rd ed. (1968); ALAN L. DOUGLAS, *The Electronic Musical Instrument Manual*, 5th ed. (1968), and *The Electrical Production of Music* (1957); and WERNER MEYER-EPPLER, *Elektrische Klangerzeugung* (1949). A discussion of computer music may be found in H. BRUN, *Über Musik und zum Computer* (1971); HEINZ VON FOERSTER and J.W. BEAUCHAMP (eds.), *Music by Computers* (1969); L.A. HILLER and L.M. ISAACSON, *Experimental Music* (1959); L.A. HILLER, *Informationstheorie und Computermusik* (1964); H.B. LINCOLN (ed.), *The Computer and Music* (1970); and M.V. MATHEWS *et al.*, *The Technology of Computer Music* (1969). Periodicals that consistently printed important articles on electronic music include *Die Reihe* (German or English edition); *Electronic Music Review* (monthly); and *Gravesaner Blätter* (German and English); for current articles, see *Darmstädter Beiträge zur Neuen Musik* (German); *Source, Perspectives of New Music, Journal of Music Theory* (all three semi-annual); and the *Audio Engineering Society Journal* (monthly). Recordings of electronic music are now so numerous that it would be almost impossible to give a good list here. Consequently the reader is referred to HUGH DAVIES (comp.), *International Electronic Music Catalog* (1968); and to the *Schwann Long Playing Record Catalog* (both regular and supplementary editions) for listings.

(L.Hi.)

Electronics

The scope of electronics has grown so wide that older definitions of it are no longer adequate. Originally coined to describe the behaviour of free electrons in evacuated containers, the term was later made to include their behaviour in gases, liquids, and solids, and ultimately was applied to a particular branch of electrical engineering; and, yet, even electrical engineering, which is concerned with the movement of electrons along conductors, may with more logic be said to be a branch of electronics, which is concerned with electrons wherever they appear. The central fact in all of this is that, as the subject of electronics, the electron is so universal and so fundamental a constituent of matter that it is virtually impossible to limit the scope of the subject and the potential of the industry.

For a discussion of the theoretical aspects of electronics, see such articles as ATOMIC STRUCTURE; ELECTRICITY; MAGNETISM; ELECTRON; and SOLID STATE OF MATTER.

For a fuller discussion of the principal electronic applications, see such articles as BROADCASTING; COMPUTERS; ELECTRON TUBE; RADAR; RADIO; SOUND RECORDING AND REPRODUCING; TELEVISION; SEMICONDUCTOR DEVICES; and TELEMETRY.

In the 1970s electronics was one of the largest industries in the United States and was a major industry in several other countries. By the end of the century it may well be the largest industry in the world.

In the United Kingdom electronics output has risen by 80% since 1961 and the electronics industry is now overtaking the cotton industry. A similar story is repeated in Japan and the countries of Western Europe.

Electronics may be the key technology of the 1970s; its development may determine the direction of technology as a whole. Whatever major field of social activity or world industry is selected, it will be found that electronics is playing an increasingly dominant role in it. It is plausible that this is nowhere better illustrated than in the field of communications.

Sound and television broadcasting covers most of the earth's surface. Fourteen years after the first satellite was put into orbit in 1957, satellite communications (*q.v.*) linked most of the world with telephone and television circuits.

Electronics and communications

Electronics plays a key role in whatever aspect of communications is considered: submarine cable; automatic telephone networks; wide-band television circuits; and information processing, storage, and retrieval. Developments have come so fast that the major problem is how society can adapt to them.

Outside the communications field, electronics plays an increasing role in industry; for example, it can transform the situation in the machine shop. A modern milling machine is designed for the difficult and costly operation of machining complex shapes in three dimensions from solid lumps of hard metal. Now, this sequence can be directed by a digital computer and electronic control in which the programmer converts design data into a series of numbers representing the coordinates of sections through the solid component, together with instructions for feed and speed. The numbers are coded on punched tape that is read by a digital computer and recorded on magnetic tape in the form of instructions to the machine tool. The tape is then fed to the electronic control unit. Errors in the machining are automatically corrected. The machine can be programmed to produce one complex part after another or to repeat any part as often as required. The larger the number of components in a given batch, the greater the saving in machine time and cost when compared with the cost of skilled machinists. More than half of the large machine tools in some countries are now numerically controlled.

History

The word electron comes down from antiquity, having been derived from the Greek word for amber, which, as the ancient Greeks observed, attracts light bodies when rubbed. The character of the electron as a negatively charged particle in modern terminology owes something to Benjamin Franklin, whose kite experiment in 1752 demonstrated the "sameness of electrical matter with that of lightning." Franklin suggested that the "electrical fire" is present in all matter, and that an excess of the normal amount appears as a "positive" charge while less than the normal amount appears as a "negative" charge. To be consistent with this nomenclature, modern electronics has ascribed a "negative" charge to the electron, whose removal from neutral atoms leaves the atoms in the state Franklin described as positively charged.

The invention of the torsion balance, a magnetic needle suspended by a fine fibre, permitted accurate measurement of electrical forces and the important discovery (1785) that the action of the electric fluid, like that of gravity, was less as the distance between the charged bodies became greater. The invention of the voltaic pile, or battery, about 1800 was followed by a century of advance in knowledge of electricity and magnetism, which included the study of the behaviour of the new force in a vacuum. From these experiments came the conclusion that the particles whose activity was revealed were the basic substance of matter. A British physicist, J.J. Thomson, called them "electrons." A classic series of experiments during the opening years of the 20th century produced measurements of the mass and charge of the electron.

In the meantime, a curious phenomenon encountered by the United States inventor Thomas A. Edison in his development of the electric light led to an important discovery. Edison noted that under certain conditions of vacuum and voltage a bluish glow appeared in his lamp, indicating the existence of an unexplained current flowing between the lamp's two filament terminals in a direction opposite to that of the main current. Intensive research, especially that carried on by Thomson, revealed that the "Edison effect" was due to the passage of electrons from the negative terminal (cathode) to the positive terminal of the filament. One of Thomson's pupils worked out the theory of the physical process involved, which he called thermionic emission. Another British electrical researcher, Sir John Ambrose Fleming, who was scientific adviser to Edison's company in Britain, had the idea of using the Edison effect for the reception of radio waves. Preparing a vacuum tube with two metal plates, he heated the negative, or cathode, plate to incandescence by means of a battery, so that electrons were emitted in a continuous stream toward the positive, or anode, plate. When Fleming's tube was inserted in the circuit of a radio receiver, the anode became alternatively positive and negative according to the fluctuations of the incoming radio waves. Electrons were attracted to it when it was positive and repelled from it when it was negative, thus permitting the current to flow in only one direction. This rectifying process, also called detection and demodulation, produced a rectified output current that could be used to operate a telephone receiver or other recording device. The signals, however, were very weak. In 1907 a young American engineer, Lee De Forest, decisively improved Fleming's tube by adding a third electrode, shaped like a gridiron, between the cathode and anode. This controlling grid enabled the tube to deliver power without absorbing power and so made amplification of radio signals possible, and this led, through a number of intermediate steps, to the large-scale development of broadcast radio in the 1920s. One other major development took place in these early years, the invention by Vladimir Zworykin, a Russian-born American engineer, of the iconoscope, a device based on the photoelectric effect—*i.e.*, the ejection of electrons from metals by the action of light—which made possible the scanning and conversion of images into electric currents that is the basis of television.

Theory of thermionic emissions

The electronics industry was no more than on the verge of its major development when World War II suddenly mobilized the engineers and scientists of Europe, Japan, and the United States. By the end of the war, the U.S. and U.K. in particular had built up massive industrial resources for the production of electronic equip-

Invention of the transistor

ment. The invention of the transistor in 1948 provided a small, cheap, and simple substitute for the vacuum tube and gave a fresh impetus to the industry, especially by greatly facilitating development of the computer and opening new possibilities in industrial automation. With the introduction first of microcircuits and then of integrated circuits in the 1960s, electronics opened still further horizons in space, in the laboratory, and in industry.

Components

Modern electronic engineering encompasses a great variety of different kinds of component parts called "hardware," as opposed to "software," the name applied to the circuits and systems that use these components. A convenient method of classification is to divide the components into those that contain a source of power, called active, and those that do not, called passive.

ACTIVE COMPONENTS

Active components include such devices as electron tubes and transistors that serve as oscillators, amplifiers, and perform other specialized functions in electronic circuits.

Vacuum and gas-filled tubes. Fleming's crude diode of 1904 was the first of a vast family of electronic tubes that generate, rectify, amplify, modulate, and control electric signals. From 1905 to 1960 the production of these devices increased from a handful in a few laboratories to hundreds of millions a year in modern factories. The greatest rate of increase occurred after World War II with the rapid expansion of television, and this rate had only just begun to decline in the early 1970s under the impact of the advent of semiconductor devices.

From the point of view of circuit function and application, electron tubes can be grouped into four classes: General Purpose; Image; Photoelectric; and Microwave (see also ELECTRON TUBE).

General purpose tubes. Such tubes include all those whose basic functions are rectification, amplification, oscillation, frequency changing, and switching. They are called diodes, triodes, tetrodes, pentodes, hexodes, etc., according to the number of elements or electrodes they contain. In every case there is a cathode, or negative electrode, from which electrons are obtained, usually by heating the electrode and an anode or plate, the positive electrode. The introduction of a third electrode in the form of a grid constitutes a triode; a fourth produces a tetrode or screen grid tube; and so on.

The graphical symbol for a vacuum diode is shown in Figure 1A. Its function is primarily to rectify; *i.e.*, to change alternating inputs into direct pulsating outputs. Diodes are rated on the basis of their peak inverse voltage; *i.e.*, the maximum negative voltage that can be used on the plate without damaging the tube. Typical values for vacuum diodes are of the order of 1,000 volts but values as high as 25,000 volts can be attained by proper tube design.

The triode can amplify as well as rectify. The amplification capability is greatest when the space between grid and cathode is kept as small as possible. The spacing is minimized by the use of so-called frame grid construction, in which the grid is wound with very fine wire on a rigid framework and is spaced only about 0.002 inch (0.05 millimetre) from the cathode. The spacing between electrodes is also important in determining the time the electrons take to travel between electrodes, called the transit time; the upper frequency limit at which the tube will operate satisfactorily is determined by this time.

The tetrode includes a second (screen) grid whose function is to screen the control grid and cathode from the anode. This improves the performance but on the other hand electrons emitted from the screen grid tend to worsen the performance. The drawback is overcome in the pentode by introducing a third grid, called a suppressor grid, to repel the secondary electrons emitted from the anode. Very high amplification factors over extended frequency ranges can be obtained with modern pentodes.

Vacuum tubes having six or more electrodes, such as hexodes and heptodes, are used mainly as frequency changers. In such devices some of the electrodes are used to generate a local oscillation and others to combine this with the incoming signal frequency and so produce a number of new frequencies, called beat frequencies. Modern practice is to incorporate the two sets of electrodes in one envelope with a common cathode.

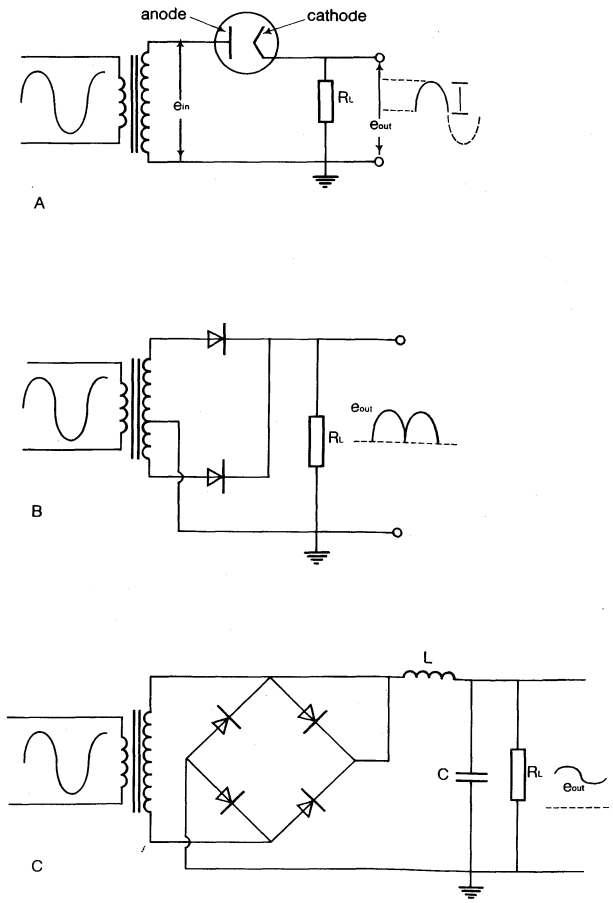


Figure 1: *Rectifier circuits.*
(A) Half-wave rectifier: vacuum diode. (B) Full-wave rectifier: semiconductor. (C) Bridge rectifier with LC filter circuit.

Image tubes. These tubes use the electrical input signal to generate and display a light image. This group includes cathode-ray oscilloscope tubes and television picture tubes. Their essential features are an electron gun for generating an electron beam; a means of deflecting and controlling the beam; and a screen that produces light when struck by the electron beam. The electron gun assembly is mounted in the neck of the tube and horizontal and vertical deflection of the beam is accomplished either electrically by passing the beam through two pairs of parallel charged plates, or magnetically by sets of coils wound round the neck. Modern cathode-ray oscilloscope tubes will respond to very small signals, and their speed of response is high. Using the best modern techniques, it is possible, for example, to display and photograph an event lasting only a thousand millionth (10^{-9}) of a second.

Features of an image tube

In the picture tube for television receivers, sometimes called a kinescope, magnetic deflection is provided by two pairs of coils formed round the neck and the absolute brightness of the spot produced by the electron beam is controlled by a grid in the electron gun. In the projection kinescope, used for theatre television, the very high intensity image in the tube is attained by accelerating voltages of up to 60,000 volts.

Photoelectric tubes. These tubes generate an electrical signal from an input of light energy. The simplest phototube consists of a cathode with a photoemissive surface (a surface which gives off electrons when struck by light) and an anode. Incident light produces electrons from the cathode which are drawn to the positive anode. Very low intensity light can be detected with a sensitive tube called a photomultiplier tube.

The most complex of the photoelectric tubes are the television camera tubes, in which the image of a scene focussed on the photosensitive surface is converted into an electrical signal—the video signal—which is transmitted to a television receiver where it is used to reproduce the original image.

Microwave tubes. Special tubes are required for the production and amplification of microwaves; *i.e.*, waves of very high frequency and short wavelength. The general principles are that the interelectrode spacings must be very small and that the inductance (opposition to the flow of alternating current) of the leads and interelectrode capacitances must be kept to a minimum. In the magnetron, the electron beam is controlled by crossed magnetic and electric fields. It is capable of generating pulses of extremely high power at frequencies up to about 1,000,000,000 (10^9) cycles per second (one gigahertz).

The klystron is a microwave receiving tube in which velocity control of the electron beam replaces the intensity control of the triode. The effects of transit time are reduced to a minimum by impressing the input signal on a part of the beam that has a very high velocity so that the transit time limitations are less serious than in a triode.

In the class of microwave tubes called travelling-wave tubes, the interaction between the signal wave and the electron beam occurs over the entire length of the tube.

Gas-filled tubes. The basic properties of gas-filled tubes depend on electric discharges in the gases. In a gas diode electrons emitted from the cathode collide with and ionize the molecules of a contained gas, such as neon, so that very large changes in current occur for very small changes in anode voltage. The gas diode can therefore be used to absorb very large changes in current with only small variations in its output voltage; *i.e.*, as a voltage regulator.

The thyratron is a gas-filled (hydrogen or mercury vapour) triode used for the control of high-power circuits. In the ignitron an electrode dips into a mercury pool that is the cathode and initiates a discharge between the ignitor and the mercury pool.

Semiconductor devices. Although semiconductor electronics (the study of materials that are neither good insulators nor good conductors) is a relatively recent development, semiconductor devices, in the form of crystal detectors, have been used since the beginning of the century (see also SEMICONDUCTOR DEVICES). The critical development was the invention of the point-contact transistor after half a century of theoretical advance in the physics of the solid state. This early device, in which contact was made to a semiconductor crystal by two closely spaced, pointed wires, was never a successful industrial product because its construction limited very severely the currents that it could handle, but it demonstrated the practical possibilities of electron flow in crystals and led to the development of the junction transistor.

Point-contact transistor

The junction transistor consists essentially of a sandwich made up of one layer of crystalline semiconductor, having an excess (called N-type) or deficit (P-type) of electrons, between two layers of semiconductor, which accepts (P-type) or donates (N-type) the electrons. It proved to have so many advantages over thermionic tubes that it led to an acceleration in the growth of electronics and to the development of a large family of active semiconductor devices that took over many of the functions of thermionic tubes.

The field-effect transistor is virtually a solid-state version of the thermionic triode and has many of the advantages of the thermionic tube without its disadvantages. It has replaced the tube in such applications as frequency meters, amplifiers, flip-flops (described later), timer circuits, and in applications in which a constant current is necessary.

The thyristor, a transistor having a thyratron-like characteristic, has been developed and is used for switching and controlling power of the order of megawatts.

Because the transistor does not need a heated cathode for its supply of active electrons, it is more efficient, smaller, lighter, and more robust than the thermionic tube equivalent. High-speed digital computers would not have been possible if transistors had not been available to replace the thousands of bulky tubes otherwise required for the large, powerful, machines.

Integrated circuits. Printed circuit techniques marked another stage in the general evolution towards smaller, more compact, and more reliable equipment. In these, solid-wire connections to individual components were replaced by printed connections on a flat circuit board on which the components were soldered. A density of 30 components per cubic inch was attained with these methods compared with about 10 per cubic inch that was the best attainable with miniature tubes and conventional equipment.

The next major development was to dispense entirely with individual containers, and transistors and their associated components, together with their interconnections, were formed in place, as thin films on glass or ceramic bases (substrates). The techniques used in producing these integrated circuits include diffusion, photolithography, etching, alloying, and metal evaporation and deposition (see also INTEGRATED CIRCUITRY). With increasing sophistication and know-how, the technique progressed to the point at which the packing density was increased to 1,000 per cubic inch by 1960 and to 100,000 per cubic inch by 1965. Packing densities up to several million per cubic inch are now possible, but these densities create a number of problems in practice, such as the dissipation of heat generated. Circuits formed in a single chip of silicon crystal are called monolithic and are now found in virtually every computer.

Components formed in place

As of the early 1970s, the latest development in this field was large-scale integration, also called array technology, in which the silicon wafer is covered with integrated circuits, which need not be identical. The circuits are tested by probes, the results are fed to a computer, and the optimum interconnection pattern for coupling the good circuits is obtained. Input and output connections are decided by the computer and added at the places indicated on the slice.

PASSIVE COMPONENTS

The important electrical characteristics of the passive components in an electronic circuit are those of resistance, capacitance, and inductance.

Resistance is a measure of the opposition a material presents to the flow of electric current. It is determined by many factors including the atomic structure of the material, its temperature, its impurity content, and the density of free electrons available to move under the influence of an electric field. The unit of resistance is the ohm, defined as the ratio of the voltage across a conductor to the current flowing through it. Thus, if an applied voltage of one volt produces a current of one ampere, the resistance is one ohm.

Capacitance, the ability to store a charge on a pair of conductors separated by an insulator, is measured as the ratio of the stored charge to the potential produced by the charge. The unit is a farad, a very large unit, usually replaced in electronics with microfarads (10^{-6} farad), or picofarads, formerly called micromicrofarads (10^{-12} farad).

Inductance, a magnetic property of conductors, is defined as the ratio of magnetic flux to the current required to produce that flux. The unit is the henry, or one weber per ampere. Both capacity and inductance depend upon the geometrical configuration of the conductors.

All conductors possess resistance, inductance, and capacitance. In circuit components, one or other of these is made to predominate, and the undesired properties are usually reduced to a minimum.

Basic properties of all conductors

Resistors. Resistors may be either fixed or variable, depending on circuit requirements. The principal forms of fixed resistors are carbon-composition, wire-wound, and metal-film. A cylindrical shape is convenient for fixed resistors and was used for most of them until the introduction of microcircuits.

In the carbon-composition resistor the resistive element is a cylinder of carbon with impurities added to control

the resistivity. The material is press-formed and wire leads are attached at each end and the assembly is encapsulated.

The metal-film resistor employs a very thin film of metal deposited on a glass or ceramic cylinder as the resistive element. Gold caps with leads attached are soldered to the film at each end and the unit is encapsulated. The effective length of the film may be varied before encapsulation by cutting a helix through the film on the circumference of the cylinder and so varying the length of the conduction path.

All these types of fixed resistors are usually colour-coded to indicate the nominal value of resistance and the allowable tolerance, or deviation, from this value.

Variable resistors are also either wire-wound or composition types. A sliding contact of some kind is mounted in such a manner that the contact position can be changed, thus varying the resistance.

From the circuit point of view, all resistors possess inductance and capacitance, which may be ignored at low frequencies but must be taken into account at higher frequencies. It is possible in resistor design to balance the effect of incidental inductance and capacitance over a limited frequency range; this provides the so-called non-reactive types.

Capacitors. The principal property of capacitors is their ability to store a charge. Generally speaking, in electronics they consist of two sets of metal plates separated by a vacuum, air, or a liquid or solid insulating material called the dielectric. For any particular dielectric, the capacitance increases as the area of the plates is increased and decreases as the distance between them is increased. Similarly, an increase in the dielectric constant of the insulating material increases the capacitance.

Mica, a mineral, has a high dielectric strength—*i.e.*, it resists high voltages—and a high dielectric constant, making it suitable for robust, compact capacitors. In these, metal foil is interleaved with layers of mica; leads are attached and the whole assembly encapsulated. For greater precision, a metal film can be deposited directly on the mica, as in the case of silver-mica capacitors. Mica capacitors are made in values from 10 to 5,000 picofarads and operating voltages from 500 to 20,000 volts.

Paper, film, and oil-filled capacitors all consist of alternate layers of metal and dielectric rolled into a cylinder. Leads are attached and the assembly encapsulated or suspended in oil.

In ceramic capacitors the dielectrics are ceramic materials such as titanates with high dielectric constant and strength. Physically small assemblies can provide quite high capacitances.

In wet electrolytic capacitors, the electrolyte, a solution of borax, phosphate, or carbonate, is used to form the dielectric and saturates the paper separating the conductors. In dry, or solid, capacitors a layer of tantalum or aluminum oxide is used to make possible extremely compact and small high-value capacitors: up to 3,500 microfarads with tantalum and 150,000 microfarads with aluminum oxide. Electrolytic capacitors have two important characteristics; the very thin dielectric restricts the maximum voltage and permits more leakage current than do thicker dielectrics, and the plate with the oxide film must always be kept positive with respect to the other plate or the capacitor will heat up and may even explode. They are, however, very useful where moderate temperatures and low voltages apply and space is limited, since high values can be obtained. Oil-filled capacitors are superior for high-voltage, high-capacitance applications.

Characteristics of capacitors

Variable capacitors are usually made with two meshing sets of metal plates, one movable and one fixed, with air as the dielectric, although this arrangement is bulky and presents mechanical problems. Variable capacitors from 10 to 500 picofarads can be made for voltages up to 5,000 depending on the separation of the plates.

Inductors. Inductance is a magnetic property and the design of inductors is based on the concept of energy as stored in the magnetic field which accompanies a current (see also MAGNETS AND ELECTROMAGNETS). Inductance depends on the geometry of the inductor, but for any given shape and cross-sectional area it is directly proportional to the square of the number of turns and the permeability (the ease with which lines of magnetic flux can pass through a material) of the core and inversely proportional to the length of winding. There is usually another important criterion, called the "Q," which is a measure of losses at any particular frequency. For direct current the losses depend on the effective resistance of the winding; as the frequency of the current is increased the losses will increase because they will occur in both the conductor and the core. The choice of ferromagnetic dust cores for higher frequencies minimizes losses in the core; the use of stranded wire reduces the effective resistance due to high-frequency current losses in the wire.

The simplest form of inductor consists of one or more turns of sufficiently heavy-gauge wire to be self-supporting so that air is the only insulator to be considered. At frequencies of the order of megahertz (10^6 hertz) it is easy to achieve Q's of 500 with this very simple design. For low frequencies, the greater inductance usually required necessitates a larger number of turns and therefore much smaller gauge wire. If this is wound as a single layer, it requires a cylindrical former or bobbin for mechanical support.

Much greater values of inductance can be obtained by using high-permeability magnetic material in the core, which, for radio frequencies, often consists of particles of magnetic powder (ferromagnetic dust) compressed into a doughnut shape, through which the wire is threaded. One advantage of this construction is that the magnetic field is confined almost entirely to the toroid. Inductances of about one millihenry to one henry are common with this design.

Very high values of inductance, up to 1,000 henries or so, usually for low or voice frequencies, can be obtained by increasing the size and number of turns and filling the core space with interleaved laminations of high-permeability silicon-steel alloy. Variable inductors are obtained either by moving a connection from one turn to another on the conductor so as to vary the effective number of turns, or by varying the position of a core of compressed magnetic powder mounted in the centre of the coil.

Transformers are special forms of inductors widely used in electronics to transform alternating power to one or more circuits at the same frequency, but usually with changed values of voltage and current. They operate on the principle of electromagnetic induction. All transformers consist of two or more windings so arranged that a change of flux in one winding induces a change of flux in the other. They are usually provided with a high-permeability material to increase the efficiency of the flux linkage and this material is laminated or otherwise divided to minimize the energy losses that would occur in solid material. Suitably designed transformers provide an accurate method of multiplying or subdividing voltages and currents since these are proportional to the ratio of the turns in the windings.

Transformers

Solid-state devices. In addition to the active semiconductor devices discussed above, electronics has produced a variety of passive solid-state devices.

Varistors. An important property of semiconductors is that a non-ohmic, or nonlinear, resistance can be obtained when a suitable contact is made between a semiconductor and a metal. Solid devices based on this behaviour, called varistors—*i.e.*, variable resistors—can be subdivided into symmetrical varistors—*i.e.*, those that have the same non-ohmic volt-current characteristics for both directions of applied voltage; and nonsymmetrical varistors—*i.e.*, those whose characteristics depend on the polarity of the applied voltage.

Nonsymmetrical varistors act as solid-state rectifiers that can be further subdivided into large-area rectifiers and point-contact rectifiers. Large-area rectifiers are generally made from either copper oxide or selenium. In the copper oxide devices, coin-shaped blanks punched from copper sheets are oxidized on one side with an oxide film, which acts as the semiconductor, and the units, after an-

nealing and drying, are assembled on a rod to form the finished rectifier.

Rectifiers. Selenium rectifiers are made by applying a thin layer of selenium to a backing washer of nickel or aluminum and heat-treating to convert the selenium to crystalline form. An outer contact of cadmium is sprayed on the free surface of the selenium layer. Selenium rectifiers are used in the small and medium power range.

Rectification can occur not only between the metal and the semiconductor but also at the contact between two semiconductors, and this is the basis of the germanium junction rectifier and the silicon junction rectifier, which have largely replaced the thermionic diode in electronics.

Zener diode as voltage regulator

A particularly useful variety is the Zener diode, in which the reverse voltage is very nearly constant for a wide range of reverse current; consequently, it can be used to provide a constant voltage where current demands vary.

Another interesting junction diode is the tunnel diode, which, owing to special characteristics, is a useful device as an amplifier or oscillator at high frequencies.

For high-power applications, silicon rectifiers consisting of units made from a mixture of silicon carbide, graphite, and ceramic applied between metallic electrodes are used. A silicon-controlled rectifier or thyristor is the solid-state equivalent of the thyratron and normally represents an open circuit but can be switched rapidly to the conducting state when an appropriate signal is applied to the control terminal. It has become popular as a fast solid-state switch over a very wide range of power.

Thermistors. A class of semiconductor devices that makes use of a large temperature coefficient of resistivity are the so-called thermistors, or thermally dependent resistors. These are made in a great variety of shapes and sizes by chemically preparing the desired mixture of semiconductors in powder form, molding the powder into a body of the desired size and shape, and heating until the grains are sintered into a compact, homogeneous mass. Metallic contacts are provided by firing on silver, gold, or platinum paint and soldering lead wires to the painted regions. The devices are enclosed in various forms of encapsulation. The resulting bead, rod, disc, or flake thermistors are sensitive to both voltage and temperature changes and have found numerous applications: in carrier-repeated amplifiers to compensate for temperature changes in the line; in microwave and high-frequency power meters; in signal limiters; in voltage regulators; in negative-resistance amplifiers; as time-delay devices; in medical thermometry; and in meteorology.

Photoconductive and photovoltaic devices. Another large field of devices, the photoconductive and photovoltaic devices, belongs almost entirely to the semiconductor area. The physical effect exploited in photoconductive devices is the liberation of extra electrons in the semiconductor by the absorption of light energy and subsequent change in the resistance. The usual method of making photoconductive cells is by preparing a film of the semiconductor over a metal or between metal electrodes. In the selenium cell a thin selenium film is applied over a metal layer and divided by a scribed line into two sections. In the lead-sulfide cell a thin film of lead sulfide is used between two electrodes and the unit is sealed into an evacuated glass tube. The point-contact phototransistor is similar to the conventional point-contact transistor but the activating electric charge is liberated by light rather than being introduced by another electrode. P-N junction and multijunction phototransistors can be produced with a current of only a microampere or so when not illuminated. They are often preferred to other cells in applications where a circuit must discriminate between "light-on" and "light-off."

Photovoltaic cells exploit the ability of semiconductor P-N junctions or rectifying contacts to generate a voltage when illuminated. The photoconductive devices described earlier also have photovoltaic properties and the design of photovoltaic cells amounts to maximizing these properties. A recently developed and most promising member of the family is the silicon solar cell, which achieves a conversion efficiency of more than 10 percent in converting sunlight into electric power.

Thermoelectric devices. Thermoelectric effects, the production of voltages between semiconductor junctions maintained at different temperatures, and the production of temperature differences by passing current through such junctions, are used in solid-state cooling units or thermal-battery devices.

Hall-effect devices. The Hall effect in semiconductors, the production of a voltage in a strip of the material carrying a current when it is placed in a magnetic field perpendicular to the current flow, permits the design of very sensitive measuring and indicating devices.

Ferrites. Magnetic properties are also involved in so-called ferroelectric and ferrite devices, which have applications as memory and storage components in computers, and as specialized components in high-frequency circuits (see also FERRITES).

SENSING DEVICES AND TRANSDUCERS

Transducers are devices that change energy from one form to another. Of principal interest are those in which either the input or the output is in the form of electrical energy.

Pickups for sound and vibration; microphones. The most familiar transducer, for converting sound waves and vibrations into electric signals, is the phonograph pickup, which is normally an electromechanical transducer actuated by modulations in the grooves of records and converting them into electrical outputs. These output signals, however small in amplitude or complex in waveform, can be reproduced at the required level with the help of amplifiers. Pickups for sound reproduction may be based on a variety of different elements; *e.g.*, piezoelectric crystals, capacitance variation, inductance variation, motion of an electrode in a tube, or a beam of light. Examples of piezoelectric crystals are: quartz, tourmaline, Rochelle salt, and the titanates; pickups embodying these crystals make it possible to measure noise and vibration, thus aiding the equipment designer in minimizing or eliminating these destructive properties.

Crystal pickups

Microphones, which constitute another well-known class of electro-acoustic transducer, respond to sound waves and deliver essentially equivalent electric waves. The many different types cover almost every conceivable application. The carbon microphone depends on the variation in resistance of carbon granules with pressure variation. The electronic microphone generates its voltages by the motion of one of the electrodes in an electron tube. The electrostatic microphone uses the variations in its electrostatic capacitance due to sound waves. The magnetic-armature microphone employs an armature to move a coil in a magnetic field to produce the electric waves. The ribbon microphone uses a moving conductor in the form of a ribbon that is directly driven by the sound waves. Standard microphones whose response is accurately known can be used to calibrate sound fields and pressures. Contact microphones can be actuated by mechanical contact of the diaphragm with the throat, as in the throat microphone; and the ear microphone is designed to fit in the ear. Extremely sensitive measurements and records of sound pressures can be made with such devices (see also SOUND RECORDING AND REPRODUCING).

Temperature sensors: thermocouples, resistance thermometers. The temperature sensors used in electronics are based on the fact that temperature changes alter electrical properties of the sensor. Thus the thermocouple is based on the thermoelectric effect, an electromotive force resulting from the difference in temperature between two junctions of dissimilar metals. A commonly used device, the vacuum thermocouple, consists of a rhodium-platinum alloy joined to a platinum wire and sealed with the leads in an evacuated tube. When a current is passed through the assembly, the thermocouple is heated and a voltage develops and can be calibrated to give the value of the current; if the value of the current is known, the difference of temperature between the hot and cold junctions can be calibrated on a meter scale. In a thermocouple thermometer the measuring junctions are placed in contact with the body whose temperature is to be measured.

The resistance thermometer, which depends on the variation of electrical resistance with temperature, utilizes a conductor that has a large change of resistance with temperature, in a form that can be placed in the region whose temperature is to be measured. The resistance must be measured independently of the leads and a bridge network is usually employed for this.

Thermistors are an example of semiconductor devices that have found increasing application as sensors for temperature changes and measurements (see also THERMOMETRY).

Stress and strain sensors: strain gauges. Pickups that convert mechanical stresses and strains into electrical currents or voltages are extensively used as the transducers in telemetry (*q.v.*). An electric strain gauge consists of a resistance wire or printed-circuit wound so that a change in stress produces a change in resistance. Electromagnetic strain gauges consist of a small armature and an iron-core coil whose spacing distance varies with the movement of the member in which the armature is fixed. Change in armature gap varies the inductance which in turn controls the frequency of a transmitting oscillator.

Stress changes resistance

Proximity sensors: capacity and induction types. Proximity sensors are based on the variation of electrical properties with the relative spacing of their constituent parts in relation to surrounding objects. Thus, the capacitance altimeter is a capacity proximity sensor in which altitude is measured by change of capacitance between two conductors on the aircraft that are affected by the distance to the ground. The redistribution of current in a conductor or coil brought about by the presence of another conductor can be used to indicate the proximity of the conductors. Variation of inductance can be made very sensitive to the proximity of a magnetic core or metal disc.

Chemical sensors—conductivity cells. The state of ionization of a solution—or its hydrogen-ion concentration—can be conveniently and accurately measured by the variation in its conductivity. For this purpose a conductivity cell, which is a liquid container with two electrodes, is connected into an arm of a bridge network and filled with the liquid to be measured. The conductivity can be measured and calibrated directly to yield the hydrogen-ion concentration, or it can be calibrated against a known standard cell.

Optical sensors: photocells, TV camera tubes. Optical sensors can now be effective over a great range from the long infrared part of the optical spectrum to the extreme ultraviolet. They are capable of very great sensitivity especially in conjunction with photomultipliers, and it is possible to select a suitable sensor for a specific part of the whole spectrum.

Practically any physical or chemical property can be converted into electron movement with a suitable transducer. Thus, transducers and electron devices have been developed for the investigation of high and low pressures, viscosity, turbidity, colour, flow-rate, humidity, hardness, roughness, polish, elasticity, and many others.

Circuitry

In this section some typical electronic circuits will be described in general terms.

POWER SUPPLY CIRCUITS

Rectifiers. Power supplies are found in virtually all electronic equipment. A power supply is usually the source of direct-current power in a circuit and thus requires the conversion of an alternating-current input to a steady direct current output. The usual methods involve a rectifier that converts alternating current into pulsating direct current and a filter unit that smooths the pulsations. Voltage regulators may also be used when accurate control of output voltage is required. The inverse process of obtaining alternating current from direct-current supplies involves inverters and vibrators.

Figure 1 shows how rectifier properties are exploited in power-supply circuits. The vacuum diode of Figure 1A is connected to the secondary winding of the transformer so that its anode is positive during the positive half cycle of the alternating voltage on the secondary winding. Thus the tube conducts during half of the cycle and current flows during this half from cathode to anode, leading to the term half-wave rectifier. A half-wave rectified voltage appears across the load resistance R_L in the output circuit.

Half-wave and full-wave rectifiers

Figure 1B illustrates a full-wave rectifier using two semiconductor diodes across the secondary winding, which has a centre-tap connected to ground. The resulting waveform across the load R_L is shown.

In Figure 1C four semiconductor diodes are used in a bridge rectifier. This arrangement has the advantage that for the same input voltage as Figure 1B, it produces about double the output voltage. The bridge rectifier produces the continuous rectified current with a superimposed alternating current—or ripple—as shown.

Filters. Filters are used widely in electronics to select or reject voltages of a given frequency or range of frequencies. They are illustrated here with reference to power supplies but they have many other important functions as well.

In Figure 1C the filter structure consists of the inductance (L), connected in series with the load (R_L), and the capacitor (C) shunted across the load. The unvarying part of the rectified current passes with relatively low loss (due to ohmic resistance in the inductor) to the load, while the alternating or high frequency part finds a relatively easy path to earth through the capacitor and does not appear in the load. Thus, the direct voltage appearing across R_L is smoother than it is at the bridge output. This process of filtering in a rectifier is called smoothing. With suitable combinations of R_L, L, and C, filter networks can be designed to accept or reject any band or range of frequencies. The band-pass filter, for example, consists of series and parallel, tuned R-L-C sections.

Voltage regulators. As noted before, gas diodes and Zener diodes have the ability to maintain a constant voltage over a large current range. This characteristic provides the most popular method of regulating the output of rectifier circuits within well-defined narrow limits. A suitable diode, such as a neon tube or a Zener diode, is connected in series with the output of the rectifier filter and the output voltage is taken from across the diode. A second method of voltage regulation is the use of vacuum tubes or transistors to sense small changes in the direct output voltage and compensate for these changes, thus keeping the final voltage constant.

Inverters and vibrators. Inverters were originally simple rotating machines which converted direct current into alternating current; such a conversion can now be achieved with thyristors, thus eliminating all moving parts. The basic unit consists of two thyristors with separate grid excitation connected either in series or in parallel with the load.

Vibrators are electromagnetic devices for turning continuous steady current into a pulsating current. Before the advent of transistors, they acquired some importance as alternating-current converters for equipment that was powered from a low-voltage battery (*e.g.*, a car radio) but their main use now is to act as choppers for low-value, direct-current signals, which can then be amplified with normal alternating-current techniques.

SWITCHING AND TIMING CIRCUITS

The functions of switching and timing can be performed electronically much better in many cases than by purely mechanical techniques, because of the greater accuracy, sensitivity, reliability, and speed of response. The ideal switch is either a perfect open circuit (infinite resistance) or a perfect conducting path (no resistance), which the ideal diode approximates. Such a diode conducts perfectly with one polarity and allows only a very tiny current flow with the opposite polarity. Switching operations formerly carried out by tubes are now largely performed by transistors and semiconductor diodes. The use of diodes as on-off switches amounts to selecting a diode with a suitable current and voltage range and selecting the operating point.

Timing and time delay circuits. These circuits are based on the ability of electronic systems to produce an

accurately controlled series of switched pulses of known frequency and duration. Thus, a circuit based on a quartz crystal that vibrates at a constant frequency can be stepped down with frequency dividing circuits to produce a low frequency of great accuracy. This is essentially the technique employed in international time standards, although the fundamental frequency is now that of a precisely measurable energy transition occurring in cesium atoms. The scope of electronics is nowhere better demonstrated than in the field of time measurement; the calibration signals from standardizing laboratories are radiated to an accuracy of one in 10^{13} and can be received within $\frac{1}{10}$ second in laboratories at opposite ends of the earth.

Time-delay operations can be carried out electronically quite simply; short periods can be timed automatically and repeated as often as desired. A large capacitor C (Figure 2) is connected in the transistor circuit in such a

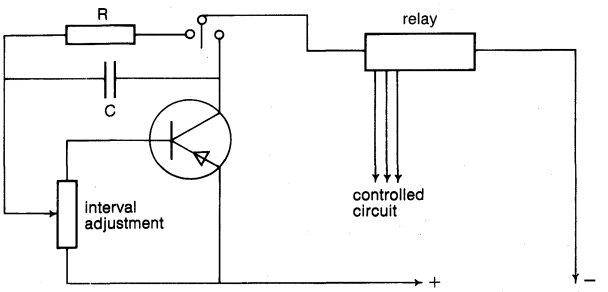


Figure 2: Interval timer.

way as to hold the relay armature down. When the current through the relay coil fails, the armature is released; the time of this occurrence can be altered simply by adjusting the potentiometer resistor.

Multivibrators. A useful source of all kinds of outputs at controlled frequencies is illustrated in the two-transistor network of Figure 3, called a multivibrator. When the circuit is running, one transistor drives the other and the frequency of the output pulses is largely controlled by the values of resistance and capacitance. The system has one stable state and one unstable state and tends to remain in the stable state until an appropriate trigger signal causes it to change. It is also called a flip-flop and is the basis of a very widely used binary counter.

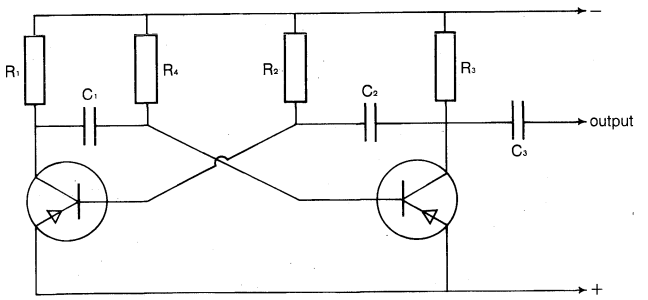


Figure 3: Multivibrator using two transistors.

Gating circuits. The logic processes in computers are arrived at by using binary counters of the type which change state at each input pulse. Large numbers of these circuits connected in chains and sequences carry out the logical operations of the computer and act as gates for the pulses passing through the chain. Thus, a so-called AND gate is one that produces an output only if all its inputs are energized and produces no output if one or more of its inputs are not energized. An OR gate is one that produces an output if any of certain selected inputs is energized but none if none of these selected inputs is energized. There are other types of gates but most logic systems may be constructed from AND and OR gate circuits together with a NOT inverter that produces the complement of the operating voltage; *i.e.*, its output is positive if input is negative and negative if input is positive.

Logic gates used in computers

AMPLIFIER CIRCUITS

The primary function of an amplifier is to increase the amplitude or power of a signal; the active elements used to do this are tubes and transistors. Amplifiers can be classified according to function, frequency range, and percentage of time of input cycle during which current flows in the active element (see Tables 1 and 2).

Table 1: Classification of Amplifiers by Frequency Range

type of amplifier	frequency range*
Subaudio frequency (VLF)	0–20 Hz
Audio frequency (AF)	20 Hz–20 kHz
Low frequency (LF)	20 kHz–300 kHz
Medium frequency (MF)	300 kHz–3 MHz
High frequency (HF)	3 MHz–30 MHz
Very high frequency (VHF)	30 MHz–300 MHz
Ultra high frequency (UHF)	300 MHz–3000 MHz
Microwave	above 3,000 MHz

*1 Hz = 1 cycle per second; 1 kHz = 1,000 cycles per second; 1 MHz = 1,000,000 cycles per second.

Coupling methods. The obvious method of obtaining any desired amplification from tubes or transistors is to connect them in sequence as required. Thus, the basic unit is the single-stage amplifier that consists of the active device and all the associated components that accompany such a stage. A series of units in sequence forms a multistage amplifier whose gain (or amplification) is approximately the product of the individual unit gains. In practice, input and output stages often differ from each other and from intermediate stages; the design is affected by considerations of cost, weight, size, stability, and so on. But a primary consideration is the technique for coupling the various stages together. These techniques can be divided into alternating-current coupling and direct-current coupling.

Table 2: Classification of Amplifiers by Percentage of Input Cycle in the ON Stage

class	percentage of input cycle tube or transistor is ON	grid voltage
A	100%	above cut-off
AB	between 50% and 100%	near cut-off
B	50%	at cut-off
C	less than 50%	below cut-off

In alternating-current coupling the link between the output of one stage and the input of the next includes either a capacitor or a transformer; although it may also include resistors and conductors there is no direct current path in the link. Thus Figure 4 shows resistance-capacitance-coupled, two-stage transistor amplifier. The frequency response of this amplifier would depend very largely on the relative values of C and R_6 but this amplifier would not be suitable for very low frequencies.

In direct-current coupling there is a direct-current path between stages, the coupling being made through a resistance network. This type of amplifier is useful for amplifying small changes in the level of small direct-current voltages and can also be used for alternating-current signals of low and medium frequency.

Main types of amplifier. The necessity of obtaining the maximum signal or power gain in a wide variety of conditions has led to a corresponding variety in the design of amplifiers.

Audio. Audio-frequency amplifiers, covering the frequency range 20 hertz to 20 kilohertz, usually have a nonselective response, and the frequency-gain characteristic is designed to be as constant as possible over the operating range. The preferred method of coupling is resistive, or inductive and resistive. The so-called high-fidelity amplifier—*i.e.*, one designed to amplify audio frequencies with a minimum of distortion—is only one link in a chain of equipment so that account must be taken of the wave shape and impedance (opposition to flow of alternating current) of the source and the impedance and character-

High-fidelity amplifiers

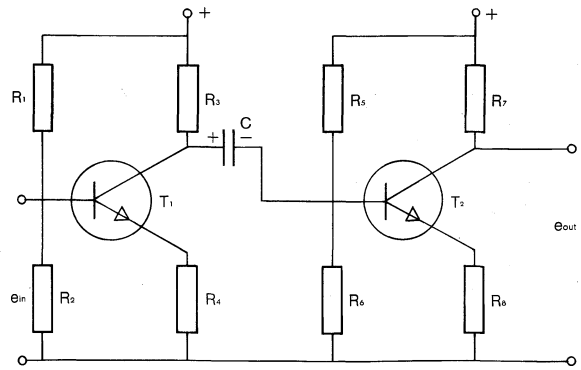


Figure 4: Amplifier using resistance-capacitance-coupled transistors.

istics of the speaker or other reproducer. Stability and reliability may entail the use of negative feedback (described later).

Video. Video amplifiers are wide-band amplifiers used to amplify the video frequencies in radar pulses and the periodic visual signals in television. The frequency range is about 30 hertz to five megahertz. As they are sometimes required to carry the direct-current component of the video signal, direct-current coupling is the method normally used.

Radio frequency. Radio-frequency amplifiers are generally selective in operation and are often tuned; *e.g.*, by tuning the grid or anode circuits or by including a band-pass filter in the coupling network. They generally use Class A for low power, Class B for high-power modulated waves, and Class C for high-power unmodulated waves (see Table 2).

Microwave. Microwave amplifiers, covering the frequency range from about 3,000 megahertz up to the infrared region where wavelengths are of the order of millimetres, employ special tubes, such as the klystrons and travelling-wave tubes mentioned earlier, and semiconductor diodes. Recent developments in electronics have led to great progress in this field. Lasers and masers (*q.v.*) use a completely new principle for the amplification of low-level high-frequency radiation by allowing it to react with electrons in such a way as to increase the energy of the radiation. This is the best available solution for amplifying faint signals from remote sources such as radio stars, without amplifying as well the thermal noise (a random noise generated in all components, its amplitude depending on temperature of the components) always present in ordinary amplifiers. The maser is operated at temperatures near absolute zero, where the inherent noise of components is negligible. Lasers represent the extreme limit of amplifying devices.

Special types. It is broadly true to say that vacuum-tube and field-effect transistor amplifiers act as voltage amplifiers with a relatively high impedance and junction-transistor amplifiers act as current amplifiers with a relatively low impedance. If a high input impedance combined with a low output impedance is required, a cathode follower—also called a grounded-grid circuit—may be used. In the case of a vacuum tube the input is applied between control grid and ground and the output is taken from an impedance between cathode and ground. In a field-effect transistor the drain electrode is grounded and for a junction transistor the collector is grounded. This is really a process of negative feedback used to achieve the desired result.

The so-called chopper amplifier mentioned above has become increasingly popular for amplifying low-value direct-current signals. Both mechanical and electronic devices are used to chop the direct-current signal into an alternating-current signal, then amplified with normal alternating-current techniques. Very low noise levels in relation to input may be obtained with these amplifiers.

Differential amplifiers have two similar input circuits connected so as to respond to the difference between two voltages or currents and effectively suppress voltages and currents that are alike in the input circuits.

The phase inverter or paraphase amplifier converts a single input into a push-pull output by combining cathode and anode coupling.

Magnetic amplifiers, which do not use tubes or transistors as amplifiers, can be employed to obtain large output power for a small input power. A direct-current or slowly varying alternating-current input or control current is made to control a large output power by changing the impedance of a coil in the output circuit. The magnetic circuit is designed to reach magnetic saturation so that both control and output windings are wound on a toroidal saturable core.

Feedback. The process of returning a fraction of the electrical or acoustical output of any stage in an amplifier to the input of that stage or of a preceding stage results in an increase or decrease in amplification depending upon the relative phase of the portion returned. This is called feedback: positive feedback if the return is positive with respect to the input signal, and negative feedback if it is negative.

Positive feedback will increase the amplification but will reduce the stability and introduce distortion. There is always a point beyond which positive feedback will result in oscillation. Thus, sufficient positive feedback in an audio-frequency system will produce an oscillatory condition known as howling.

Negative feedback will reduce the amplification but will increase the stability and bandwidth of the amplifier, and reduce distortion. It will also increase the input impedance and decrease the output impedance: the cathode-follower amplifier already described is essentially a method of negative feedback.

Negative feedback may be used in alternate stages in multistage amplifiers to improve gain stability. Each pair of stages is isolated from the other.

High-gain direct current coupled feedback amplifiers, called operational amplifiers, can perform a number of operations such as multiplication, addition, and integration of an input signal. They are widely used in measurement and control applications as well as in analogue computers. Feedback amplifiers are also used to stabilize the voltage output of a rectifier power supply to a greater extent than is possible with a simple Zener-diode circuit.

OSCILLATORS

Electronic circuits can generate alternating-current signals of every variety of waveform and over a wide range of frequency and can amplify them as well. The only convenient way of generating high-frequency voltages is with vacuum-tube or transistor oscillators, widely used for this purpose in signal generators for testing and measurement; in timing and counting equipment; in radio and television transmitters and receivers; for dielectric and induction heating; and in radar systems.

There are two broad classes of oscillator: harmonic oscillators and relaxation oscillators. Harmonic oscillators generate a nearly pure single frequency and employ resonant circuits such as tuned circuits of inductance and capacitance, microwave cavity resonators, and piezoelectric crystals. Relaxation oscillators generate complex waveforms such as square and triangular waves; they are less stable than harmonic types.

Variable frequency L-C and R-C oscillators. Positive feedback will produce oscillation; the frequency may be determined by a feedback network made up of variable inductance and capacitance or variable resistance and capacitance.

Figure 5 is a simplified circuit diagram of a typical variable frequency L-C oscillator: the transistor Colpitts oscillator. Positive feedback is achieved from the collector electrode via R_1, L, and C_1, C_2, with the emitter grounded. The inductance is tuned by the capacitor C_3 and the feedback ratio of the system is determined by the ratio of C_1 to C_2.

An example of a simple oscillator circuit employing a conventional amplifier stage and an R-C feedback network is the transistor phase-shift oscillator, in which the necessary phase shift is obtained by three sequential resistance-capacitance circuits.

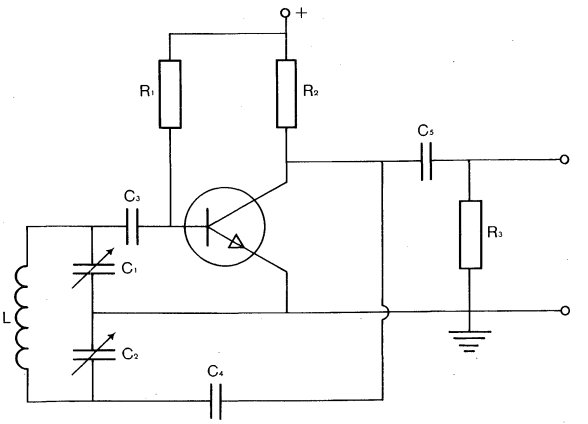


Figure 5: Colpitts oscillator circuit.

Relaxation oscillators. Circuits that employ the nonlinear or negative-resistance characteristics of the active element are termed relaxation oscillators. The neon-tube oscillator—using a gas-discharge tube—is a simple example. A capacitor connected across the tube is charged through a resistor and the voltage across it increases until the breakdown voltage is reached, at which point discharge begins and the voltage drops to a low maintaining voltage. The resistor is made too large to supply the current to maintain this discharge, so the tube turns off and the capacitor begins to recharge. Repeated cycles produce a sawtooth waveform.

A silicon-controlled rectifier, or thyristor, can produce results similar to those of the neon tube and the breakdown voltage can be adjusted by varying the gate (control) current.

Fixed-frequency oscillators. The usual form of fixed oscillator is the crystal oscillator that makes use of a piezoelectric device (*q.v.*); that is, a crystalline material that develops an electric charge or potential difference across some of its faces when the crystal is subjected to mechanical strain. Quartz is almost universally used because it is hard, robust, and strong and has a small temperature coefficient (small change in frequency as the temperature changes). The general principle is to connect the crystal in the feedback circuit. In a Pierce oscillator, the circuit is identical to that of the Colpitts oscillator of Figure 5 except that the crystal replaces the resonant circuit. Frequency stability of 100 parts per million (10^6) is typical and even greater accuracy and stability can be achieved by putting the crystal in a temperature-controlled oven.

Quartz crystal oscillators

Transit-time oscillator. The transit time of electrons in vacuum tubes—that is, the time required for the electrons to travel between tube elements—becomes of major importance in the generation of high frequencies. The corresponding motions of electrons in semiconductor solids can be exploited to produce microwave oscillators. Thus, the Read diode, developed in 1958, is used as a transit-time oscillator for microwave applications. Acoustic-wave amplifiers involve the use of a travelling-wave interaction between drift electrons and acoustic waves in a solid such as cadmium sulfide. The so-called Gunn effect has been utilized to produce microwave oscillations in bulk cadmium sulfide and, in addition, has led to the development of small, compact, robust microwave oscillators.

SPECIAL CIRCUITS

Mixer-modulators, frequency-changers, multiplexers. The mixing and separation of two or more signals or carriers is a fundamental requirement in radio (*q.v.*) and communication systems and in sound recording and reproduction (*q.v.*).

In the reception of radio signals with a heterodyne receiver, the incoming signal is mixed with the signal from a local oscillator to produce the intermediate-frequency signal. In vacuum-tube receivers special tubes are used to perform the mixing function. Frequency-converter tubes, widely used before transistor development, perform this function by including the local oscillator in the tube envelope. In transistor versions of these circuits the mixer or frequency changer may be one of two types. It may consist of a separately excited mixer requiring a second transistor as a local oscillator; or it may be a self-oscillating mixer in which a single transistor combines the functions of mixer and oscillator. The main advantage of the self-oscillating type is that it is cheaper.

Multiplexers are of two types: frequency-division multiplexers and time-division multiplexers. The first are designed to transmit two or more signals over a common path by using a different frequency band for each signal. Each channel is identified with a separate subcarrier frequency and each message modulates the subcarrier identified with its channel. At the receiver terminal the channels are separated by frequency selection and the subcarrier is filtered out, restoring the original message signal.

In time-division multiplexing (see TELEPHONE AND TELECOMMUNICATIONS SYSTEMS) different time intervals are used for different signals. The timing and channel operations may be obtained by a multivibrator ring circuit containing one modulator per channel, and all modulators are connected in tandem and fired by a triggering pulse.

Limiting and clamping circuits. A frequent requirement is to prevent a particular characteristic of the output of a circuit from exceeding a predetermined value, or to hold it at that value. Limiting and clamping circuits are used for these functions.

The widely used so-called Schmitt limiter or trigger is a pulse generator in which a pair of transistors with positive feedback produces an output pulse of constant amplitude as long as the input voltage exceeds a certain value.

One of the simplest limiter circuits is of the so-called neon-lamp type, which consists of a series resistor and a neon lamp. The output taken across the lamp never exceeds the breakdown voltage, so that the instantaneous values of amplitude above this value are clipped.

Figure 6 shows another simple example of the use of transistors and semiconductor diodes in chopping and limiting applications.

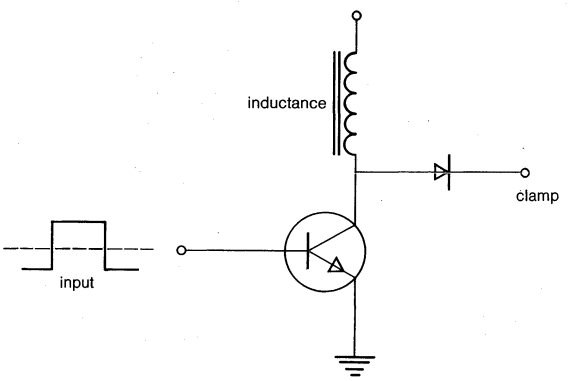


Figure 6: Transistor chopper.

Wave-shaping circuits. Circuits that generate pulses of current or voltage are of two basic types: (1) the self-contained generator, which includes a pulse oscillator; and (2)) one that acts on an input signal to produce a train of pulses at the output. A square-wave generator is useful in testing the response of video amplifiers and in switching applications in computers and telemetering testing. Sawtooth and triangular pulses can be generated by relaxation oscillators fed from a constant-current transistor.

BIBLIOGRAPHY. General reference works on electronics include C. SUSSKIND, *The Encyclopedia of Electronics* (1962); S. HANDEL, *A Dictionary of Electronics*, 3rd ed. (1971); L.E.C. HUGHES and F.W. HOLLAND (eds.), *Electronic Engineers' Reference Book*, 3rd rev. ed. (1967); and G.W.A. DUMMER and J.M. ROBERTSON, *The Electronics Book List* (1964). A classic work in the field is SIR J.J. THOMSON and G.P. THOMSON, *Conduction of Electricity Through Gases*, 3rd ed., 2 vol. (1928–33, reprinted 1969). See also G.W.A. DUMMER, *Modern*

Electronic Components, 2nd ed. (1966); E.J. ANGELO, *Electronic Circuits*, 2nd ed. (1964); J.N. SHIVE, *The Properties, Physics and Design of Semiconductor Devices* (1959); and D.B. LANGMUIR and W.D. HERSHBERGER (eds.), *Foundations of Future Electronics* (1961).

(S.H.)

Electron Tube

An electron tube is a device, usually in a sealed glass or metal enclosure, that generates and controls a stream of electrons in such a manner as to perform a useful function in an electronic circuit. This function can be the amplification of a very weak radio or audio signal, the generation of electrical oscillations such as audio tones or radio frequencies, the production of X-rays, or numerous other functions that will be discussed in the course of this article. Electron tubes include both vacuum tubes, in which the air or gas has been exhausted from the enclosure, and gas tubes, in which a small amount of gas has been introduced into the enclosure.

This article discusses the basic operating principles of electron tubes, the history of their development, how they are constructed and manufactured, and some of their applications.

BASIC TUBE OPERATION

All types of tubes operate under the same basic conditions, which can best be understood by looking first at a diode, a two-terminal device that conducts electricity more easily in one direction than in the other: voltage is impressed on the tube, normally between the plate (or anode) and cathode (see Figure 1A). If the plate is at a positive (+) voltage (potential) with respect to the cathode, electrons emitted by the cathode will be attracted to the plate and current will flow. If it is at a negative (—) potential, no current will flow. The diode thus serves as a sort of valve (the name given it in the U.K. and some other countries), permitting current to flow when the plate is positive and blocking current when the plate is negative. It can thus convert ac (alternating current, a flow of electricity that reaches maximum in one direction, decreases to zero, then reverses itself and reaches maximum in the opposite direction) to dc (direct current, current that flows in only one direction), a useful characteristic in many electronic circuits, as will be seen later.

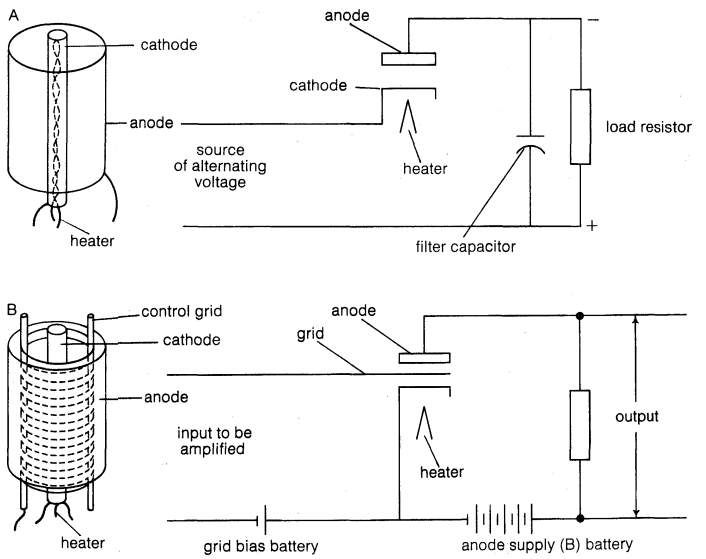


Figure 1: (A) Internal parts and schematic rectifier circuit of thermionic diode. (B) Internal parts and schematic amplifier circuit of thermionic triode.

If a third element in the form of a fine metal mesh or screen, called the control grid, is inserted between the cathode and plate, it can be used to control the flow of electrons (see Figure 1B). The resulting tube is termed a triode and is useful as an oscillator or amplifier. Other grids may be added for various operating refinements, producing what are called multi-element tubes. Tubes of this nature generally operate with a high vacuum. Gas-filled tubes are usually used as switches; for the tube to pass current (switch on), conditions must be such that the gas is ionized. When not ionized, current will not flow (switch off). Specialized tubes may use the stream of electrons for other purposes, as for the generation of X-rays in the X-ray tube, or to produce a visual display in the television picture tube.

DEVELOPMENT OF THE ELECTRON TUBE

Sir William Crookes, an English physicist of the mid-19th century, was the earliest pioneer of the vacuum tube when he experimented with the passing of high voltages across electrodes in an evacuated tube. Electron theory developed from his research with what came to be known as Crookes' tube. The Röntgen X-ray tube followed in 1895, and the cathode-ray tube in 1897. Thomas Edison also contributed to the vacuum tube, though he was not aware of it at the time. Working with the light bulb in 1883, Edison noticed that if he placed a metal plate inside the bulb and connected a positive voltage to the plate, current would flow to it when the lamp was lit. He also observed that if the plate was negative, current would not flow. Edison did not understand this effect and did not think it important, though he did record it in his journal. Some 16 years later, Sir J.J. Thomson of England experimented with the Edison effect, as it was called, and described what Edison had discovered: that current can be made to flow between two electrodes in a vacuum if one of the two is heated. He established that the particles in these experiments were indeed electrons, of a definite electric charge and mass, irrespective of their origin or the methods for their release.

Using the theory that Thomson developed, another Englishman, J.A. Fleming, constructed a workable, practical diode tube. His name for the diode—"valve"—came from his observation that the diode "acted as a valve, permitting current to flow or be shut off."

Development of the triode

Shortly thereafter, Lee De Forest of the U.S., working with Fleming's valve, sought to develop a tube that in addition to passing current in only one direction would amplify a small electrical signal voltage. His investigations led to the triode tube, consisting, as the name implies, of three elements. The third element, which De Forest called a grid, could, when placed between filament and plate, control the current flow through the tube. This single development may have had as much to do with the growth of electronics as any other discovery or invention. For the first time, signals could be amplified. De Forest originally called his tube the Audion, a name that soon fell into disuse.

The term filament, originally used to describe the emitter of electrons, took on a new meaning shortly after the development of the first tubes. It was found that some materials emit electrons more readily than others. Also, it was discovered that two different elements could be used, one to emit electrons (the cathode) and the other to supply the heat (the filament or heater).

MODERN TUBES

Electron emitters. In most modern tubes, electrons are produced by thermionic emission from a cathode. This is the name given to the process in which a suitable metal such as tungsten is heated to a high temperature. As the electrons in the heated metal become more excited, they acquire sufficient energy to escape from the metal surface. When pure tungsten is used as the cathode, it is heated to about 2,200° C (3,992° F) to emit electrons.

Cathode materials

Other materials besides tungsten can be used as electron emitters. With thoriated tungsten (tungsten with a thin layer of thorium on the surface), emission takes place at about 1,700° C (3,092° F). Such materials as strontium and barium oxides, used as coatings, begin to emit electrons at still lower temperatures, some as low as 700° C (1,300° F). These lower temperatures give higher efficiency.

In applications such as receiving tubes in radio and television sets, where large plate currents are not required,

oxide-coated cathodes are used almost exclusively. A separate filament heats the cathode, which usually consists of a nickel sleeve coated with oxides. There is no electrical connection between the two. This is called an indirectly heated cathode. When a filament serves as the cathode it is called a directly heated cathode.

One of the big advantages of the indirectly heated cathode is that the filament can be connected to an alternating-current source without introducing hum, as would happen if directly heated filaments were used. In a transmitter tube and other high-powered tubes, directly heated cathodes of tungsten are necessary, since oxides simply cannot produce the large quantities of electrons required.

Another type of cathode emitter is the photocathode used in photoelectric diode tubes (see Figure 2). The

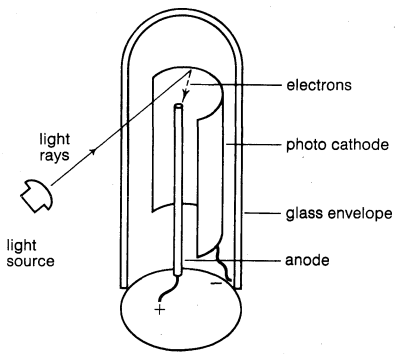


Figure 2: Basic photoelectric diode tube.

photocathode is coated with a material such as the element cesium that gives off electrons when light strikes it. No heat is required; such tubes are thus classed as cold-cathode tubes. Photoelectric cells or tubes are used for detecting or measuring light, as in light-beam burglar-alarm systems and in devices that measure light intensities. Cold cathodes are also employed in a special type called a field effect tube, in which a strong electric field draws electrons from the cathode. This type of tube is not very common; it was used as a rectifier tube in some of the very early ac-operated radio receivers.

Multi-element construction. Following the diode and triode, further developments led to the tetrode tube (four elements), in which a screen grid is inserted between the control grid and plate, and then the pentode or five-element tube, in which a suppressor grid is added between the screen grid and plate. Each of these tube types has its own advantages and disadvantages.

Tetrode. The tetrode has a major advantage over the triode: its screen grid serves as a shield between the control grid and plate, minimizing interaction between them and permitting much higher amplification than is possible from the triode. It has two disadvantages, however: it reduces the useful plate current when the plate and the screen grid are at nearly the same voltage, and it results in an increase in the effects of secondary emission. This is the term describing electrons that are knocked off the plate by an incoming electron striking it at high velocity. In the case of the triode, secondary emission is no particular problem, since electrons that are knocked off the plate simply return to it because of its high positive potential; but in the tetrode, the screen grid is nearby and also operating at a high positive potential, so that many of these secondary electrons are attracted to the screen grid and flow in its circuit, rather than in the plate circuit, where the output should flow for greatest circuit efficiency. Primarily for this reason, tetrode tubes are now seldom used, except in some transmitters. The high amplification factor and the lower plate-to-grid capacitance make the tetrode particularly useful as an amplifier of radio-frequency signals.

Pentode. The pentode tube represents a major improvement (see Figure 3). The pentode's fifth element, the suppressor grid, is at a negative voltage with respect to the plate and screen grid; therefore, any electrons that are knocked off the plate by secondary emission are repelled by this grid, forcing them to return to the plate. A few of the secondary-emission electrons may pass through the suppressor grid, but the numbers are so small that they do not affect the operation of the tube. Pentodes are used mainly as rf voltage amplifiers in radio and TV receivers. They are also suitable as power amplifiers, having greater amplification than triodes.

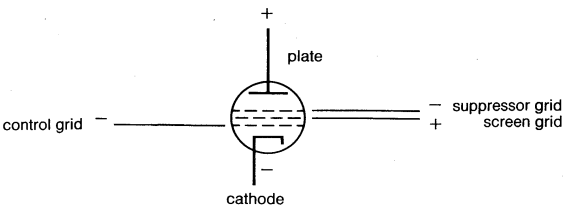


Figure 3: Pentode tube.

Function of the suppressor grid

Special tube types. *Multisection tubes.* In addition to the tubes described, electronic devices sometimes use other highly specialized tubes. Such is the pentagrid (five-grid) converter tube, used in most low-cost superheterodyne tube radio receivers as a combination oscillator-mixer and first detector—two important functions (see RADIO)—thus avoiding the necessity for two separate tubes and keeping the associated circuitry simple. In the superheterodyne receiver, the incoming modulated signals are amplified and then mixed (heterodyned) with an oscillator signal generated in the receiver. The frequency of this local signal is controlled by the tuning control so that it differs from the incoming signal frequency by a fixed amount, called the intermediate frequency. This intermediate frequency is further amplified and then detected (rectified) to recover the intelligence (audio, television, etc.) on the incoming signal.

Many versions of multisection tubes—more than one set of elements in the same physical envelope—have been developed, including dual triodes (two triodes in one envelope) and triode-pentodes. Some tubes, often called compactrons, have been introduced, though not widely used, containing three triodes in one envelope, or several pentodes and a triode, and other combinations.

Other specialized tubes, not used in receiving equipment, include high-power tubes used in transmitters. These must be cooled, either by air or water. High-power transmitting types that require water to keep them cool use metal envelopes. Openings allow the water to circulate around the outside of an inner envelope and out the other side of the tube. Such tubes have a double envelope, the inner containing a vacuum and housing the elements of the tube, the outer permitting the water to enter and circulate.

In the X-ray tube, another specialized type, a target of a heavy metal such as tungsten is bombarded by electrons and produces X-rays, which are used for X-raying teeth, broken bones, and the like.

Klystron and magnetron. The end of the 1960s saw a large increase in the number of special television, high-frequency, and high-power tubes, for example, the klystron and magnetron. These tubes operate on principles different from those of ordinary tubes. They are most useful for frequencies in the microwave region (above 1,000 megahertz, or 1,000,000,000 cycles per second), where they are capable of producing high output power. Both magnetrons and klystron tubes generate a very narrow bank of frequencies because the built-in cavity resonator acts as a tuned circuit for the operation of the tube; it can thus serve as a frequency control device. Regular tubes have external resonant (tuned) circuits so that the frequency can easily be changed. The size of the magnetron cavity establishes the resonant frequency.

The klystron, used to generate and amplify microwave signals for high-frequency communications systems, does permit some small adjustment of frequency. It contains an electron gun—a cylindrical electrode that produces a narrow beam of electrons—in place of a regular cathode (see Figure 4). The input radio frequency is connected to the input cavity and the output is taken from the output cavity. A beam of electrons, forced through the cavities

by an accelerating voltage, is allowed to drift in the space between the cavities. The input radio frequency produces a varying voltage on the first cavity, thus speeding up some electrons and slowing down others. This action produces a bunching effect in the drift space; the bunched electrons at the second cavity produce an amplified version of the radio-frequency input to the first cavity. When used as a generator or oscillator, some of the output signal is fed back to the input (positive feedback) in such a manner as to sustain oscillations.

From *Basic Electronics* by Glade Wilcox. Copyright © 1960 by Glade Wilcox. Reprinted by permission of Holt, Rinehart and Winston, Inc.

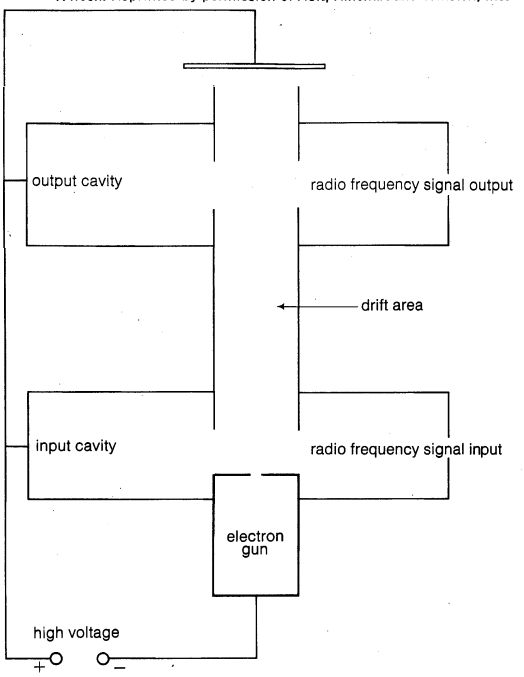


Figure 4: The klystron tube.

Although, for simplicity, the illustration shows only two cavities, it is possible to have a number of them—as many as six in a single klystron tube. Amplification is increased with extra cavity resonators, along with cost, size, and complexity.

The ignitron. Special tubes called ignitrons have been developed for use by industry as rectifiers for high-power applications, such as converting ac to dc for electric locomotives (see Figure 5). The ignitron is a mercury-pool rectifier in which the initial ionization is provided by an arc started between a special electrode, called the igniter, and a liquid pool of mercury, the cathode. The igniter must be used to start the arc, so that the tube can be controlled by small pulses of current to the igniter and hundreds of amperes can be controlled by a small input. The heavy currents are carried by the arc discharge through the ionized mercury, rather than by thermionic emission as in most of the lower power devices.

Other gas tubes, such as gas-filled diodes, are on a smaller scale. These tubes are similar to regular vacuum-tube diodes used in receiving equipment except that after the air has been evacuated, a small amount of mercury vapour is allowed to enter the envelope. Ordinarily a filament (to keep mercury vapour from condensing) and a plate are used, as in a regular diode tube. In these smaller gas diodes, ions of the gas, rather than electrons, carry the current.

Gas-tube diodes come in many sizes, from small receiver types to large types used in power supplies for radio and television. The type used is determined by the power requirements. In addition to handling heavier current, the gas-tube diode has an internal resistance lower than that of a vacuum tube; that is, less power is wasted within the tube itself.

The thyratron. Another type of gas tube is the thyratron. Basically this is a triode or pentode tube, with a

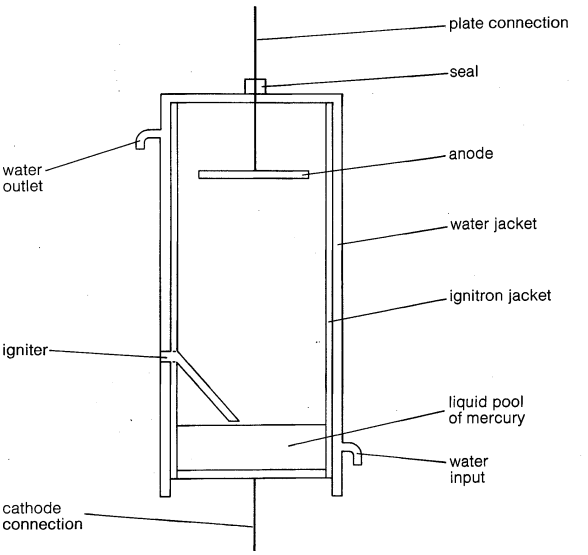


Figure 5: Cutaway view of ignitron construction (see text).
From *Basic Electronics* by Glade Wilcox. Copyright © 1960 by Glade Wilcox. Reprinted by permission of Holt, Rinehart and Winston, Inc.

small amount of gas added. Though a triode gas tube has a cathode, control grid, and plate, the control grid does not exercise the same control as it does in a vacuum-tube triode. Once the control grid of the gas-filled triode allows current to flow, the tube conducts fully, regardless of any change in the voltage on the control grid. The only way current flow may be stopped is by reducing the positive voltage on the plate to near zero. For this reason, thyratrons and other gas-filled tubes are operated with alternating current rather than direct current on the plate. The alternating current shuts the tube off at the end of the positive half-cycle of each alternation. If the grid potential is adjusted so that the tube conducts as soon as the plate becomes positive, the tube conducts for approximately one-half, the positive half, of the input cycle.

If conduction for a lesser period of time is desired to satisfy a circuit requirement, a higher potential is impressed on the control grid, permitting conduction to take place only with a higher plate potential. The tube stops conducting when the positive half-cycle on the plate has passed, as before. In practice, it is possible to operate the tube so that it conducts from only a tiny fraction to almost half of its input cycle, thus providing a widely varying current for such applications as dc motor speed control. The grid can be operated with alternating voltage on it as well as on the plate. The phase of the grid voltage can be made adjustable so that the conduction time of the tube can be predetermined and the desired output current obtained.

Thyratrons are often used in arc welding equipment. Through their use, precise amounts of heat can be produced at the weld. If more heat is needed the thyratron can be adjusted to conduct for a longer portion of the positive half-cycle. When currents are required that are greater than the rating of a particular thyratron, larger devices can be used, controlled by the smaller.

Television tubes. With the advance of television in the 1960s, many tubes were designed specifically for television receivers, such as tubes for the rf amplifier. One such tube, only a quarter of an inch in diameter, is capable of excellent low noise amplification.

Another special tube utilizing the phenomenon of secondary emission to provide large amplification in television camera tubes and other applications is the photomultiplier (see Figure 6). Basically a phototube, the photomultiplier amplifies the signal produced at the photocathode up to 10,000,000 or more times.

Photomultiplier utilizes secondary emission

This tube operates through the use of a number of small plates called dynodes. Each dynode, starting at the cathode, is operated at an increasingly higher positive voltage with the collector (or final plate) the highest. Electrons from the cathode strike the first dynode plate, knocking

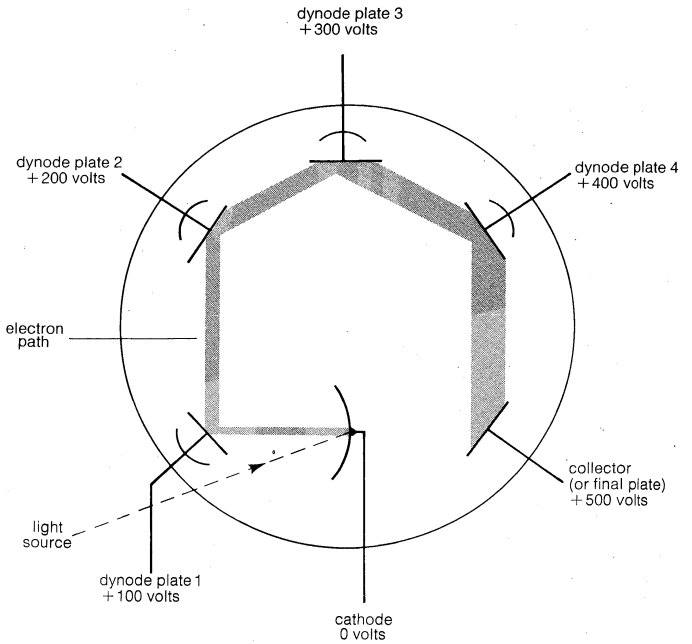


Figure 6: The photomultiplier tube (see text).
From *Basic Electronics* by Glade Wilcox. Copyright © 1960 by Glade Wilcox. Reprinted by permission of Holt, Rinehart and Winston, Inc.

out electrons by secondary emission. The plate material and circuit conditions are selected so that more electrons are knocked out at the plate than strike it, thus producing amplification. These electrons are attracted to the second dynode plate, where they knock out even more electrons, which are attracted to the next dynode plate, and so on until the final plate is reached.

Such tubes are particularly useful for the measurement of very small light intensities. Light so dim that it cannot be seen with the unaided eye can be measured using these tubes and associated equipment. They are available with various colour band sensitivities, from ultraviolet to infrared, so that any particular spectrum may be investigated by the researcher. Photomultiplier tubes are also often used in nuclear radiation-measurement equipment in which the radiation particle produces a very small flash of light or scintillation in a special type of crystal. The photomultiplier tube can be used to detect this light and amplify the resulting signal to a useful electronic pulse that can be recorded, counted, and analyzed to determine the characteristics of the material producing the radiation.

Electron-gun tubes. A very important group of tubes are those that belong to the electron-gun family. Included are cathode-ray tubes used in oscilloscopes, which display varying electrical voltages and currents as a function of time, and picture tubes used in television sets. An unusual type is a tube with a phosphor—the layer of luminescent material applied to the inner face of the tube—that has a long persistence; that is, it retains an image, or glow, a long time. Some of these long-persistence tubes are used in radar, others in computer memory units.

All cathode-ray tubes use an electron gun. Basically the gun resembles a regular electron tube with cathode, control grid, and several other elements, but without a plate. In the cathode-ray tube, however, the entire gun is used to accelerate the electrons from the cathode and focus them into a narrow stream. They leave the gun in just this form and end up striking a very small spot on the face or screen of the cathode-ray tube.

A major difference between the cathode-ray tubes used in oscilloscopes and those used in television sets is the way in which the electron stream is deflected to different areas on the screen. The oscilloscope type of cathode-ray tube almost always uses electrostatic deflection, a process in which two pairs of parallel metal plates are placed inside the cathode-ray tube, one pair horizontal and the other vertical. The electron beam passes between these plates and is deflected by any difference in potential between the two plates of each pair. In this way the beam can be moved in any direction so that it hits the desired spot on the face of the tube.

Television picture tubes use electromagnetic deflection; that is, the stream of electrons leaving the electron gun is deflected by a magnetic field derived from a horizontal and a vertical deflection coil, called a deflection yoke, positioned around the neck of the picture tube. Through the use of properly shaped voltage pulses to the yoke, the beam is deflected to the proper position on the screen.

Phosphor light colours

There are two differences between cathode-ray tubes and picture tubes. One is the colour of light produced by the phosphor, which in black-and-white picture tubes is white. In oscilloscopes, the phosphor generally produces green light, though other colours are sometimes obtained. The second difference is that of persistence mentioned above. Most, but not all, cathode-ray tubes for oscilloscopes have a short persistence. Television picture tubes have a longer persistence.

The colour picture tube has three different colour phosphors, green, blue, and red, arranged on the face of the picture tube in the form of tiny dots, placed with one colour at each corner of an imaginary triangle. Three electron guns, one for each of the three colours, shoot electron beams at the proper phosphors.

TUBE CONSTRUCTION AND MANUFACTURE

Common vacuum tubes are made up of a variety of parts and elements. The grid or grids are made on a special grid-winding machine. Cathodes are made of many different materials and combinations. Filaments are also made as a separate part of the tube, generally of fine tungsten wire. Similarly, the plate, or anode, of the tube is made of a suitable material, depending on the application.

Manufacturing of electron tubes is highly automated. The various elements are assembled with micra or other insulating material to provide proper spacing and rigidity, for the distance and tolerance between elements are often only a few thousandths of an inch. A jig holds all elements in place until they are ready to be placed inside the envelope. Often heat is applied to drive off gases from the finished product.

Removing residual gas

A getter—an alkali metal that reacts chemically with and eliminates any remaining gases—is also placed inside the tube. The elements are spot-welded to suitable pins that serve as connections through a header on the outside of the tube.

It is necessary that the base permit the pins to pass through it, allow no air to enter, and provide a good seal. For many years the manufacturing problem was extremely difficult, but today's methods of very close heat control, excellent tolerance control, special base glues, and automated equipment have brought the problem under control.

After the elements have been spot-welded to connection pins, the entire inner electrode assembly is heated and placed inside the envelope, which usually has an extended tubelike appendage at the top, bottom, or side, through which air is pumped out. This appendage is then heated and pinched off, making a permanent, vacuum-tight seal. The getter, if used, is then flashed (ignited) to combine with any remaining gases and produce a good vacuum.

A getter is not always used, for in very high frequency tubes it may increase the interelectrode capacities between the various elements and degrade tube performance. It is also often omitted in high-power tubes to prevent a short developing between elements.

During the entire manufacturing process, great care is exercised to keep the elements of the tube from being contaminated by dust, dirt, or chemicals. Gloves are used by personnel handling the internal parts of the vacuum tube.

The manufacture of large, high-powered tubes, or the special klystrons and other ultrahigh frequency devices, is not as automated. Much handwork is still necessary in the proper fitting and assembly of the parts.

Rugged-ized tubes for special needs

Several tube manufacturers make special versions that are rated as "long life," or "premium," tubes, most often used in applications in which breakdowns are likely, such as in the amplifier in an undersea cable, in a remote, unattended radio relay station, or in an industrial or military application where they are subjected to extreme shock and vibration. Apart from their ruggedized character and extreme care in manufacture, these premium tubes are often much like the regular versions of the same device. In some cases, the heater or filament is specially designed to be longer lasting than the regular version.

In the basing of tubes, many different forms are used. Early tubes used a screw-in base, but since about 1930, tubes, except for high-power and high-frequency types, have generally been of the plug-in type. The most common bases today are the seven- and nine-pin miniature types, which have small wires emerging from the glass envelope through a glass base. These small pins enter a socket that serves to hold the tube in place, to permit connections to be made, and to allow it to be easily replaced.

Other types of bases include the earlier octal basing, which was a standard eight-pin socket. The tube was placed in a bakelite plug, with fine wires emerging from the base of the glass envelope. These wires were placed in metal prongs on the bakelite socket and soldered in place. The bottom of the glass envelope was then glued to the bakelite. Many television picture tubes still use this type of basing, though new equipment is no longer designed for its use.

Special receiving tubes may have different types of base. Materials are used that do not absorb energy at high frequencies to minimize losses.

APPLICATIONS AND FUTURE PROSPECTS

From its beginnings, the electronics industry has built its devices around the versatile electron tube, whose applications today assume staggering proportions. Low-power tubes are found as rectifiers, amplifiers, oscillators, and in other special applications. The great majority of television sets are still exclusively tube types. The average colour television set uses 20 to 30 tubes, the number often being a measure of the quality of the set.

Radio transmitters of broadcast quality employ large numbers of electron tubes, serving basically the same function as they do in the receiving equipment. In fact, many of the smaller, regular tubes are used in transmitters as well as receivers. In addition, transmitters utilize high-powered amplifier tubes (generally tetrodes and pentodes) to provide the necessary power output. Gas-filled diodes are commonly used as rectifiers to provide the high-voltage, high-current dc required by the high-power amplifier stages.

High-power and special tubes find applications in many other devices, such as industrial controls, radar, computers, radio-frequency heating, weather equipment, and automation of factories.

The transistor (see SEMICONDUCTOR DEVICES) has in many cases changed the nature of electronic equipment, such as the computer, enabling it to be made much smaller and with much less heat dissipation. Transistorized radios are now made in enormous numbers. Yet the "end of the tube," once thought to be in sight, now seems further off. The transistor has been best able to supersede the electron tube in applications where small power is required, small size is imperative or desirable, or in applications in which the transistor can do the job more cheaply.

The transistor versus the tube

The time will no doubt come when the tube-operated television set will be a thing of the past. Special problems exist, however, in some of the television circuitry that makes transistor application difficult. All-transistor colour sets have achieved some success, however, in the early 1970s.

In applications in which higher powers are required, transistor inroads have been slow for many reasons. In high-frequency applications, very low capacities and inductances are required. Transistors, when fabricated to meet this requirement, are expensive. When high powers are needed at high frequencies, as in transmitters, transistors are inadequate because they cannot provide sufficient power while operating in the upper frequency range. Transmitters that have lower power outputs, such as those used in mobile equipment, have transistorized receivers (for low battery drain and dependability) but often have tube type transmitters, or a combination of tubes and transistors, with tubes doing the high-power work.

On the other hand there has been great progress in solid-state diodes. The tube rectifier is rarely used in receiving equipment any longer. Even in high-current applications, solid-state diodes are supplanting the vacuum tube. It is now conceivable that solid-state diodes will completely take over the diode rectifier tube market in the near future.

Yet the impact of transistors and other solid-state devices on electron tubes has not been nearly as great as predicted. Much of the reason for the continued health of the electron tube lies in the great increase in black-and-white and colour television sets all over the world. Where old-fashioned radio receivers used only 5 or sometimes 6 tubes, the television set, even black-and-white, uses 12 to 15 tubes, with colour sets using 20 to 30 tubes. Stereophonic radios, tape recorders, and other devices also continue to create a market for tubes, though the transistor is rapidly taking over.

The market for tubes

In the case of the klystron, magnetron, and other special-purpose tubes, no real challenge has yet been forthcoming from transistors. The nature of these tubes and their uses make competition from the transistor unlikely for some time. Transistors available today can operate at frequencies of several hundred megahertz (million cycles), but klystrons operate at 20,000 to 30,000 megahertz. The tube designed specificially for this operation promises to remain unchallenged for some time.

BIBLIOGRAPHY. *Reference Data for Radio Engineers*, 5th ed. (1968), a complete volume on vacuum tubes, including design information, practical circuit design and operation; F.E. TERMAN, *Radio Engineers' Handbook* (1943), the classic text on electron tube theory from a mathematical and theoretical standpoint, as used by engineers; D.A. SMITH, *ABC's of Vacuum Tubes* (1967), a comprehensive text on the vacuum tube.

(D.A.S.)

Electroplating

Electroplating is the act of coating with metal by means of electric current. Plating metal may be transferred to conductive surfaces (metals) or to nonconductive surfaces (plastics, wood, leather), after the latter have been rendered conductive by such processes as coating with graphite, conductive lacquer, electroless plate, or a vaporized coating.

While some metal coating procedures date back to ancient times, modern electroplating started in 1800 with A. Volta's discovery of the voltaic pile, or battery, which made noteworthy quantities of direct current electricity available. At about the same time, the battery was employed to deposit lead, copper, and silver. After a nodule of copper had been deposited on a silver cathode, the copper could not be removed. In the same year, zinc, copper, and silver were deposited on themselves and on a variety of basis metals (the metals on which the plating is applied), such as gold and iron.

Figure 1 shows a typical plating tank containing copper sulfate ($CuSO_4$) solution. A dynamo supplies electric current which is controlled by a rheostat. When the switch is closed, the cathode bar, which holds the work to be plated, is charged negatively. Some of the electrons from the cathode bar transfer to the positively charged copper ions (Cu^{++}), setting them free as atoms of copper metal. These copper atoms take their place on the cathode surface, copper plating it.

Concurrently, as shown in the drawing, the same number of sulfate ions (SO_4^{2-}) are discharged on the copper anodes, thereby completing the electrical circuit. In so doing, they form a new quantity of copper sulfate that dissolves in the solution and restores it to its original

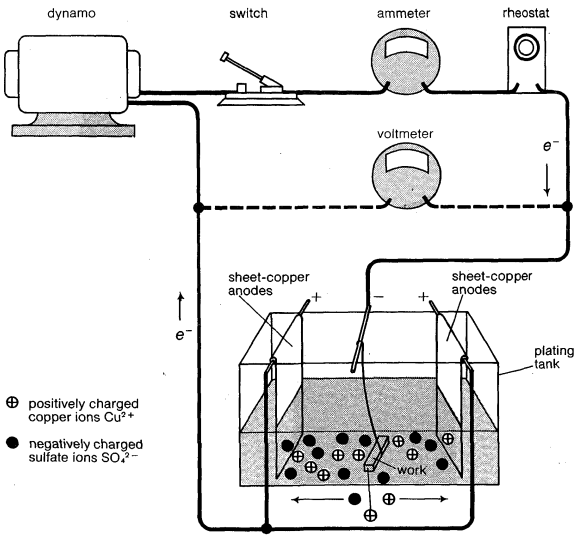


Figure 1: Electroplating circuit.

composition. This procedure is typical of nearly all ordinary electroplating processes: the current deposits a given amount of metal on the cathode and the anode dissolves to the same extent, maintaining the solution more or less uniformly. If this balance is perfect and there are no side reactions or losses, a 100 percent cathode efficiency and 100 percent anode efficiency could possibly be realized.

Cathode and anode efficiency

If the metal surface of the cathode is chemically and physically clean, the discharged atoms of copper are deposited within normal interatomic spacing of the atoms of the basis metal and attempt to become an integral part of it. In fact, if the basis metal is copper, the new copper atoms will frequently arrange themselves to continue the crystal structure of the basis metal, the plate becoming more or less indistinguishable from and inseparable from the basis metal.

Cleaning and pickling. If the surface of the basis metal is contaminated by a film of grease, oil, rust, or oxide, or even if the surface is strained or smeared by cold working, the plating will cover these intervening materials. The adhesion then may vary widely from completely adherent to completely nonadherent. It is thus necessary to clean the basis metal perfectly before plating. Films of oil and grease are removed by such methods as vapour degreasing, hot-alkaline cleaning (soak, cathodic, or anodic), or scrubbing with pumice. Rust and oxide films are removed by acid dipping or pickling (soaking in acid), and, if possible, a thin layer of metal is dissolved away by etching to remove surface strains and expose a fresh surface.

Plating processes. If suitable solutions of different metals are mixed, it is possible to plate a wide variety of alloys of metals. By this means plated brass can be made more or less indistinguishable from cast brass. It is also possible, however, to deposit alloys or compounds of metals that cannot be produced by melting and casting them together. For example, tin–nickel alloy plate has been used commercially for its hardness and corrosion resistance, which are superior to that of either metal alone. The deposit consists of a tin–nickel compound (Sn-Ni), that cannot be produced in any other way.

Other common alloy plates include bronze, and gold, with varying properties, such as different colours or hardnesses. Magnetic alloy plates of such metals as iron, cobalt, and nickel are used for memory drums in computers. Solder plate (Sn-Pb) is used in printed circuit work.

Electroplating on a commercial scale was begun around 1840–41 and was accelerated by the discovery of cyanide solutions for plating silver, gold, copper, and brass. A cyanide–copper solution, for example, gave adherent deposits of copper directly on iron and steel. It is still used for this purpose and also for the initial plating on zinc die castings. The copper sulfate solution described above corrodes these metals, giving nonadherent deposits.

Nickel plating

Nickel plating on a commercial scale was introduced in Boston, in 1869. It required pure anodes and salts and a neutral or slightly acid nickel solution. Previous workers had used alkaline nickel solutions in which the nickel anodes did not dissolve properly and which provided only thin films of metal. Nickel plating achieved immediate popularity, and an important American railroad became known as the Nickel Plate Road as a symbol of luxury and efficiency.

While there were many attempts to discover a dependable chromium plating process, they were unsuccessful until 1924. Most of the early work had involved the simple trivalent and bivalent salts of chromium. It took a long time to discover dependable and practical conditions for chromium plating from chromic acid solutions, with insoluble lead anodes and a catalyst of sulfate ion equal to 1 percent by weight of the chromic acid. In this case, the solution is maintained by chromic acid additions and not by means of soluble anodes.

Many innovations were introduced into electroplating practice around the turn of the century. Electrolytic cleaning and the bulk plating of small parts in tumbling barrels were introduced. Zinc plating iron parts for rust protection gradually established itself, followed by cadmium plating for the same purpose around 1920. Tin plating was well known from the middle of the 19th century onward, but it was not until the World War II period that both acid and alkaline baths came to be used on a large scale for the plating of steel strip for tin cans, thus economizing on the amount of tin required compared to hot tinning.

Electroplating is a large and growing industry with sophisticated engineering and equipment requirements. The metals that can be readily plated from aqueous solutions at high current efficiencies near 100 percent can best be surveyed from Figure 2. It shows these metals in a single

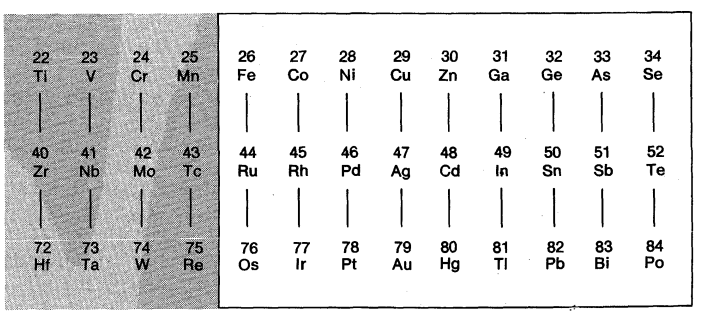


Figure 2: Periodic chart of metals readily plated.

rectangle in their proper relationship to each other. The only metal in common use outside of the rectangle is chromium, which is commonly plated at low current efficiencies of about 10–20 percent. Of the 27 metals inside the rectangle, about 15 are more or less common in plating, specifically, iron, cobalt, nickel, copper, zinc, ruthenium, rhodium, palladium, silver, cadmium, tin, iridium, platinum, gold, and lead. The others can be deposited easily but have not found much use in this way either due to cost or availability or lack of useful properties.

A few metals outside of the rectangle are sometimes used for plating or electroforming purposes. Thus aluminum is frequently deposited in this way from organic or nonaqueous solutions and also fused salt baths. Some of the refractory metals (molybdenum, tungsten, tantalum, and niobium) have also been applied in this way by workers using fused (melted) salt baths.

To obtain a smooth, bright finish it was formerly necessary to polish and buff both the basis metal and the plate. The cost of this processing was generally as much as or more than the cost of plating. The trend has accordingly been to the constantly increasing use of bright plating solutions and even "levelling" deposits (especially of cop-

per and nickel) that actually fill in scratches and other defects in the basis metal and make possible the production of bright finished articles with little or no buffing or polishing. While occasional use was made of bright plating solutions after the initial use of carbon disulfide to brighten silver plate around 1847, it was only with the introduction of bright nickel plating in the 1930s that the period of intensive developments in this direction began. Another development giving similar results, particularly with copper, was the use of periodic reverse current for electroplating. Intensive work on "levelled" deposits dates from the introduction of semibright nickel plating in the early 1950s.

Bright plating solutions

The introduction of chromium plating in 1925 had stimulating repercussions all through the plating industry. Chromium was essentially a bright plate and retained its brightness indefinitely. Chromium plate found a ready market in the automotive and appliance fields, in which the merits of the combination plate nickel–chromium or copper–nickel–chromium were soon proven. The requirements for closer control procedures in bath composition, temperature, and current density were reflected in better control and development of other processes.

The so-called hard chromium plating likewise created a new way of improving wear resistance of machine parts and improved operation due to good frictional and heat resistance properties. Worn or undersized parts were built up with chromium plate. The success of chromium plating increased the field for engineering applications of other metals such as nickel, copper, silver, and gold.

The Manhattan Project for the production of the atomic bomb in World War II could not have succeeded without a durable nickel plate lining in the pipes and valves for handling the uranium hexafluoride vapours. Nickel was required and there simply was not enough nickel available to make solid nickel pipes.

The success of hard chromium plating is partly due to several intrinsic properties of the process. If the solution is heated to the proper temperature of about 50° C (122° F) or more, it does a great deal of its own cleaning and produces deposits of a very high degree of adhesion under almost all conditions. The chromium deposits also "heal" torn places and machining defects in the surface of the basis metal to which they are applied, due to excellent smoothing or levelling ability. Part of the decreased friction on these deposits results from this property.

The low current efficiency of chromium plating baths was nearly doubled (though it remained in the range of 15 to 25 percent) by the use of combined catalysts of both sulfate and complex fluoride ions (for example, silicofluoride, SiF_6^{2-}) in self-regulating high speed (SRHS) baths described in 1950. These baths also gave harder and brighter chromium deposits of increased utility.

Chromium plating has been unique in the extremely wide variety of deposits that are produced for different purposes by altering the bath composition, temperature, current density, and post treatment of the deposit. Ordinary deposits are cracked unless extremely thin (less than 0.00002 inch; 0.000508 millimetre) but are completely adherent, and the cracks heal and fill in completely as deposition proceeds, while new cracks form on the surface. Use is currently being made of this pattern of extremely fine cracks to obtain decorative plates of improved corrosion resistance, described as microcracked or microporous chromium plate.

Microporous chromium plate

A dull, crack-free chromium plate was introduced in 1953, and a bright, crack-free plate in 1957. Porous chromium plate is a type of heavy chromium plate, generally more than 0.004 inch (0.1016 millimetre) thick, used in diesel engines and other internal combustion engines for oil retention and to get good "breaking in" qualities by etching or otherwise roughening the chromium surface and then finally honing or polishing it prior to use.

While decorative chromium plate has for a long time been specified to be about ten-millionths of an inch thick and commonly requires several minutes plating time to produce, a large-scale use has been found for radically thinner chromium coatings: to replace the tin on tin cans. This development originated in Japan. A rapidly moving strip-steel can stock is plated with about 0.0000003–0.0000005 inch (0.0000076–0.0000127 millimetre) of chromium in about one-third of a second, a coat that provides a good basis for an adherent coat of lacquer. The product is called tin-free steel (TFS), used predominantly for beverage cans. About 10 percent of all "tin" cans in the early 1970s were being made of TFS, and the proportion was expected to rise to about 30 percent.

While nonmetallic materials have been plated since the mid-19th century, a period of rapid growth in the utilization of electroplated plastics began in 1963 with the introduction of ABS plastic (acrylonitrile-butadiene-styrene), which was readily plated. Thus the utilization of plated plastics grew from 9,000,000 pounds in 1968 to about 35,000,000 pounds in 1970. The plastic part is first etched chemically by a suitable process, such as dipping in a hot chromic acid–sulfuric acid mixture. It is then sensitized and activated by first dipping in stannous chloride solution and then in palladium chloride solution. It is then coated with electroless copper or nickel before further plating. A useful degree of adhesion is obtained (about 5 to 30 pounds per inch; 1 to 6 kilograms per centimetre), but in no way comparable to the adhesion of metals to metals.

Applications. Copper plating is used extensively to prevent case hardening of steel on specified parts. The entire article may be copper plated, and the plate ground off on the areas to be hardened. Silver plating is used on tableware and electrical contacts; it has also been used on engine bearings. The most extensive use of gold plating is on jewelry and watch cases. Zinc coatings prevent the corrosion of steel articles, while nickel and chromium plate are used on automobiles and household appliances.

BIBLIOGRAPHY. W. BLUM and G.B. HOGABOOM, *Principles of Electroplating and Electroforming (Electrotyping)*, rev. 3rd ed. (1949), classical textbook, although somewhat out-of-date; J.B. MOHLER, *Electroplating and Related Processes* (1969), a good elementary and introductory text; F.A. LOWENHEIM (ed.), *Modern Electroplating* (1963), a good advanced text with emphasis on bath compositions and functions of constituents; A.K. GRAHAM and H.L. PINKERTON (eds.), *Electroplating Engineering Handbook* (1962), an advanced text with emphasis on plant and equipment; D.G. FOULKE and F.E. CRANE (eds.), *Electroplaters' Process Control Handbook* (1963), special emphasis on the analysis of solutions and the testing of deposits; H.W. DETTNER and J. ELZE (eds.), *Handbuch der Galvanotechnik*, 3 vol. (1963–69), the outstanding treatise, entirely in German; A. BRENNER, *Electrodeposition of Alloys: Principles and Practice*, 2 vol. (1963); G. DUBPERNELL, *"A Periodic Chart for Electroplaters," Proc. 33rd Conv. Am. Electroplaters Soc.*, pp. 244–257 (1946), "The Story of Nickel Plating," *Plating*, 46:599–616 (1959), "The Development of Chromium Plating," *ibid.*, 47:35–53 (1960); G.G. KAMM, A.R. WILLEY, and N.J. LINDE, "Surface and Corrosion Characteristics of Tin-Free Steel-Chromium Type for Beverage Containers," *J. Electrochem. Soc.*, 116:1299–1305 (1969); W. GOLDIE, *Metallic Coating of Plastics*, 2 vol. (1968–69), the outstanding text on the subject; G.M. WENGLOWSKI, "An Economic Study of the Electrochemical Industry in the United States," in J.O'M. BOCKRIS (ed.), *Modern Aspects of Electrochemistry, No. 4*, pp. 251–306 (1966); R.L. BRUMMET and H.E. ARNETT, *Business and Economic Evaluation of the Metal Finishing Industry* (1967).

(G.Du.)

Elementary and Secondary Education

As nations and peoples develop their language, codify their law, and create new patterns of work, their educational systems tend to become increasingly complicated in terms of levels and programs. In keeping with the specialized demands of modernizing societies, the educational system provides a source for the human resources and the knowledge necessary for continuing development. The educational system also assists in the preparation for specific occupational tasks and attempts to promote commitments to (and, ideally, constructive criticisms of) the values of society. In contemporary nations three major educational levels have emerged, each of which may be subdivided into numerous types of institutions.

Definitions of primary, elementary, and secondary education

The terminology used to identify the educational stages, however, may be confusing. The first major stage of the formal education system, for example, is called primary education in some countries and elementary education in others. Historically these terms have had quite different meanings. Thus during earlier periods in Europe, when highly selective educational systems existed, primary schools frequently referred only to those lower schools preparing students for advanced education. By contrast, elementary schools were terminal institutions that provided literacy and some general education for larger groups of children. Currently the choice between the phrases elementary education and primary education is largely a matter of national preference. The earlier distinction has largely disappeared, since most nations now have a single system for the first stage of schooling. In the United States, for example, the term elementary school is widely used, whereas the term primary school (in its various linguistic equivalents) is more common in western Europe.

Secondary schooling likewise presents certain definitional problems. Not only do most nations have a number of distinct kinds of secondary schools but also the demarcation of secondary schools from lower or higher institutions varies from nation to nation. Thus the sixth form in English education is typically considered part of secondary education, yet the academic study at this level would be more equivalent to the introductory years of American universities. Moreover, as part of secondary education there may be general secondary schools, various vocational secondary schools, and teacher-training schools. Any of these types may be considered terminal or as preparatory for higher education. They may emphasize arts, language, or science, or they may offer little specialization. Vocational secondary schools may provide only a single course of study, such as agriculture, or they may prepare pupils in any of a number of vocational skills. Teacher-training schools—for preparation of teachers in lower grades—is conducted at the secondary level in many nations. The trend, however, is to require postsecondary school training for future teachers.

SOME MAJOR MODELS IN EDUCATION

As the foregoing suggests, generalizations about the structure of elementary or secondary education may be misleading. National traditions, ideological preference, influences of climate and terrain, the level of industrialization, and other ecological considerations all influence the organization and control of education, the school program, and even the schedule for attendance. Certain influential models have existed historically, however, and it is well to identify these before entering into a discussion of recent trends (see also EDUCATION, HISTORY OF for a discussion of the historical roots common to much of European education as well as education in other areas).

Although the patterns of elementary and secondary schooling everywhere have taken on distinctive national colourations, cautious generalizations may be made about those systems found in certain regions. In western Europe, for example, the first stage of schooling is usually only four to five years in length, followed by a much longer period of secondary education. Traditionally these two levels have been separated by a rigorous and decisive examination. Secondary education usually consisted of a variety of terminal vocational institutions and one (or one public and one private) clearly demarcated academic path providing preparatory work for universities. Increasingly, however, with the extension of compulsory education and under egalitarian pressure for postponement of commitment to a particular course of study, the lower years of secondary education are being opened to all. Further, the fees formerly charged in secondary schools have now been largely abolished.

Because of their wide influence throughout the world, it is appropriate to sketch briefly the educational patterns of a few major European nations and the United States. For Europe an examination of France, Germany, England, and the U.S.S.R. should suffice.

The French system. Contemporary France has a first cycle of education for all children 6 to 11 years of age. Much has been written about the belief of the French people in the principle of national culture. The task of the French elementary school is to lay the foundation for the perpetuation of national culture by providing all children, through a common language, with knowledge of the common heritage of the nation.

Following this five-year period of general education, children enter one of a variety of secondary education programs. During the first two to three years of secondary schooling, however, children are in an orientation or observation stage. This period is used to determine which type of further education is most appropriate for the pupil. For the least academically talented boy or girl a three-year terminal program in general education is offered that will carry the pupil until the end of his compulsory education at age 16. A second group of pupils, having passed an aptitude test at the age of 11, are assigned to a short course in general education leading to a variety of vocational courses. The most academically talented boys and girls at the age of 11 enter *lycées*, where they specialize—with various options in classical languages, modern languages, and the sciences. Each program of studies at the secondary level awards its own diploma, certificate, or degree. The most prestigious of these is the *baccalauréat* offered only to students in the *lycées* upon successful completion of a series of examinations. The *baccalauréat* thus caps secondary education; however, it is also considered the first university degree and is a requirement for entrance to most programs of higher education. The *baccalauréat* bears some similarity to the *bachillerato* offered in Spanish-speaking nations at the end of general secondary education. This degree, however, should not be confused with the bachelor's degree offered in American and British universities.

Lycées

The West German system. In West Germany, schools are the responsibility of the governments of the states, or *Länder*, and the structure of elementary and secondary education is not entirely consistent throughout the nation. In all states, however, the period of elementary education covers eight or nine grades. All children spend the first four to six years in a neighbourhood elementary school. After this period, three basic options are available to the pupils. They may, after counselling by the elementary school teacher and upon the request of the parents, be placed in a *Realschule*, a *Gymnasium*, or a *Hauptschule*, the last representing a continuation of elementary education.

Those pupils attending the *Hauptschule* proceed with their study of language, arithmetic, geography, history, science, music, art, and physical education. Industrial arts for boys and home economics for girls are usually offered. After completion of a program of studies at the *Hauptschule*, which lasts four or five years, the pupil typically enters apprenticeship training.

In Germany, the term secondary school refers to institutions offering courses leading to the "Certificate of Maturity" (the *Reifezeugnis*), a qualification for entrance to an institution of higher education. The *Realschule* offers pupils further general education, some prevocational courses, and English-language study. At the age of 16, pupils conclude their program of studies and transfer to a vocational school or enter apprenticeship training.

If academically qualified, the pupil may also transfer to the *Gymnasium*.

Gymnasium

The *Gymnasium*, the third alternative for German youth, offers rigorous academic preparation for higher education. Like the *lycée* in France and the grammar school in England, the *Gymnasium* is designed for those pupils who have shown the most academic promise; and its curriculum emphasizing languages, mathematics, natural sciences, and social sciences requires a high degree of diligence throughout all of the nine grades. Unsuccessful pupils in the *Gymnasium* may be transferred to the *Hauptschule*. At the age of 16, moreover, pupils may terminate their academic studies and enter a vocational school.

Students in the *Gymnasium* must pass an examination, the *Abitur*, entitling them to the Certificate of Maturity, if they are to be admitted to a German university. The content of the *Abitur* is adjusted to the focus of studies, such as classical languages, modern language, or mathematics-science, chosen earlier by the student.

The English system. In England, the first stage of education is called primary education and includes pupils from ages 5 through 11. The great majority of these pupils pass through a sequence of infant and junior schools that together comprise the full system. Primary education is concerned with physical, spiritual, and moral development as well as the intellectual growth of children. The infant schools in particular tend to provide a relatively unstructured learning atmosphere for the child to experience and experiment with the world around him. The junior schools attempt to provide a sound introduction to disciplines as a preparation for further learning. These schools, however, have been moving in the direction of becoming more child centred, with increased emphasis on activity and greater accommodation to the interest of the children.

England, since World War II, like most other European nations, has increased the variety of secondary school programs available to adolescents. The allocation of pupils completing their primary education to different types of secondary schools takes place largely by means of an examination given at the age of 11 or shortly thereafter. The majority of students attend the secondary modern school, which offers a broad, general education until the age of 15. On the basis of their scores on the "eleven-plus" examinations, pupils may enter secondary technical schools and prepare to become skilled technicians.

The third school in the English tripartite system of secondary education is the grammar school, traditionally the most prestigious of the secondary institutions and the one most associated with university preparation.

Comprehensive schools

Some attempts have been made to incorporate the programs associated with modern, technical, and grammar schools in multilateral or comprehensive schools. (Indeed, the structure of secondary education has become both a pedagogical and a political issue of some magnitude in England.) This reform of educational structure is sometimes viewed as a move in the direction of the American-style comprehensive education. It should be noted, however, that a single comprehensive school for adolescents is not merely an American phenomenon; it is also well developed in Sweden and is under consideration in a number of other nations.

The students seeking university entrance must successfully complete a series of examinations that result in the General Certificate of Education. These examinations have two levels: ordinary and advanced. Entry to a university requires a prescribed combination of passes on the ordinary and advanced level in such subjects as English, foreign language, science, and mathematics. The sixth-form curriculum is largely oriented toward preparation for the advanced level examination and provides intense specialization.

In England a separate and highly significant group of independent schools exists. These primary and secondary schools are financially self-supporting. The best known of the independent schools are the "public schools," some of which, notably Eton and Harrow, have long maintained a distinguished international reputation. These schools for centuries have prepared students academically for higher education, typically at Oxford or Cambridge University, and ultimately for leadership in English life. Although a controversial element in English education and frequently accused of reinforcing invidious social distinctions, these institutions remain popular.

The Soviet system. Contemporary U.S.S.R. has an educational pattern that differs distinctively from that just described. The first stage of education takes place in an eight-year general-education school. This school offers the usual core subjects of language, mathematics, science, history, and geography. Work in art, music, and physical education is also provided. A unique feature is the provision of labour experience in order that pupils may acquire simple skills in working with wood and metal, so that they may become involved in socially useful work.

After graduating from the eight-year school, Soviet youth, if qualified, have a number of secondary schools open to them, including polytechnical or general secondary schools and various vocational or technical schools. Although in previous times there existed in the U.S.S.R. the same sharp division between the first (universal) and second (highly selective) stages of education common in other parts of Europe, in recent years the selection process has been extended over a number of years. The object now is to provide prolonged guidance for the pupils and place them into increasingly differentiated programs of study throughout the last few years of the elementary school and the first two or three years of secondary school. Many of those leaving the elementary school go directly into an occupation. Most critical for students wishing to continue their education is the selection at 17 or 18 years of age for higher education. The universities and higher institutes admit only 15 to 20 percent of the graduates of the secondary level institutions. Most graduates of the polytechnical or general secondary schools are required to take jobs for two or three years, after which those qualified may proceed to higher education. For those not admitted to regular educational programs at the secondary or higher levels there still are opportunities for continued education through a vast network of part-time educational programs.

Further comment needs to be made regarding polytechnical education. Polytechnical education is not the same as the Western vocational education in the Soviet view, for the latter is seen as referring to a much narrower set of experiences. Polytechnical education as understood by Soviet educators and political leaders is part of general education or citizenship training because all citizens are ideologically viewed as "workers." Thus all students must acquire both theoretical and practical training in the fundamentals of production in some agricultural or industrial pursuit. The specific educational programs that should comprise polytechnical education are, however, subject to continuing discussion and debate.

The U.S. system. The United States also offers a distinctive educational pattern—one that some observers see as representing, at least structurally, the trend in most industrialized societies. Without heritage of a distinction between primary education and elementary education (in the senses noted at the beginning of this article), the United States historically has had a single system of elementary and secondary education for the great majority of its students.

In the United States, education is the responsibility of the individual states; therefore some variation in structure exists. Typically, however, children at the age of six enter a neighbourhood school. Depending on the preference of the community, elementary education may extend from six to eight years. Although great variation exists, the elementary schools in the United States tend to provide a more relaxed learning environment and a more informal relation between pupil and teacher than found in most other parts of the world.

Comprehensive high schools

In the United States, over 80 percent of the secondary age groups attend secondary school. The American goal for decades has been to provide secondary education for all within a single type of institution popularly referred to as the comprehensive high school. Although the precise meaning of this term is frequently not clear, comprehensive schools usually include a broad program of general and specialized curricula and accommodate a wide range in the academic abilities of the pupils. Thus in the same American secondary school some students plan to pursue university study and others seek terminal secondary education. At times, however, this distinction is more apparent than real, for American schools are strongly influenced by residential patterns, which in some communities make for a high degree of homogeneity in the aspirations of parents and students.

Even though the principle of secondary education for all youth is well accepted in the United States, criticism is made regarding the academic standards in the secondary schools. Some persons have urged the schools to strive harder for academic excellence and have pointed to European secondary schools as models.

Students, at least those anticipating a university education, in the future may expect more stringent requirements in the basic disciplines. In spite of continuing controversy, however, there appears to be little possibility that a substitute will be created for the comprehensive school.

FACTORS AFFECTING EDUCATIONAL CHANGE IN THE 20TH CENTURY

A number of factors have influenced educational change in the 20th century, probably the most important of them being (1) the extent and rapidity of technological change, (2) the rate of population growth, (3) the growth in concern for equity, and (4) the increasing popularity of long-range, comprehensive planning.

Technological change. Technological developments during the 20th century have produced industrialized societies in many parts of the world and a few emerging postindustrial societies. With its more complicated division of labour, industrialization has meant a growth in the number of occupations requiring extensive formal educational preparation. The number of people who must be able to manage and to manipulate the technical and bureaucratic apparatus in an industrial society is large and requires a constant flow of literate persons and a stream of technically competent persons.

Relation between the occupational and educational structures

In the more advanced nations the occupational structure has already changed from a triangular shape, with little room at the top for highly skilled and professional tasks, to a diamond shaped structure, with a reduction in the demand for bottom-level unskilled labour and a large growth in the need for middle-level skills. Some observers see the most technologically advanced nations already moving toward an occupational structure shaped like an inverted triangle with the great bulk of occupations requiring advanced educational preparation. The direction of change in this occupational structure implies growth at every level of the educational system and indeed hints at the need for organized schooling for learning throughout one's life. To the extent that the schools react to or anticipate employment needs, the trends suggest that schools will be continually revising curriculum, adjusting facilities, and retraining teachers.

In the less developed nations technical skills frequently are in short supply. In the early stages of industrialization, consequently, rapid expansion of secondary and higher education may be called for. On the other hand, social demand for schooling may, at the same time, effect a rapid expansion of elementary schooling. Given sparse resources, the choice of educational priorities—that is, the choice of the best educational investment in terms of national goals—is indeed critical. Unfortunately, given the limited knowledge about the complex processes of economic and social development and the equally limited insight into the functioning of educational systems, there are few guidelines for these nations.

In the processes of economic and social development it is generally assumed that providing the requisite knowledge as well as the necessary certification for advanced skills will be the task of the educational system. In quantitative terms, therefore, the trend everywhere is toward expansion of educational enrollments at the secondary and higher levels. To a large extent the trend also is toward increased specialization in preparation for any of a growing number of technical and professional fields. The difference in rate of economic and technological change between the developed and underdeveloped nations, moreover, is significant. Educational expansion in the less developed nations has frequently outstripped the occupational opportunities for graduates. Generally, in western Europe and North America, on the other hand, the output of the educational systems has not kept abreast of the expanding technical and professional demands of the economies.

The swiftness of technical change requires not only new and increased amounts of technical knowledge but also great flexibility regarding its transmission. In a relatively static society, learning is comparatively fixed and limited. The transmission of knowledge from one generation to the next is thus a comparatively easy task. Under conditions of technological and social change, schools must adapt quickly to new demands. The rapid obsolescence of skills and knowledge, moreover, means that learning and formal instruction must take place during much of one's lifetime and cannot be limited merely to childhood and adolescence.

Demographic change. A second major influence on contemporary education is the nature of demographic change, more particularly the rate of population growth. It is obvious that the structure, size, and growth of a society's population directly affects education. If, for example, population is growing at a rate more rapid than the rate of expansion of educational plants and personnel, the enrollment ratios, literacy rates, median-school years completed, and other measures of educational attainment will tend to fall.

Population growth and education

The effects of population growth on education, though visible in most parts of the world, are most dramatic in the less developed nations. In Latin America, for example, the average population growth rate is more than 3 percent a year, while that in South Asia exceeds 2.6 percent. Due to improved health measures that have reduced infant mortality, moreover, the youth population is growing more rapidly than the population as a whole. Thus, merely to maintain, in the coming years, the present proportion of children and youth in school will require great efforts, particularly on the part of the nations of Asia and Latin America. Further, even if in the near future birth-rates are forced to drop, little immediate relief can be expected because the children who will attend the elementary and secondary schools for the next several years have already been born.

Rising expectations. Quantitative changes in education are also being strongly influenced by the phenomenon of increased social demand—a phenomenon widespread in contemporary societies. Rising expectations among the parents and the youth of less developed societies, stimulated by glimpses of the wealthier outside world or by the privileges of the small, educated groups at home, have given rise to an intense desire for education. In the more advanced societies, the expanded social demand for education may be traced to the growth in the affluence of large segments of the population and to the increased propensity of the privileged and less privileged alike to view education as a requisite for respectable status and a good job.

Egalitarianism

Associated with the increased demand for education is an egalitarian overtone—at times a dominant force, at times merely a subtle pressure. In much of Europe and North America, for example, many of the "educational reforms" have been a direct outgrowth of social reform aimed at extending social opportunities. The possibilities for status and success in the society are viewed as being closely related to the kind and level of education attained. This realization has led many nations, particularly the more industrialized ones, to seek to promote equality in educational opportunity. At a minimum, equity in schooling has come to mean an extended period of participation by all children and youth in a common educational culture.

Need for planning. The rapidity of technological and social change, the growing faith in man's ability to predict the consequences of such change, and the desire to make optimum use of limited human and natural resources have given great stimulus to planning in education. As in national planning, views of educational planning may be placed along a continuum. At one end of the continuum lies educational planning as the process of fulfilling explicit manpower targets established by the central government, and at the other end of the continuum

lies educational planning as a system of loose guidance to facilitate the individual's vocational and other choices. Although there remains a certain vagueness about the term educational planning, it may be distinguished from educational administration by (1) comprehensiveness of coverage, (2) a longer time perspective, and (3) more attention to economic and social supports and constraints on education.

The major elements taken into consideration in educational planning may be classified under four categories: (1) population change, (2) economic growth and economic development, (3) social demand, and (4) social equity. All nations do not give these elements equal emphasis. Many of the less developed nations concerned with limited resources are most concerned with adapting education to satisfy the needs of economic development for new knowledge and skills. The more industrialized nations, on the other hand, view as priorities the desires of all parents and students for more education and the demand among disadvantaged groups for greater equality of educational opportunity. Regardless of preferences, however, the increased interest in educational planning represents a growing optimism that the educational system is manipulable and manageable. Such confidence has led some reformers to use the educational system as a means of achieving planned economic and social change.

In summary, the most visible educational changes in the 20th century are (1) the diffusion and expansion of schooling and (2) the increasingly significant role of formal education in social and economic change. Technological change has demanded ever increasing levels of education for the population. Social pressure has resulted in mass education becoming at least the long-range goal of every contemporary nation. Although in the poorer nations universal schooling is as yet an unfulfilled goal, rudimentary schooling is being offered to large numbers of young children who in earlier periods went unschooled. In the more advanced nations of western Europe, and in Australia, Japan, Canada, and the United States nearly all children are provided free education at least through the lower grades of secondary school. In these nations literacy approaches 100 percent.

SCHOOL OBJECTIVES, CURRICULA, AND TEACHING METHODS

In most of the less developed nations the formal education of the majority of citizens ends after a few years of elementary education. It has been estimated that half the world's adult population remains illiterate despite the many continuing national and international efforts to cope with this problem. Only in the more industrialized nations has some portion of secondary education customarily become part of the common schooling of the majority of citizens.

Basic and elementary education. Despite the many cultural and political differences among nations, the objectives and curriculum at least of elementary education tend to be similar. Nearly all nations are officially committed to mass education, which is viewed as eventually including a full elementary education for all. An increasing agreement may therefore be found among nations to the effect that preparation for citizenship is one of the major objectives of elementary education. In terms of curriculum this objective suggests an emphasis on language competence, arithmetic skills, and basic social studies and science. The proportion of school time devoted to each of these areas may vary from nation to nation, but taken as a whole they typically comprise the bulk of the curriculum.

Some observers viewing the less developed nations have suggested that such a curriculum does not cover enough ground and that community or vocational skills should be included. Many innovations in this direction have been attempted. Local crafts are taught in the basic schools of India, for example; agriculture has been introduced in some of the primary schools of Uganda, Kenya, and Tanzania; and introductory skills in wood and metal working are included in the curriculum of mainland China. The specific reasons for teaching hand skills in the elementary school vary from nation to nation. In a few nations such teaching is viewed as providing vocational or prevocational skills; in some nations it is hoped that the products of the children's work will be marketable and hence defray the expenses of schooling; and in still other nations handwork is seen as a way of instilling an appreciation for labour. Objections have been voiced to these changes, however, by some parents and students who fear that the introduction of "practical" subjects may lessen the possibility of entrance to secondary schools—a widely held pupil aspiration.

Experiments in elementary curricula and methods

Perhaps more experimentation and innovation in curriculum and teaching methods have taken place at the elementary level than in secondary education. In the more progressive schools throughout the world attempts are being made to synthesize subjects previously taught separately. Examples of this are efforts to combine reading with literature and history with geography. In North America and western Europe, innovations such as team teaching, the use of more teachers' aides or clerks, and the establishment of ungraded schools have been tried. Most significant and pervasive, however, has been the widespread and growing acceptance of the principle that elementary education should focus on activity and experience as well as on subject matter. This principle as yet is reflected only in some of the elementary schools of a few nations but may well lead to a worldwide movement involving more informality in the learning environment, more student-initiated projects, and a greater variety in the content of elementary schooling.

Although not in any way opposed to the trend noted above, one remarkable similarity may be noted among the world's educational systems in the general sequence of learning. Elementary schools in most nations introduce the child first to the local environment and then systematically bring him into contact with larger environments. The way in which this is done, the relative emphasis on factual knowledge, the relative attention to the concept of culture, and the degree of pupil involvement in the design of the learning experiences may vary greatly from nation to nation or even between educational systems. Nevertheless, the principle that a child's learning should move from the immediate and familiar to the distant and unfamiliar appears to be widely accepted.

A special comment should be made on the Communist educational model, which offers somewhat unique objectives and curriculum. Although most nations of the world, in principle at least, share a concern for mass education, the fervour of Communist nations for infusing ideology and skills throughout all age groups, young and old, is rarely equalled. A dramatic case in point is offered by the example of mainland China, which claims to have moved the literacy rate of the Chinese people from 20 percent to approximately 90 percent between 1950 and 1970. In China, and also in Cuba, the curriculum even from the earliest grades is also linked to work, more particularly to manual labour. In addition to the usual reading, writing, arithmetic, and ideological lessons, elementary school children tend school gardens, care for the school buildings, and construct and repair simple furniture and home devices. Older children are involved in more complicated work projects in the school and community. Older children are, moreover, likely to spend their vacation working in farming areas—at the same time bringing literacy training to rural adults. To further enhance the status of work, teachers may be required to participate in harvests, and skilled farmers and craftsmen may in turn be used in the school to teach their respective skills.

Great variation then exists intranationally and internationally regarding such characteristics as the degree of stress on textbook content, the extent of teacher domination of classroom learning, and the relation between teachers and pupils. Some of these distinctions will be further elaborated later when considering a classification of types of schooling.

Secondary education. It is at the secondary level that the greatest contrast in objectives and curricula may be

found. Secondary schooling tends to be selective and therefore, with some exceptions, is not viewed as a part of citizenship preparation. Rather, secondary education typically is viewed as preparation for middle-level vocational and technical positions or as intellectual training for admittance to higher education. The *Gymnasium* of Germany, the *lycée* of France, and the grammar school of England, with their tradition of high-quality literary and scientific preparation, have greatly influenced the patterns of secondary education in Asia, Latin America, and Africa.

Subject-matter organization

Curriculum in general secondary education is strongly influenced by the academic disciplines, and most subjects offered in secondary schools could be easily classified under the humanities, social sciences, and natural sciences. Somewhat less attention is typically given to the fine arts, music, and physical education. There is, however, great variation among nations in the level of specialization required of students. The academic curriculum of the secondary schools in several European nations, for example, demands rigorous and extensive study in the chosen course. Scholarship is traditionally defined as competence within prescribed subject-matter boundaries. Teachers are viewed, and view themselves, as specialized scholars and thus have more in common with university professors than elementary schoolteachers.

The same orientation, if not the same level of academic accomplishment, may be said to exist in much of Latin America, Asia, and Africa. In Latin America, for example, most general secondary schools are called *liceos* or *colegios* and offer academic programs similar in nature to, although somewhat briefer in length than, those found in the general secondary schools of western Europe. One exception to this pattern may be found in Chile, which has lengthened the period of elementary schooling by moving the seventh and eighth grade from the secondary school to elementary education. The similarity between this reform and those attempts in western Europe to prolong the period of common education is quite apparent.

The model of secondary education that contrasts most strikingly with this pattern again is to be seen in the communist nations. The polytechnical concept of secondary education as found in the U.S.S.R. has already been mentioned. A more drastic attempt to link instruction and work is found in China, where a respect for labour and a suspicion of the humanities and empirical social sciences has led to a rejection of the traditional academic model. Most of the secondary school population in China attends part-time, vocationally-oriented middle schools. Secondary school students may operate small factories and experimental farms producing a variety of goods and foods for the nation. Even the general middle schools, which are the most academic in orientation, expect all students to involve themselves regularly in approved work projects.

Changing views of "intellectual" and "academic" education

Recent years have seen other modifications and occasionally the outright rejection of the intellectual objective of secondary education. In the United States, secondary education increasingly is becoming part of common education and therefore serves general citizenship objectives as well as preparation for advanced education. In western Europe, at least the lower years of secondary schooling are seen in much the same light. Furthermore, in the less industrialized nations there is dissatisfaction with the seeming remoteness of academic curricula in the secondary schools. Here proposals have been forthcoming to streamline curriculum by eliminating some of its historical and classical content and to define more broadly "intellectual" and "academic" to include experiences directly related to employment.

In the industrialized nations the direction of secondary education is clearly toward a unified pattern, which will represent at least a partial realization of the objectives of the European educators who for years have promoted the idea of *école unique* and *Einheitsschule*. The United States and Sweden have pioneered in promoting comprehensive secondary schools, where all students and all curricula would, in effect, be under one roof. Other European nations are experimenting with their own variants of comprehensive schooling. Alternatives to this approach may be found, for example, in the French *cycle d'orientation*, which provides for students, in their early secondary schooling, a guidance and diagnostic period before they must select the path for further education.

Methodology. Teaching techniques and teaching styles reflect variations in objectives and curriculum. The American educator Lawrence Thomas, in classifying types of schooling, has suggested the following categories: (1) memorizing, (2) training, (3) developing intellect, and (4) problem solving.

These types may be viewed as characterizing, or at least having implications for, the entire educational process. They are perhaps most useful, however, in offering perspective on methods of instruction and the selection and use of curricular materials.

Following the interpretation of Thomas, schooling oriented toward memorizing tends to place heavy reliance on this sequence: lecture–recitation–examination. The threat of failure is the chief source of motivation, and all learning activities are dominated by the teacher. The academic schools in many of the developing nations appear to fit this pattern. Schooling characterized as training usually implies that the teacher serves as a model, conducts drill, and supervises assignments or the exercise of prescribed skills. Most secondary vocational schools reflect this pattern. Schooling for developing the intellect features the teacher dominating the learning process through lecturing, presenting demonstrations, questioning the student about his ideas, and in turn inviting questions and debate. The student is expected to analyze predefined problems. The *Gymnasium*, *lycée*, and grammar schools of Europe are good examples of this pattern. Schooling as problem solving emphasizes experiences that provide students with the skills and attitudes to locate, define, and analyze problems found in school materials or in community life. Extra class activities, such as student clubs and student government, are considered vital if they serve to reinforce problem-solving skills. A few of the elementary schools of the United States and England illustrate this approach, as do experimental programs in science teaching in a number of nations.

Other classifications could be devised for the various methods of instruction employed in the classrooms of elementary and secondary schools throughout the world. Within the same school, different teachers or even the same teacher might employ a variety of teaching techniques. Nevertheless, the four types described serve well as frames of reference for understanding what is expected of children and youth in their schooling and, to some extent, in later life.

Summary. There has been a multiplication of objectives for elementary and secondary education in contemporary times. In addition to preparing pupils for advanced education, elementary and secondary schools increasingly provide a general education for a large proportion of the world's future citizens. These schools also provide youth with a growing array of vocational skills as the educational systems continually seek a degree of congruence with the needs of employment. Clearly these objectives may, at times, be in conflict. Institutions of higher education want only students who have demonstrated a high level of academic ability; the labour market seeks the most skilled or the most trainable young persons. Thus, regardless of the principle of equal opportunity in education, selection remains an important characteristic of the schools.

Although the general objective of providing a secondary education for all youth has not been fully achieved in any society, the wider participation of the various social strata is slowly being accomplished. How a student body representing a spectrum of social backgrounds and measured aptitudes may affect the design of curriculum and teaching methods is not, however, entirely clear. The concept of the curriculum not as a fixed body of traditional knowledge but as a changing set of experiences, skills, knowledge and values is achieving popularity. The view

of the teacher not as the fount of specialized knowledge but as a companion in the exciting but often confusing adventure of learning is gaining many adherents but is not yet widely reflected in elementary and secondary schools throughout the world.

CURRENT AND EMERGENT TRENDS

Viewed globally, several of the factors already described as having influenced the pattern of education earlier in the century continue to be potent determinants of educational change. One factor is represented by the expansion of enrollments resulting from population growth and increased public desire for the fruits of schooling. A second factor, most apparent in the industrialized societies but visible in most other nations, is the social pressure for equalizing educational opportunity between social classes, ethnic and racial groups, rural and urban populations, and the sexes. Planning through the use of new managerial techniques, systems analyses, and operations research constitutes a third factor. A fourth factor, arising partly in response to the three already mentioned and partly a product of educational theory and research, involves curriculum and pedagogical changes. Related to such changes but representing a fifth factor of a different magnitude is the dissatisfaction voiced by a few social critics with any of the existing models of formal education and the call for the initiation of new structures, programs, and financial arrangements for education.

There is little to add to earlier comments on the quantitative effects of population growth on elementary and secondary schooling. It should be noted, however, that education as a means of population control is receiving increased attention. Extraschool instruction in fertility control appears to be successful in several of the developing nations, for example, particularly those with relatively high levels of literacy. This has led to recommendations that in nations with rapid population growth fertility education should become an important part of the curriculum for girls in elementary and secondary schools.

Efforts toward social equity

Trying to create an educational system to satisfy the goals of social equity continues to cause difficulty. The complexity of the idea of equality of opportunity is, perhaps, only now being fully understood. As applied to the educational system, for example, should equality mean (1) providing a free education up to a certain level; (2) providing a common curriculum for all children; (3) providing equivalent inputs in terms of per-pupil expenditure, school plant, and the like; or (4) providing a racial and ethnic composition in the schools proportionate to that found in the greater society? Until recently it was widely assumed that expansion of enrollment and extension of the period of compulsory education would automatically reduce differentials in amount of schooling associated with social classes or with racial or ethnic groups. Experience has shown, however, that at least up to a point, educational expansion benefits the middle and upper classes much more than it does the lower classes or disadvantaged social groups.

As the definitions of equity vary, so will the ameliorative policies. Current knowledge is not yet adequate, however, to design programs to meet many of the possible objectives of equity. Many nations still hope that an increase in the rate of school participation of poorer segments of the population can be achieved largely through curriculum and structural changes in the schools. A few nations—the best example is perhaps the United States—have initiated large-scale efforts at extraschool compensatory education for certain of the more disadvantaged elements of the population. Work with preschool children from poor families in the United States suggests early instruction in conceptual and language skills may somewhat improve the ability of these children to cope successfully with the skills and knowledge taught in elementary schools. Unfortunately, studies appear to indicate that unless extraschool compensatory education is continued, paralleling regular schooling, the relative academic advantage of the children from the more privileged social groups is quickly reasserted.

In terms of planning and evaluating education for an assumed or desired future, some rather formidable tools are being developed. Economic models and high-speed computers allow intricate simulation models to be employed and permit the examination of large quantities of data regarding educational inputs and desired outputs. The utility of these and other managerial and analytic techniques is hampered, however, by a lack of insight into the dynamics of educational systems. It is difficult, for instance, to plan for certain outputs or to evaluate the consequences of a particular educational policy without more understanding of the teaching–learning process, the effects on learning of different administrative styles, the potentials and limitations of instructional technology, and the functioning of the peer group. Although progress is being made in these directions, no one can yet predict the results on the learner of a given combination of inputs in the educational system.

Efforts toward new curricula and methods

Although innovations in curriculum and in teaching techniques are too numerous to describe in detail, a few proposals can be identified that suggest the general directions of change. One of the newer approaches in curriculum and in instructional techniques has been called the inquiry or discipline approach. In this approach the function of instruction is to help the pupil develop several modes or styles of inquiry rather than acquire a store of information per se. The proponents of this approach argue that, given the rate of increase in knowledge, it is extremely difficult for a teacher to attempt continually to update content or information; moreover, the explosion of knowledge creates an impossible burden of subject matter on the student. Thus, instead of teaching the discipline directly, some argue that the methods or style of inquiry of the discipline should be taught to the student. Students, in effect, would learn science by acquiring the scientific method, would learn history through the historical method, and so on. This approach has acquired many supporters, particularly among the teachers of mathematics and the sciences in the United States. One of the major advantages of focussing on the structure and methodology of the discipline lies in the ability of pupils to develop skills of analyzing and explaining phenomena. Limitations, however, are significant: to teach the methods of scientific inquiry that currently exist may frustrate a person's ability to learn new methods as they arise. Further, a focus on method and style of inquiry does not deal directly with the real substantive problems and issues of life.

It is this latter point that conflicts with the demands of young radicals in many parts of the world. The call by youth for relevance in education has most dramatically affected the universities but nevertheless is being heard at lower educational levels. In the United States, for example, many young blacks are demanding programs of black studies in the secondary schools as well as in the colleges. At every level the more vocal pupils are asking the schools to give greater attention to social issues. But the radicals are not seeking rational, disciplined analyses of contemporary problems. Such formal study will, many of them argue, only lead to social paralysis. Education to be relevant must be personal and intuitive and must deeply involve the student.

Whether or not a disciplinary approach is used in the selection and study of subject matter, the trend is toward more active student involvement in the learning process. The idea that children and youth must engage in the act of discovering knowledge for themselves is gaining support among educators and teachers.

Trends toward newer technologies

The idea of discovery may be related to the more general trend toward individualizing and personalizing instruction. Techniques such as team teaching, flexible scheduling, and the provision of individualized instructional technology (headphones for language tapes, programmed learning, desk computers, and so forth) support this trend. Yet, the innovations in curriculum and teaching methods being described have not been widely implemented even in the wealthier nations. High costs, inadequate teacher preparation, and lack of appropriate ad-

ministrative leadership are among the many obstacles to implementation. Many educators and teachers, moreover, remain committed to schooling as an accumulation of a rather clearly defined body of knowledge, and such persons are unconvinced that the newer instructional approaches do indeed lead to the attainment of such an objective.

Improved educational technology may lead to greater efficiency in achieving the objectives of elementary and secondary education. Advocates of a greater use of technology in the schools point out that educational systems make less use of media, computers, and other applications of technology than other sectors of society. Advocates point out that, far from being a dehumanizing force, technology can free the teacher from the more routinized work. Through instructional technology, pupils proceeding at an individual rate may acquire the basic skills, diagnose their own difficulties, and evaluate their own progress. The teacher can then become a coordinator of learning activities rather than a distributor of knowledge. Through the wise use of technology, furthermore, the teacher can be provided the opportunity to concentrate on the more advanced intellectual, aesthetic, and ethical objectives.

In terms of the final trend—namely, the consideration of alternatives to formal education—the reasons for dissatisfaction with present structures and programs are relatively clear; the nature of the needed substitutes, however, is more difficult to discern. By way of illustration, some U.S. administrators and economists concerned both with rising educational expenditures and with the general lack of efficiency in the educational process have argued that vocational and professional education, and perhaps all of formal schooling, should become, in effect, a private business. Education assumed to have high community benefits might be financed by the government, but other education would be privately financed. Thus, vouchers would be distributed to parents and children based on the cost per child of that amount and kind of education deemed beneficial to society. Critics of this proposal have pointed out (1) the difficulty, if not impossibility, of separating out those aspects of education that have primarily societal effects and (2) the likelihood that such an arrangement would lead to even greater racial and social-class segregation. In any case, all schemes for the direct purchase of professional services for education have thus far largely remained matters of academic discussion.

In western Europe and North America a few drastic innovations in schooling have moved from the drawing board into operation. These vary greatly in organization, curriculum, and teaching methods but tend in the direction of creating a very permissive environment with maximum opportunity for children to experience the world about them in their own way. In the less developed nations the dissatisfaction of political leaders with current educational arrangements generally lies less with the possible constraints on individual expression or even on individual creativity than with the failure of educational investment to result in rapid economic growth. As a result, some nations are examining the agressive policy of China in linking formal schooling with productive work. Although ideological differences may keep many nations from adopting the Chinese educational model, work-study and study-work programs are on the increase. Particularly in the less industrialized nations, the educational systems tend to be poorly adjusted to the changing needs of employment. Consequently, there is an increase in experimental programs for training youth within business and industry and an increase in arrangements linking industry and the educational system in partnership. The point of view appears to be growing that, although elementary and secondary education may provide certain prevocational skills and possibly certain generalized vocational skills, specific occupational training best takes place on the job.

The future is likely to bring numerous and often radical experimentation and innovation in elementary and secondary education. Some nations and communities may relegate a number of the current educational functions to other agencies. In some cases, for example, the responsibility for the teaching of certain skills may be removed from the schools and given to the place of employment, the peer group, or specified tutors. In any case, forceful steps will most assuredly be taken in many nations to ensure more equitable representation of the less privileged social groups in the educational process. And, regardless of which arrangement is chosen, the notions of relevance, efficiency, and flexibility will serve as major guides.

BIBLIOGRAPHY. Among the recent standard descriptive works on comparative education treating elementary and secondary education are J.F. CRAMER and G.S. BROWNE, *Contemporary Education*, 2nd ed. (1965); E.J. KING, *Other Schools and Ours*, 3rd ed. (1967); and I.N. THUT and D. ADAMS, *Patterns of Education in Contemporary Societies* (1964).

Since most scholars in comparative education specialize in one or more geo-cultural areas, the resulting studies, while often not explicitly comparative, offer some of the best factual descriptions and some of the more penetrating insights into education. These publications range in focus from the Handley factual country reviews sponsored by the U.S. OFFICE OF EDUCATION to the more elaborate field investigations represented in P.J. FOSTER, *Education and Social Change in Ghana* (1965); and A.M. KAZAMIAS, *Education and the Quest for Modernity in Turkey* (1966). The series on "Education in Europe," prepared by the Council for Cultural Co-operation of the Council of Europe, provides a useful guide to the structure and curriculum of elementary and secondary education.

Concern for more meaningful comparisons of education across cultures has given rise to attention to methodology. The first book in the field devoted entirely to method was G.Z.F. BEREDAY, *Comparative Method in Education* (1964). Recent writings in comparative education increasingly have emphasized the empirical and social sciences. Examples of this trend are H.J. NOAH and M.A. ECKSTEIN, *Toward a Science of Comparative Education* (1969); and J. FISCHER (ed.), *The Social Sciences and the Comparative Study of Educational Systems* (1970).

(D.A.)

Elements, Geochemical Distribution of

Knowledge of the geochemical distribution of elements involves elucidation of the relative and absolute abundances of the chemical elements in the Earth and in its various parts—the crust, interior, atmosphere, and hydrosphere. This comprises a major part of the science of geochemistry, which is the study of the distribution of the chemical elements in space and time and the laws governing this distribution (see GEOLOGICAL SCIENCES). Basic knowledge in this area was largely accumulated during the 19th century. The concept of a limited number of chemical elements had been established by 1800, and the appearance of the periodic table (*q.v.*), or system, in 1869 provided a new insight into the limitations on the number of elements. Concurrent with these advances in chemical understanding, from about 1850 onward there was a steadily increasing output of analytical data on the Earth's rocks, minerals, and waters, mainly from laboratories in Europe and North America. The output from North America was materially increased following the establishment of the United States Geological Survey in 1879 and the appointment of Frank W. Clarke as chief chemist in 1884.

Basic work of Clarke and Goldschmidt

Clarke's name will always be linked with the study of the geochemical distribution of the elements—indeed, the term clarke was proposed as the unit for the average percentage of an element in the Earth's crust by Soviet scientists and has been generally adopted. In 1889 Clarke wrote the first of his many publications on the geochemical distribution of the elements. He assembled many chemical analyses of rocks from different continents, calculated average values, and showed that the overall chemical compositions of continental areas are remarkably similar. By combining these averages he obtained values for the abundances of the commoner elements in the continental crust of the Earth, values that have not been materially changed in spite of the vast increase of available data since that time. He also estimated abun-

dances for many of the less common elements; these estimates were based in many instances on very limited and imprecise data and subsequently have been improved.

A further development of great significance was the assemblage of comprehensive data on the abundances of individual elements in terrestrial materials and in the cosmos (based on solar and meteorite abundances) by the Norwegian geochemist Victor Moritz Goldschmidt during the 1930s. Goldschmidt's tables provided the basis for modern research on the geochemical distribution of the elements, and his compilation of data on cosmic abundances was the key to later theories on element synthesis in stars and supernovae.

Goldschmidt also contributed to the understanding of elemental distribution within the Earth through his geochemical classification of the elements into lithophile, siderophile, chalcophile, and atmophile. Lithophile elements are those with a strong affinity for oxygen; they are concentrated in the crust or lithosphere as silicate and oxide minerals. Siderophile elements are principally metals that alloy readily with iron; Goldschmidt explained their scarcity in the Earth's crust by their concentration in the nickel–iron core (the siderosphere). Chalcophile elements are those with strong affinity for sulfur; they occur mainly as sulfides. And atmophile elements are gases, such as nitrogen, argon, and other rare gases, which are unreactive and hence accumulate in the atmosphere. (Goldschmidt also proposed a group of biophile elements, for those that concentrate in living matter—essentially carbon, hydrogen, oxygen, nitrogen, sulfur, and phosphorus.)

This article treats the distribution of chemical elements in the Earth's core, mantle, and crust, with specific reference to the principal groups of rocks, ore deposits, mineral fuels, and soils. It also covers the geochemical distribution of elements in the Earth's atmosphere and hydrosphere. The significance of the data on element abundances that are presented is enhanced by inclusion of information on the cosmic or absolute abundance of elements at the outset. For further information on the chemical composition of extraterrestrial bodies, see MOON; METEORITES; and articles on the individual planets. For additional detail on the Earth's interior, see EARTH, STRUCTURE AND COMPOSITION OF. And see the articles OCEANS AND SEAS; OCEANS, DEVELOPMENT OF; FRESHWATER, GEOCHEMICAL PROPERTIES OF; ATMOSPHERE; and ATMOSPHERE, DEVELOPMENT OF, for discussion of the chemistry and chemical processes of the Earth's shells of air and water.

ABSOLUTE OR COSMIC ABUNDANCES OF THE ELEMENTS

In order to understand the processes whereby elements are distributed or concentrated in specific environments or regions, it is necessary to have some knowledge of what may be called their absolute abundances, namely, their overall abundances within the entire universe. Such knowledge cannot be obtained from the analyses of terrestrial rocks. The rocks of the Earth's crust are extremely varied in composition, and they represent the end-products of a long series of chemical differentiation processes from the primordial matter that aggregated to form the Earth. Even this primordial matter, if it could be sampled, would not give the absolute abundances of the chemical elements; it already was enormously depleted in hydrogen and helium with reference to the universe as a whole. The best source of information in this regard is the composition of the Sun, other stars, and the Moon.

The Sun and other stars. Goldschmidt derived his table of absolute abundances from a careful evaluation of spectrographic data (emitted light that is correlative with chemical composition) on the Sun and other stars, combined with analyses of meteorites, which he recognized as probably being a good average sample of the nonvolatile (nongaseous) matter of the solar system. More recent work has largely confirmed these premises; in particular, it has been shown that elemental abundances in a specific group of meteorites, called the Type I carbonaceous chondrites, are in excellent agreement with abundances determined in the Sun. It is important to note that more than 99 percent of the mass of the Sun —and probably of the universe as a whole—consists of hydrogen and helium.

When the cosmic abundances of the elements are plotted against their atomic numbers (Figure 1), a number of significant relationships can be observed. The abundances show a rapid decrease for elements of lower atomic number (to about atomic number 30), followed

From L.H. Ahrens, *Distribution of the Elements in Our Planet* (1965); McGraw Hill Book Co.

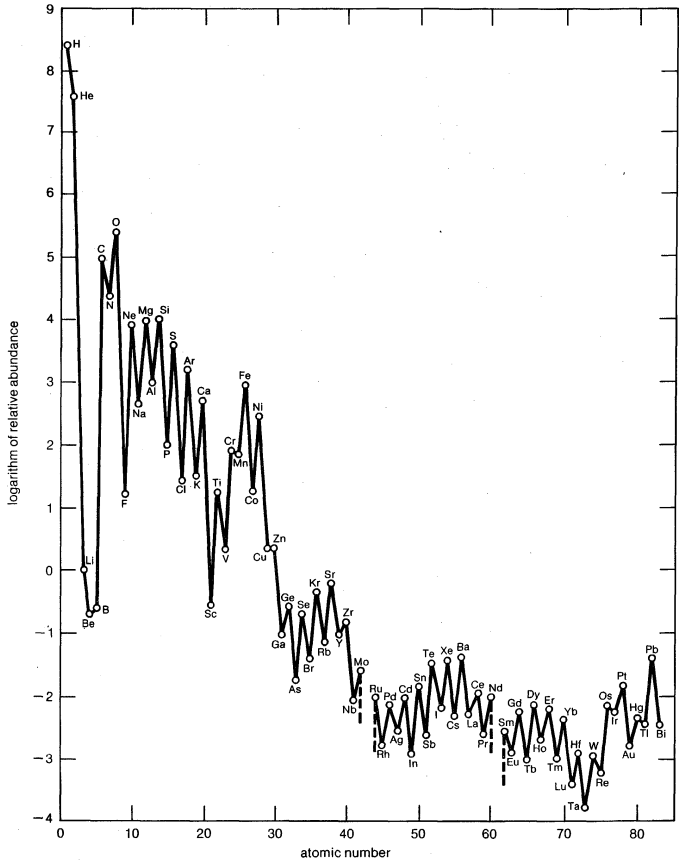


Figure 1: Relative abundances of the elements.

by fairly uniform abundances for the heavier elements. Elements of even atomic number are more abundant than their neighbours of odd atomic number on either side; this regularity was first observed in analyses of terrestrial rocks in 1914 and 1917. Only ten elements—hydrogen, helium, carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, and iron, all of which have atomic numbers less than 27—show appreciable abundance; of these, hydrogen and helium far outweigh the other 8. These relationships indicate that the absolute abundances of the elements depend on nuclear rather than chemical properties and are related to the inherent stability of their atoms. On this basis the decrease in relative abundances with increasing nuclear complexity can readily be explained. The absence of elements 43, 61, 85, and 87 from the Earth is due to an almost complete instability of any nuclear arrangement for these atomic numbers (note that they are all odd-numbered). The remarkably low abundances of lithium, beryllium, and boron among the low-numbered elements can be explained by the ready disintegration of atoms of these elements under bombardment by high-energy particles, such as takes place in the Sun and other stars.

Element abundances on the Moon. Until recently, more was known about element abundances in distant stars than in the Earth's nearest neighbour, the Moon. The lunar landings have provided samples that have been intensively analyzed in many laboratories throughout the world. The data for the Apollo 11 material, collected in

The Apollo data

Table 1: Abundances of the Elements in Different Geochemical Materials*†
(figures are in parts per million unless otherwise indicated)

atomic number and element	carbonaceous chondrites (type I)	Moon (Apollo 11)	Earth's crust	granite (G-1)	basalt (W-1)	shales	sandstones	carbonates	metamorphic rocks	seawater
(1) Hydrogen	2.2%	200	1,400	380	590	—	—	—	1,000	10.7%
(3) Lithium	1.3	12	20	24	12	66	15	5	22	0.17
(4) Beryllium	0.04	2	2	3	0.8	3	0.X	0.X	1.3	6×10^{-7}
(5) Boron	5	2	10	1.5	15	100	35	20	—	4.5
(6) Carbon	3.7%	140	250	260	220	1.0%	—	11.3%	0.37%	28
(7) Nitrogen	2,700	100	20	8	14	600	—	—	—	0.7
(8) Oxygen	45.3%	40.0%	46.6%	48.2%	44.6%	50.5%	53.4%	49.8%	48.0%	85.7%
(9) Fluorine	190	140	650	700	250	800	180	240	500	1.3
(11) Sodium	0.55%	0.33%	2.8%	2.4%	1.6%	0.96%	0.33%	0.04%	2.6%	1.08%
(12) Magnesium	9.6%	4.5%	2.1%	0.23%	4.0%	1.5%	0.70%	4.7%	1.4%	0.13%
(13) Aluminum	0.85%	5.6%	8.1%	7.4%	7.9%	8.0%	2.5%	0.42%	7.7%	1×10^{-2}
(14) Silicon	10.3%	19.2%	27.7%	34.0%	24.9%	7.3%	36.8%	2.4%	30.4%	2.9
(15) Phosphorus	1,400	500	700	390	610	700	170	400	650	9×10^{-2}
(16) Sulfur	6.2%	1,700	260	170	130	2,400	240	1,200	600	904
(17) Chlorine	260	14	130	70	200	180	10	150	100	1.94%
(19) Potassium	0.05%	0.14%	2.6%	4.6%	0.53%	2.7%	1.1%	0.27%	2.6%	392
(20) Calcium	1.06%	8.0%	3.6%	1.00%	7.8%	2.2%	3.9%	30.2%	2.9%	411
(21) Scandium	5	75	20	3	34	13	1	1	—	4×10^{-5}
(22) Titanium	0.04%	5.9%	0.44%	0.16%	0.64%	0.46%	0.15%	0.04%	0.31%	1×10^{-3}
(23) Vanadium	57	50	140	16	240	130	20	20	53	2×10^{-3}
(24) Chromium	2,200	2,100	100	22	120	90	35	11	99	2×10^{-4}
(25) Manganese	1,700	1,900	950	230	1,300	850	X0	1,100	530	4×10^{-4}
(26) Iron	18.4%	14.3%	5.0%	1.36%	7.7%	4.7%	0.98%	0.38%	3.1%	4×10^{-3}
(27) Cobalt	480	25	25	2.4	50	19	0.3	0.1	21	4×10^{-4}
(28) Nickel	1.04%	15	70	1.5	78	68	2	20	23	7×10^{-3}
(29) Copper	140	11	50	13	110	45	X	4	14	1×10^{-3}
(30) Zinc	320	15	70	45	82	95	16	20	—	5×10^{-3}
(31) Gallium	10	4.5	15	18	16	19	12	4	14	3×10^{-5}
(32) Germanium	34	0.1	1.5	1	1.7	1.6	0.8	0.2	—	6×10^{-5}
(33) Arsenic	2.0	0.05	1.8	0.8	2.4	13	1	1	—	3×10^{-3}
(34) Selenium	27	0.8	0.1	—	0.11	0.6	0.05	0.08	—	1×10^{-4}
(35) Bromine	5	0.1	2.5	0.5	0.4	4	1	6	—	67
(37) Rubidium	2.3	3.4	90	220	22	140	60	3	118	0.12
(38) Strontium	8	170	390	250	180	300	20	610	340	8.1
(39) Yttrium	1.6	120	30	13	25	30	40	30	—	1×10^{-5}
(40) Zirconium	11	370	170	210	100	160	220	19	400	3×10^{-5}
(41) Niobium	0.5	21	20	20	10	11	0.0X	0.3	—	2×10^{-5}
(42) Molybdenum	1.6	0.5	1.5	7	0.5	2.6	0.2	0.4	—	1×10^{-2}
(44) Ruthenium	0.7	—	0.01	—	<0.4	—	—	—	—	7×10^{-7}
(45) Rhodium	0.2	—	0.005	<0.005	<0.005	—	—	—	—	—
(46) Palladium	0.6	0.006	0.01	0.002	0.01	—	—	—	—	—
(47) Silver	0.4	0.008	0.05	0.04	0.05	0.07	0.0X	0.0X	—	3×10^{-4}
(48) Cadmium	1.0	0.004	0.2	0.06	0.3	0.3	0.0X	0.04	—	1×10^{-4}
(49) Indium	0.09	—	0.05	0.025	0.07	0.1	0.0X	0.0X	—	—
(50) Tin	1.6	0.6	3	4	3	6	0.X	0.X	—	8×10^{-4}
(51) Antimony	0.15	0.007	0.2	0.4	1.1	1.5	0.0X	0.2	—	3×10^{-4}
(52) Tellurium	3.3	0.02	0.01	—	<0.2	—	—	—	—	—
(53) Iodine	0.6	—	0.5	—	<0.05	2.2	1.7	1.2	—	6×10^{-2}
(55) Cesium	0.19	0.10	2	1.5	1.0	5	0.X	0.X	—	3×10^{-4}
(56) Barium	4	200	400	1,200	180	580	X0	10	1,070	2×10^{-2}
(57) Lanthanum	0.19	18	30	100	12	34	30	X	—	3×10^{-6}
(58) Cerium	0.63	54,	60	170	23	67	92	12	—	1×10^{-6}
(59) Praseodymium	0.09	11	8	17	4	8.8	8.8	1.1	—	6×10^{-7}
(60) Neodymium	0.42	46	30	55	17	31	37	4.7	—	3×10^{-6}
(62) Samarium	0.13	15	6	9	4	6.9	10	1.3	—	1×10^{-6}
(63) Europium	0.05	1.9	1	1.3	1.1	1.4	1.6	0.2	—	1×10^{-7}
(64) Gadolinium	0.24	20	5	5	4	6.2	10	1.3	—	7×10^{-7}
(65) Terbium	0.04	3.6	1	0.6	0.8	1.0	1.6	0.2	—	1×10^{-7}
(66) Dysprosium	0.22	25	3	2.5	4	4.5	7.2	0.9	—	9×10^{-7}
(67) Holmium	0.06	4.9	1	0.5	1	1.2	2.0	0.3	—	2×10^{-7}
(68) Erbium	0.14	14	3	2	3	3.3	4.0	0.5	—	9×10^{-7}
(69) Thulium	0.02	1.9	0.5	0.2	0.35	0.2	0.3	0.04	—	2×10^{-7}
(70) Ytterbium	0.13	13	3	1	2.2	2.6	4.0	0.5	—	8×10^{-7}
(71) Lutetium	0.02	1.9	0.5	0.2	0.35	0.7	1.2	0.2	—	2×10^{-7}
(72) Hafnium	0.26	13	3	6	2	2.8	3.9	0.3	—	$<8 \times 10^{-6}$
(73) Tantalum	0.02	1.7	2	1.6	0.7	0.8	0.0X	0.0X	—	$<3 \times 10^{-6}$
(74) Tungsten	0.14	0.3	1	<0.8	0.45	1.8	1.6	0.6	—	$<1 \times 10^{-6}$
(75) Rhenium	0.04	—	0.001	0.0006	0.0003	—	—	—	—	8×10^{-6}
(76) Osmium	0.45	—	0.005	<0.00006	0.0003	—	—	—	—	—
(77) Iridium	0.40	—	0.001	—	0.0003	—	—	—	—	—
(78) Platinum	0.9	—	0.01	<0.01	0.02	—	—	—	—	—
(79) Gold	0.18	—	0.005	0.005	0.005	—	—	—	—	1×10^{-5}
(80) Mercury	1(?)	—	0.08	0.25	0.11	0.4	0.03	0.04	—	2×10^{-4}
(81) Thallium	0.14	—	0.5	1.3	0.13	1.4	0.8	0.0X	—	1×10^{-5}
(82) Lead	2.9	1.2	13	49	8	20	7	9	—	3×10^{-5}
(83) Bismuth	0.13	—	0.2	—	—	—	—	—	—	2×10^{-5}
(90) Thorium	0.04	2.0	7	52	2.4	12	1.7	1.7	—	4×10^{-8}
(92) Uranium	0.01	0.5	2	4	0.5	3.7	0.5	2.2	—	3×10^{-8}

*Omitting the noble gases. †Values incorporating the symbol X are known only to the order given (*i.e.*, 0.0X indicates a value in the range 0.01 to 0.09).

the Sea of Tranquility (Mare Tranquillitatis), are given in Table 1. Analyses of Apollo 12 collections are similar for most of the elements. Comparison of the analytical data with those for carbonaceous chondrites (a type of meteorite that provides a good average sample of nonvolatile solar system material) shows that the lunar material has undergone marked geochemical fractionation (segregation of elements). Lithophile elements, those with an affinity for oxygen, are markedly enriched in the lunar material, whereas chalcophile and siderophile elements (with affinities for sulfur and iron) are markedly depleted. Although the lunar data have been obtained from only a few sites, they suggest that the material of the Moon has undergone a complex geochemical history.

TERRESTRIAL DISTRIBUTION OF THE ELEMENTS

The study of earthquake waves passing through the body of the Earth has shown that the interior is not uniform;

it consists of distinct shells separated by concentric discontinuities at which the velocities of the passing waves change. The two major discontinuities that are universally recognized are the Mohorovičić Discontinuity, which divides the Earth's crust from its underlying mantle, and the Wiechert–Gutenberg Discontinuity, which separates the mantle from the core. The latter discontinuity exists at a depth of 2,900 kilometres (1,800 miles); it is marked by a sudden increase in density, from about 5.7 at the base of the mantle to 9.7 at the top of the core. The only reasonable interpretation of this discontinuity is that the mantle consists of silicates and oxides of the common elements (largely magnesium and iron), and the core consists of metallic iron alloyed with minor amounts of other elements (analogous to the nickel-iron in meteorites). The Mohorovičić Discontinuity varies in depth from place to place; it averages about 33 kilometres (20 miles) below the continents and about 8 kilometres (5 miles) below the bottom of the deep oceans. It too is marked by a density increase from crust to mantle—a comparatively small one, from about 3 to 3.3.

To the three spherical divisions—crust, mantle, and core—two more should be added: the hydrosphere, which is the discontinuous shell of fresh and salt water, on and within the crust; and the atmosphere, the ocean of air that surrounds the Earth, gradually thinning into the vacuum of outer space. The principal characteristics of these "geospheres" are set forth in Table 2.

Table 2: Volume and Masses of Earth Shells

	thickness (km)	volume ($\times 10^{27}$ cm^3)	mean density (g/cm^3)	mass ($\times 10^{27}$ g)	mass (percent)
Atmosphere	—	—	—	0.000005	0.00009
Hydrosphere	3.73 (mean)	0.00135	1.03	0.00139	0.024
Crust	20 (mean)	0.010	2.8	0.029	0.5
Mantle	2,880	0.897	4.5	4.010	67.1
Core	3,471	0.175	11.0	1.936	32.4
Whole Earth	6,371	1.083	5.52	5.976	100.00

Correlation with density and other physical properties

The Earth's core. The evidence for the composition of the core is all indirect because no means have yet been devised for directly sampling the deep interior of the Earth. The moment of inertia of the Earth indicates that there is a concentration of mass around the centre, and seismic data have shown that below the Wiechert–Gutenberg Discontinuity the density of the material is high, ranging upwards from 9.7. The only heavy element with high cosmic abundance is iron, and because an iron–nickel alloy is an important meteorite component, it is reasonable to conclude that the Earth's core consists largely of metallic iron with a minor admixture of other elements. This conclusion is supported by geophysical evidence that indicates that the mean atomic number of the material of the core is about 22. The atomic number of iron is 26, so this implies that the core also contains an admixture of elements of lower atomic number. Sulfur, with atomic number 16, and carbon, with atomic number 6, are relatively abundant in meteoritic matter, and the presence of minor amounts of these elements in the core would effectively reduce the mean atomic number. Some authorities have advocated silicon (atomic number 14) as the major alloying component in the core, but this seems less likely; if silicon were the sole alloying element, then the core would have to contain more than 30 percent silicon in order to reduce its mean atomic number to 22. In addition, free silicon requires extremely reducing conditions (lack of oxygen), and the presence of ferrous iron in the mantle is inconsistent with this requirement.

With present information it is not possible to give definite figures for the abundances of the elements in the Earth's core. It is certainly made up largely of metallic iron, however, probably with some nickel, a little cobalt, and appreciable amounts of such lighter elements as carbon and sulfur.

The Earth's mantle. The mantle comprises that part of the Earth between the Mohorovičić and the Wiechert–Gutenberg discontinuities. It makes up 83 percent of the volume of the Earth and 67 percent of its mass and is thus of decisive importance in determining the bulk composition of the planet. In estimating elemental abundances in the mantle, however, the same difficulty as with the core arises: direct sampling is not feasible. Much more geophysical data are available for the mantle, however, and some volcanic eruptions have brought rock fragments to the surface that have certainly been derived from this zone. The most remarkable of these materials are the diamond-bearing inclusions found in the famous pipes, or volcanic necks, that are mined in South Africa and Siberia. The presence of diamond, the high-pressure form of carbon, implies a depth of origin of at least 100 kilometres, but these inclusions are rare. The common type of mantle-derived inclusion is peridotite, a silicate rock consisting largely of olivine, $(Mg,Fe)_2SiO_4$, with minor amounts of orthopyroxene, $(Mg,Fe)SiO_3$, and diopside, $CaMg(Si_2O_6)$.

Implications of the occurrences of diamond

Geophysical information indicates that below a depth of about 1,000 kilometres the mantle behaves as an essentially homogeneous material, but above this level its physical properties are more varied, and there is evidence for second-order discontinuities. This region above 1,000 kilometres is frequently referred to as the upper mantle, and in recent years has been the object of a concentrated research effort by geologists and geophysicists all over the world. The significance of the upper mantle is that processes originating there have dramatic effects on the surface—in the form of volcanic eruptions and some earthquakes—and less dramatic but equally important effects within the crust, such as the introduction and concentration of some elements, possibly leading to the formation of ore deposits. Increased knowledge of the upper mantle thus has both scientific and economic appeal.

Geophysical data on the properties of the upper mantle suggest that it must consist essentially of magnesium-iron silicates, probably largely olivine in the region immediately below the crust. Olivine is not stable under very high pressures, however; it is converted to a different phase of about 10 percent higher density and with a structure like the mineral oxide spinel ($MgAl_2O_4$). This conversion would occur in the mantle at depths of around 400 kilometres, and a second-order discontinuity at that depth can plausibly be ascribed to this conversion. Pyroxenes also undergo transformations to phases of greater density at the high pressures within the mantle. Thus the mantle, although composed of material of familiar chemical composition, consists, in its lower part at least, of different minerals than those in the upper part.

Many estimates of the composition of the upper mantle have been made in recent years; some of these are presented in Table 3. On the whole, the similarities are more important than the differences. All agree that the principal components are oxides of silicon, magnesium, and iron. The differences are mainly in the minor components such as aluminum oxide, calcium oxide, and the alkalies, and are determined largely by theoretical considerations and the weight given to specific aspects of the geophysical and geochemical data.

Although fairly reliable estimates exist for the abundances of the major elements in the mantle, little is known of minor and trace elements. Knowledge of the crystal structure of possible mantle minerals indicates that many minor and trace elements will not be readily incorporated, however. They are therefore likely to concentrate in liquid material in the mantle and be carried upward in solution, eventually being transported into the crust. It is thus probable that the mantle is relatively depleted, and the crust relatively enriched, in minor and trace elements. This is certainly true for uranium and thorium, because the amount of these elements in the crust is almost sufficient to account for the total amount of heat flowing out of the Earth.

The Earth's crust. The crust is a comparatively thin shell on the surface of the Earth and makes up less than 1 percent of its total mass. Its geochemical significance is only marginally related to its bulk, however. It has been

Table 3: Estimated Composition of the Earth's Mantle
(weight percent)

	estimated from peridotite nodules	calculated from three parts peridotite and one part basalt	based on the composition of oceanic basalts	estimated from peridotite nodules	calculated from a meteoritic Earth model
Silicon dioxide (SiO_2)	42.86	45.16	45.1	44.2	48.09
Titanium dioxide (TiO_2)	0.33	0.71	0.5	0.1	0.13
Aluminum oxide (Al_2O_3)	6.99	3.54	4.1	2.7	3.02
Chromic oxide (Cr_2O_3)	0.18	0.43	0.3	0.30	0.55
Ferric oxide (Fe_2O_3)	0.36	0.46	2.0	1.1	—
Ferrous oxide (FeO)	8.97	8.04	7.9	7.3	12.71
Manganese oxide (MnO)	0.14	0.14	0.2	0.15	0.43
Nickel oxide (NiO)	0.20	0.20	0.2	0.20	—
Magnesium oxide (MgO)	35.07	37.47	36.7	41.3	31.15
Calcium oxide (CaO)	4.37	3.08	2.3	2.4	2.32
Sodium oxide (Na_2O)	0.45	0.57	0.6	0.25	1.13
Potassium oxide (K_2O)	0.003	0.13	0.02	0.015	0.13
Phosphate (P_2O_5)	—	0.06	0.1	—	0.34

subjected to extensive investigations, and it provides the raw materials on which civilization depends. It is the most diverse of the geospheres, being a complex mosaic of many rock types—igneous, sedimentary, and metamorphic—each with a wide variety of chemical and mineralogical compositions. The surface is veneered with soils, related in composition to the rocks from which they formed, but with important modifications because of the smaller grain size, the presence of organic matter, and an intricate complex of living organisms. Ultimately, man's welfare and indeed his survival depends on the wise utilization of the materials in the crust. Modern civilization has been erected upon the exploitation of fuels and ore deposits, which are simply geochemical concentrations of useful elements.

Igneous rocks. Clarke estimated that 95 percent of crustal rocks are of igneous origin (formed from molten silicate masses, or magmas). Sedimentary rocks occur as a thin veneer on an igneous or metamorphic basement, except where locally thickened in mountain belts. The primordial rocks of the crust must have been essentially igneous, and the first sedimentary rocks were derived from them by processes of weathering and erosion. Metamorphic rocks are formed from both sedimentary and igneous rocks by transformations due to heat and pressure at depth in the crust; unless very intense, these transformations do not totally obliterate the primary igneous or sedimentary features. Igneous rocks show a wide range of composition; the principal component, silica (SiO_2), ranges from about 35 percent to 80 percent among the commoner igneous rocks, and other components also show a wide variation. They thus illustrate some quite extensive geochemical fractionations of the elements, the fractionations that may have economic significance if they bring about the findings of workable ore deposits.

Geochemical distinction between oceanic and continental crust

In 1924 a comprehensive review of igneous rock composition based on compilation of over 5,000 superior analyses was published. This was in many respects the ultimate refinement of Clarke's initial review of 1889. It confirmed that the averages of analyses from different continental areas are essentially identical. It also revealed significant geochemical differences between the continental and oceanic crusts. The average of igneous rock analyses from the oceanic islands is notably lower in silica and alkalies, and higher in magnesium and calcium oxides, than the continental averages. This is simply a reflection of the fact that most oceanic islands, such as Hawaii, consist almost entirely of basalts (averaging about 50 percent silica), whereas continental areas include large granitic masses, with silica contents around 70 percent. In terms of volumes, igneous rocks consist predominantly of two great types, granitic and basaltic. The former essentially are confined to the continents and the latter occur in both continents and ocean basins. The other types of igneous rocks, while many and varied, are quantitatively insignificant and hardly affect the averages. Thus, for the major elements, the average of over 5,000 analyses of igneous rocks is not significantly different from the simple average of two individual rocks, a granite (G-1) and a basalt (W-1). This can be seen from Table 1, by comparing G-1 and W-1 values with those in the column headed "Earth's crust."

The figures for the specific granitic (G-1) and basaltic (W-1) rocks are included in Table 1 because they have been analyzed for practically all the elements in different geochemical laboratories throughout the world. The rock G-1 was a granite from Westerly, Rhode Island, and W-1 was a basaltic rock (specifically a diabase) from Centerville, Virginia. Several hundred pounds of each of these rocks were crushed to a fine powder in the laboratories of the U.S. Geological Survey and samples distributed to analytical laboratories throughout the world, in order to obtain as many analyses as possible. The results were then critically examined, and the figures given in Table 1 are considered to be the best available in terms of accuracy and precision. G-1 and W-1 are undoubtedly the most thoroughly analyzed rocks and now serve as basic geochemical standards. It must be borne in mind that they are individual rocks, however, and cannot be considered to be averages of all granites and all basalts. Indeed, it is clear that G-1 is unusually rich in some trace elements (*e.g.*, thorium).

The eight principal components of the crust

Perhaps the most significant feature of the composition of the earth's crust is that it is dominated by comparatively few elements. Only eight—oxygen, silicon, aluminum, iron, calcium, magnesium, sodium, and potassium—are present in amounts greater than 1 percent, and these eight make up almost 99 percent of the whole. Of these, oxygen comprises almost 50 percent by weight. The dominance of oxygen is even more marked when weight percentages are converted to atomic percentages, as follows: oxygen (weight percentage, 46.6; atomic percentage, 62.2), silicon (27.7; 21.2), aluminum (8.13; 6.47), iron (5.00; 1.92), calcium (3.63; 1.94), magnesium (2.09; 1.84), sodium (2.83; 2.64), potassium (2.59; 1.42). This comparison, of course, merely emphasizes the fact that the crust consists almost entirely of oxygen compounds,

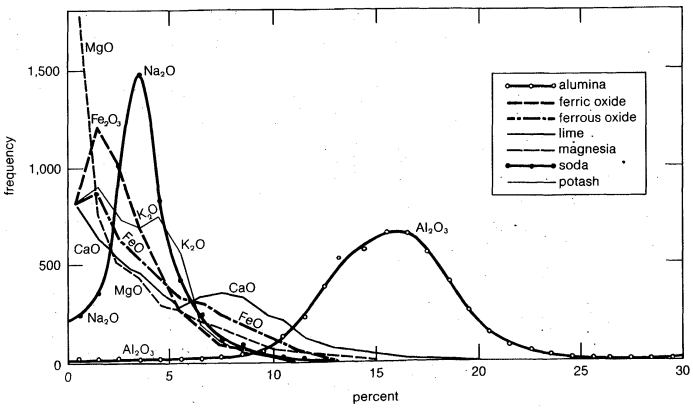


Figure 2: Frequency distribution of the major oxides in analyses of igneous rocks.

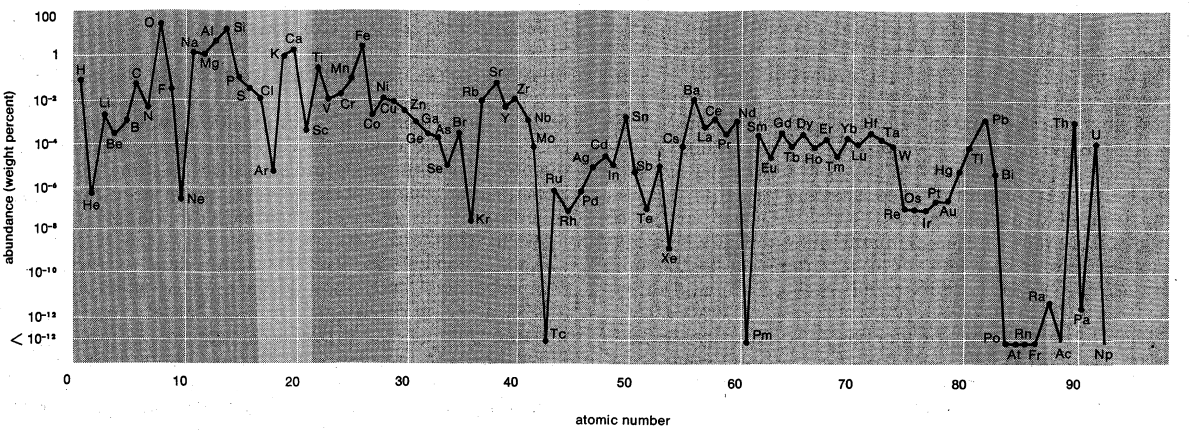


Figure 3: Crustal abundances of elements of atomic numbers 1 to 93.
From B. Mason, *Principles of Geochemistry*, 3rd ed. (1966); John Wiley & Sons, Inc.

mostly silicates and aluminosilicates of iron, calcium, magnesium, and the alkali metals. As Goldschmidt remarked, the lithosphere may well be called the oxysphere. Clarke and his collaborators calculated that the average mineralogical composition of igneous rocks is: quartz 12.0%, feldspars 59.5%, pyroxene and hornblende 16.8%; biotite 3.8%, titanium minerals 1.5%, apatite 0.6%, and other accessory minerals 5.8%.

So far only the major elements have been discussed, those present in amounts greater than 1 percent by weight. Consideration must now be given to the minor and trace elements. The data are given in Table 1, and are presented diagramatically in Figure 3.

A cursory examination immediately reveals some intriguing features in the abundance pattern. The predominance of the even-numbered elements over the neighbouring odd-numbered ones is still apparent but not so regular as in the cosmic abundances. Chemical fractionations taking place during the evolution of igneous rocks from primordial matter have clearly modified this basic relationship. The odd–even relationship is most prominent in the rare-earth elements (atomic numbers 57–71), which are chemically so similar that they are little fractionated by geochemical processes.

Minor and trace elements

A particularly noteworthy feature is that some unfamiliar elements, such as rubidium, are relatively abundant, whereas others, such as most of the industrial metals except iron and aluminum, are actually of very low abundance. Thus boron, familiar to every housewife in the form of borax cleansers and boric-acid antiseptics, and well-known in the ancient world, is an element of extremely low abundance, much lower, for example, than zirconium, and lower even than hafnium, an element first discovered as late as 1923. Mercury, another element known to the ancient Greeks and Romans, is also of extremely low abundance. On the other hand, the so-called rare-earth elements are not extremely rare. Evidently one must make a clear distinction between abundance and availability. Some elements may be of low abundance but are readily available, because geochemical processes result in their concentration in specific deposits which can be commercially exploited. Other elements that are relatively abundant may be widely dispersed in small amounts and seldom or never occur in concentrated large deposits. A typical example is titanium, which is present in practically all igneous rocks in amounts ranging up to several percent, but which, in spite of its useful properties as a metal, is still not widely used. This is partly because of its dispersed nature and partly because of the technical difficulties in extracting the element from the minerals in which it occurs.

The distribution of minor and trace elements in igneous rocks, for those of lithophile affinity, is largely controlled by their ionic radii or size. Minor and trace elements with radii similar to those of major elements can substitute for these elements in the common minerals of the igneous rocks. The crystal structures of these minerals act as sorting mechanisms, accepting those atoms of appropriate size and rejecting others. Thus rubidium, with a radius of 1.47Å (one angstrom [Å] = 10^{-8} centimetres) is incorporated in potassium feldspar, $KAlSi_3O_8$, because its radius is close to that of potassium (1.33Å). The next-higher alkali element, cesium, with a considerably larger radius (1.67Å), is not accepted into the feldspar structure; it remains in the igneous liquid during the crystallization of the major minerals until its concentration increases to such an extent that it can form the independent mineral pollucite ($CsAlSi_2O_6$).

The factor of ionic size, coupled with geochemical affinity (lithophile, chalcophile, or siderophile) is a key to the distinction between abundance and availability. Elements that are similar in size and geochemical affinity to major elements are dispersed in small amounts in common minerals; *i.e.*, rubidium in potassium feldspar, gallium in aluminum minerals, and germanium in silicate minerals. Elements that do not readily enter the common minerals of igneous rocks remain in the residual melt as crystallization proceeds. Fractional crystallization of magmas (igneous melts) normally results in a residual liquid of granitic composition. Under suitable conditions this residual liquid solidifies as a coarse-grained rock known as a pegmatite. Pegmatites are famous for their content of rare and unusual minerals, which contain many of the minor and trace elements. They are the commercial sources of lithium, beryllium, scandium, yttrium, the rare earths, cesium, niobium, and tantalum, all elements that concentrate in the residual liquid because of their specific geochemical properties.

Chalcophile elements are all of rather low abundance, and the minerals that they form, mainly sulfides and some arsenides, are not stable at the high temperatures of igneous crystallization. Sometimes these elements are found in granites and pegmatites—molybdenite (MoS_2) is a typical example. More frequently they are removed from the crystallizing magma as hot aqueous solutions and are deposited as metalliferous veins in the surrounding rock. Sometimes they may reach the surface as components in thermal springs; mercury has been deposited (as native mercury and as cinnabar, HgS) by some of these springs, occasionally in sufficient amounts for profitable mining.

Concentration of elements by sedimentary processes

Sedimentary rocks. The decomposition of pre-existing rocks by weathering, the transportation and deposition of the weathering products as sediments, and the eventual formation of sedimentary rocks might be expected to produce a gross mixture of materials, thereby working against further geochemical differentiation of the elements. This is not the case; sedimentary processes frequently produce remarkable concentrations of the elements, leading to almost pure deposits of certain minerals. Some sandstones, *e.g.*, contain over 99 percent quartz and some limestones over 99 percent calcium carbonate. The ultimate is reached in salt deposits, with extensive beds of anhydrite ($CaSO_4$), gypsum ($CaSO_4 \cdot 2H_2O$), halite (NaCl), and other compounds. Goldschmidt compared the sedimentary process with a quantitative chemical analysis, involving the successive separation of specific elements or groups of elements.

Quartz (SiO_2) is highly resistant to weathering and accumulates as deposits of sand. When consolidated these deposits form sandstones, an important group of sedimentary rocks. Under special conditions almost any mineral may be deposited in sand-sized grains, but most minerals are eventually decomposed by weathering. A few resistant ones may survive and be sufficiently concentrated to form economic deposits known as placers; the most familiar are probably the gold-bearing sands, important sources of this element, but sand deposits may have economic concentrations of zirconium (as the mineral zircon, $ZrSiO_4$), titanium (as rutile, TiO_2, and ilmenite, $FeTiO_3$), tin (as cassiterite, SnO_2), and others.

The aluminosilicates of igneous rocks, mainly the feldspars, $(K,Na)AlSi_3O_8$ and $(Na,Ca)(Al,Si)_4O_8$, are relatively easily decomposed by weathering. The alkali elements and calcium are largely carried away in solution, whereas the aluminum and silicon are quickly redeposited as insoluble clay minerals. When consolidated, these minerals form shales and mudstones. The ferromagnesian minerals undergo a more complex decomposition, sometimes leading to the deposition of iron-rich sediments consisting largely of hydrated ferric oxide; such sediments are valuable iron ores in many countries.

Calcium is carried away in solution mainly as calcium bicarbonate, $Ca(HCO_3)_2$. Most of it eventually reaches the sea, where it is utilized by a vast variety of organisms as skeletal material in the form of calcite and aragonite (polymorphs—different forms—of $CaCO_3$). Accumulation of skeletal materials after death of the organisms has formed extensive deposits of limestone throughout geological time. Magnesium in seawater can react with calcium carbonate to form dolomite, $CaMg(CO_3)_2$, and in this way some magnesium is removed from solution and deposited in sediments.

Much of the magnesium, however, remains in seawater, which is essentially a dilute solution of magnesium, calcium, sodium, and potassium chlorides and sulfates, with many other elements in small amounts (Table 1). Under special geological circumstances bodies of seawater can be cut off from the open ocean, and under arid conditions the water will evaporate and extensive salt beds be deposited. Such conditions have occurred in different regions throughout geological time, and the resulting salt deposits are economically important as sources of sodium, potassium, calcium, magnesium, chlorine, and sulfur (see EVAPORITES).

The three major groups of sedimentary rocks are sandstones, shales, and the carbonate rocks (limestones and dolomites). Much less geochemical research has been devoted to sedimentary rocks than to igneous rocks, and the data for their contents of minor and trace elements are therefore less extensive. The figures in Table 1 show that the minor and trace elements generally tend to be more concentrated in shales than in the sandstones and carbonate rocks.

Average composition of sedimentary rocks

The problem of arriving at an average composition for all sedimentary rocks is still largely unresolved, largely because of uncertainty in the relative amounts of shales, sandstones, and carbonate rocks. From geochemical arguments Clarke estimated the relative percentages of these three groups as 80:15:5, respectively. Actual measurements of sedimentary rocks suggest that these figures overestimate the amount of shales and underestimate that of limestones, however. Thus, a compilation of the recorded amounts of shales, sandstones, and limestones in over 700,000 feet of sedimentary rock formations gave relative percentages of 46:32:22, respectively. The identification of a formation as a limestone, a sandstone, or a shale, however, is likely to be gross; shales usually contain considerable sand, sandstones may carry much clay, and the term limestone is applied to many rocks with 50 percent or less of carbonate. It does appear that limestones are more prominent in the geological record than might be expected from geochemical calculations, however; this probably reflects the fact that shallow-water environments are the great places of carbonate deposition, whereas the ocean deeps are the repository primarily of clay-rich sediments.

Metamorphic rocks. Comparatively few investigations have been made of the elemental composition of metamorphic rocks. Many of these rocks retain the geochemical features of their parent igneous or sedimentary materials, and their bulk composition has been little changed despite complete recrystallization and the production of new minerals and structures in some instances. Some metamorphic rocks, however, have been markedly modified by the removal of some components and the addition of others.

The Canadian Shield

The Geological Survey of Canada has performed a comprehensive study of a large area of the Canadian Shield, a region of complex geology largely made up of metamorphic rocks. From a collection of more than 8,000 bedrock samples, the average abundances of all the major elements and a number of minor and trace elements were determined; the figures are given in Table 1. As might be anticipated, the average composition is not very different from the average composition of igneous rocks. It does show a somewhat higher silicon content, probably reflecting a preponderance of granitic over basaltic rocks and a relative abundance of quartz-rich sedimentary rocks in the original makeup of the Canadian Shield. The general validity of these abundance figures for metamorphic rocks has been confirmed by a similar study of the average composition of metamorphic rocks in the U.S.S.R., which has given closely comparable results.

Ore deposits. An ore deposit, in its simplest terms, is a portion of the Earth's crust from which some industrial raw material can be extracted at a profit. As such, its characteristics are as much economic as geochemical. Nevertheless, its formation required the operation of geochemical processes to produce the concentration of a specific element or elements in a particular place. Economics decide whether this concentration is rated as an ore deposit or merely as a deposit of scientific interest. The economics may change with time, depending upon price, availability of transportation, cost of labour, and other factors.

Some general principles can, however, be enunciated. Proceeding from the average abundance of an element in the crust, and the minimum abundance that can be profitably exploited under normal circumstances, a factor of enrichment necessary to produce an ore deposit can be derived (Table 4). The economic control is immediately evident in the approximate relation between the factor of enrichment and the price of the product sought. The most extreme example of this is in diamond mining, where the product sought may be present in the rock mined in as low a concentration as 1 part in 50,000,000. Ease of extraction, of course, plays an important role in this. Diamonds are readily separated from the great mass of waste rock by a relatively simple and inexpensive process. Magnesium is commercially extracted from seawater, where its concentration is 0.13 percent, rather than from the common rock dunite, where its concentration is about 25 percent, because of the ready availability of seawater and the cheapness of the extraction process.

Table 4: Concentration Factors for Ore Bodies of the Commoner Metals

metal	percent in Earth's crust	minimum percent profitably extracted	enrichment factor necessary for an ore body
Aluminum	8.13	30	4
Iron	5.00	30	6
Manganese	0.10	35	350
Chromium	0.02	30	1,500
Copper	0.007	1	140
Nickel	0.008	1.5	175
Zinc	0.013	4	300
Tin	0.004	1	250
Lead	0.0016	4	2,500
Uranium	0.0002	0.1	500

Ore deposits may be found in all types of rocks—igneous, sedimentary, and metamorphic—and seawater

is also a significant source of such elements as sodium, chlorine, magnesium, and bromine. There are many possible processes of geochemical enrichment leading to the formation of ore deposits, and they are often the end result of a complex series of such processes acting over a long period of time. The economic importance of ore deposits has ensured their thorough study by all techniques of geological and geochemical research, but much controversy still exists regarding the origin of many of the more complex deposits.

The most readily understood ore deposits are those of sedimentary origin. They have been formed at the surface of the Earth by processes that can usually be observed operating at the present time and that can readily be simulated in the laboratory. Salt deposits are one kind whose origin is clearly amenable to such an approach. As long ago as 1849 an Italian scientist initiated laboratory studies on the evaporation of seawater and elucidated the sequence of crystallization of the different salts. Comparison of the results with the mineralogy of salt deposits revealed gross similarities but also important differences; these differences can be explained by a variety of mild metamorphic reactions resulting from burial of these deposits under overlying sediments.

Sedimentary iron-ore deposits

Some sedimentary deposits are not readily explicable by such an approach, however. The most extensive and economically important are the vast Precambrian iron-ore deposits, which are a major source for the hundreds of millions of tons of steel produced annually. They occur on all the continents (except perhaps Antarctica) and are uniformly of great age (about 1,900,000,000 years or older). Probably the most extensive and best exposed of these are in the Hamersley Range of Western Australia, where individual beds of iron ore are continuous over hundreds of square miles in a horizontally bedded sequence of iron ore and quartzite thousands of feet thick. The conditions that gave rise to these deposits were apparently unique to this early period in Earth history, because similar deposits are not known in younger geological formations. It has been argued that the explanation lies in an oxygen-free reducing atmosphere in early geological times, under which iron could readily be transported in solution as ferrous compounds to the ocean or large lakes, where deposition eventually took place, perhaps through the agency of primitive organisms. As soon as free oxygen appeared in the atmosphere, 1,000,000,000 to 2,000,000,000 years ago, the geochemical cycle for iron was profoundly modified, and this type of transportation and deposition ended forever.

Processes other than fractional crystallization from igneous melts also give rise to magmatic ore deposits. Economic deposits of the oxide mineral chromite ($[Fe,Mg][Cr,Al]_2O_4$), for example, occur almost entirely as bands or lenses in magnesium-rich igneous rocks. Chromite evidently crystallizes early from a magma, and, being of higher density than the liquid, it sinks to the bottom of the magma chamber and becomes concentrated as almost pure bodies of this mineral. Some accessory minerals of igneous rocks are important sources of metallic elements, but the rocks cannot be mined directly because the grade is too low. If these minerals are chemically and mechanically resistant, weathering and transportation may eventually concentrate them into workable deposits. A large proportion of the world's zirconium, hafnium, rare earths, and thorium, and some iron and titanium, come from such deposits in river and beach sands.

A large number of important ore deposits occur in metamorphic rocks. The ultimate origin of these deposits is frequently obscured by the complex processes they have undergone. If it can be established that the enclosing metamorphic rocks were of sedimentary origin, the question then arises whether the ore material was deposited along with the sediments or was introduced by circulating solutions during the metamorphism or possibly at some later time. The answer is seldom clear-cut, and such deposits continue to excite lively controversy among geochemists and economic geologists.

Mineral fuels. The mineral fuels—coal, petroleum, and natural gas—may be described as a special type of economic deposit. Geochemically they represent the concentration of carbon and hydrogen by processes that were initially biological in nature. Coal is essentially the product of accumulation of land plants in large amounts, and petroleum and natural gas are the products of marine organisms (although the origin of some petroleum and natural gas under nonmarine conditions cannot be entirely excluded). The origin of petroleum and natural gas presents a more difficult problem than coal because they are fluids and thus are free to migrate from their place of origin.

Formation of coal

The formation of coal is a relatively straightforward geochemical process that can readily be traced through its successive stages. The first requirement is a geological one—the rapid accumulation of plant material under conditions that inhibit its decomposition, followed by its burial under inorganic sediments such as shales and sandstones. The great coal-forming period in the Northern Hemisphere followed the Devonian Period (345,000,000 to 395,000,000 years ago), when abundant land plants first appeared, and has been named the Carboniferous Period (280,000,000 to 345,000,000 years ago). During this period, large areas in North America and Europe were evidently low-lying swamps that supported a lush vegetation. This vegetation died, accumulated in successive layers, and was partly decomposed by bacteria and other organisms to form peat. Burial of peat deposits under inorganic sediments brought an end to the period of bacterial decomposition, and the further changes to coal were essentially a mild metamorphism caused by an increase in temperature and pressure.

Chemically, this mild metamorphism was in large part the expulsion of carbon dioxide and water from the coal-forming substance. The main trend in the change from peat through lignite to bituminous coal and anthracite is the decrease in oxygen content and the increase in carbon (Table 5). If carried to its ultimate conclusion the product would be pure carbon in the form of graphite. This occurs comparatively rarely, but evidence for it is the presence of small amounts of graphite in many metamorphic rocks.

Table 5: Average Composition of Fuels

	carbon	hydrogen	nitrogen	oxygen
Wood	49.65	6.23	0.92	43.20
Peat	55.44	6.28	1.72	36.56
Lignite	72.95	5.24	1.31	20.50
Bituminous coal	84.24	5.55	1.52	8.69
Anthracite coal	93.50	2.81	0.97	2.72

Source: Clarke, *The Data of Geochemistry.*

Coal also contains inorganic material that appears as ash when it is burned, and some coal ashes show a remarkable concentration of unusual elements. This was demonstrated by Goldschmidt in 1933, when he found appreciable amounts of germanium in some coal ashes. The Hartley Seam of the Durham Coalfield in England contains so much germanium that the ash has a brilliant yellow colour because of the presence of the oxide (GeO_2).

The source of these minor and trace elements (Table 6) and their mode of incorporation in the coal are still not fully understood. There are three possibilities: (1) these elements were taken up by the plants during growth; (2) they were carried into the coal swamp as a component of the inorganic sediments; or (3) they were absorbed during or after the coal-forming processes from circulating solutions. The first possibility is not favoured, because growing plants seldom incorporate appreciable amounts of nonorganic elements. The second possibility also is unlikely, because there is no correlation, or rather an inverse correlation, between ash content and trace element concentrations. This leaves the third possibility as the most likely one. The presence of a large amount of carbonaceous matter means that the coal-forming environment is a highly reducing one, which will favour the precipitation of some elements; the presence of hydrogen

sulfide and sulfide ions will cause the precipitation of chalcophile elements (with affinity for sulfur); and complex organic compounds are noted for their absorptive, or chelating, capacity for metallic ions. Thus several individual reactions are potentially available for the fixation of foreign elements in the coal substance.

Table 6: Rare Elements in Coal Ash

	average content in coal ash (g/ton)	average content in Earth's crust (g/ton)	factor of enrichment
Boron	600	10	60
Germanium	500	1.5	330
Arsenic	500	2	250
Bismuth	20	0.2	100
Beryllium	45	2.8	16
Cobalt	300	25	12
Nickel	700	75	9
Cadmium	5	0.2	25
Lead	100	13	8
Silver	2	0.1	20
Scandium	60	22	3
Gallium	100	15	7
Molybdenum	50	1.5	30
Uranium	400	2.7	150

As pointed out above, the origin of petroleum is not as readily elucidated as the origin of coal, because petroleum can migrate from the region in which it was formed. Indeed, the very occurrence of a commercial oil field probably implies the concentration of the petroleum from a large volume of source rocks into a relatively much smaller reservoir.

The fact that petroleum is almost always found in marine sedimentary rocks has long been a basic argument in favour of a marine origin for this material. It is certainly true that some oil has been found in igneous and metamorphic rocks, but migration from a sedimentary source bed is a reasonable explanation for these occurrences. Proof of a marine origin has been forthcoming in recent years by the sensitive analyses of recent marine sediments, which show that they contain small amounts of petroleum hydrocarbons, evidently generated either directly by marine organisms or by their subsequent decomposition.

Composition of petroleum

Natural petroleum is a complex mixture of hundreds of different hydrocarbons, but its bulk composition is remarkably constant, about 85 percent carbon and 15 percent hydrogen. It may include small amounts of organic compounds containing oxygen, sulfur, and nitrogen. Its content of other elements is exceedingly small. Petroleum ash, unlike coal ash, is not noted for its trace element content. Some petroleum ash contains appreciable amounts of vanadium, however, and has been utilized as a source of this element. A class of nitrogen-bearing organic compounds known as porphyrins include a metal atom in their molecular structure; usually this atom is iron, but other elements in this region of the periodic table, especially vanadium, nickel, and copper, may play a similar role. The vanadium content of some petroleum ash probably originates as a vanadium porphyrin in some of the organisms involved in petroleum formation.

Petroleum is always accompanied by natural gas, but many natural gas fields have no petroleum associated with them. This can probably be ascribed to greater possibility for migration for a gas as compared to a liquid. It is also possible that some natural gas is generated from coal deposits. Natural gas consists largely of methane, but small amounts of more complex hydrocarbons may be present, and it may also contain unwanted components such as nitrogen, carbon dioxide, and hydrogen sulfide. Natural gas containing hydrogen sulfide is known as "sour gas" and for long was an undesirable material because of the noxious nature of this compound; recently, however, it has been found profitable to extract the hydrogen sulfide by converting it to sulfur and then utilize the hydrocarbons.

Natural gas is the sole source of one element, helium, the industrial demand for which has steadily increased in recent years. Comparatively few occurrences of natural gas contain sufficient helium for the extraction to be commercially profitable. Currently, the Western world's need for helium is largely met by its extraction from wells in western Texas. The explanation for this local concentration of helium in some natural gas fields is still a matter for discussion; the only reasonable source is from the disintegration of radioactive elements in the crust, but the mechanism of concentration remains something of an enigma.

Helium in natural gas

Soils. Soil is a thin veneer that forms a discontinuous cover on the land areas of the Earth. Its volume and its mass are small in comparison to the major geospheres, but it is of vast importance to man. Superficially it might be considered merely as comminuted (pulverized) and decomposed bedrock; however, this viewpoint takes into account only its inorganic components and completely neglects the complex of organic compounds, living organisms, water, and included gases that gives the soil its characteristic properties and its value as the abode of life. Comminuted bedrock alone is not soil in any real sense; the Antarctic continent, where not ice-covered, has a surface of comminuted bedrock, as does the Moon, but neither material can properly be termed soil.

Soils result from the weathering of rocks, and hence their composition might be expected to reflect the composition of the rocks from which they were formed. This is true only in a very broad sense, however. Environmental factors play an important part in soil formation. The same parent rock may give rise to very different soils under different conditions. Climate, topography, vegetation, biological activity, and time are all important factors in determining the nature of a soil. Climate is probably the most important of these, as can be demonstrated by contrasting the soils developed on the same rock type under tropical and temperate conditions. In general, the soil in the humid tropics will be different in texture and composition and much less fertile, as a result of the intense leaching brought about by high rainfall, high temperatures, and the almost complete removal of organic matter by micro-organisms.

The complex of inorganic compounds, organic compounds, water, and air that makes up the soil is in a continual state of change. Water tends to dissolve and remove the relatively soluble elements such as calcium, magnesium, sodium, and potassium, and the comparatively insoluble elements—aluminum, iron, and silicon—are thereby relatively enriched in the soil. The enrichment of iron is frequently manifested by a red-brown or yellow-brown colour caused by an accumulation of iron oxides. The most reactive part of the soil is the complex of clay minerals and organic matter, which is largely responsible for its agronomic characteristics. True soil does not exist without the presence of colloidal and organic matter. The relative absence of soils in desert areas reflects the fact that chemical and physical weathering of rocks alone does not necessarily result in soil formation. Most soil processes are directly or indirectly biological in nature. Organisms and organic compounds produced by their vital activities or their decomposition are effective agents for dissolving and extracting many elements from the inorganic constituents of the soil, thereby making them available for plant growth.

Minor and trace elements in soils

Although soils do differ in composition, the range of variation in the major elements is rather small. Minor and trace elements may show considerably greater variability, as is illustrated by Figure 4. The importance of certain trace elements in the soil for the healthy growth of plants, and through the plants, of the animals that graze on them, has become increasingly apparent in recent years. Most soils contain these trace elements in sufficient amounts, but when deficiencies are present, puzzling diseases appear, which in the past have rendered large areas of otherwise suitable land unavailable for farming. On a large area in the North Island of New Zealand, for example, although it grew satisfactory pasture, sheep and cattle failed to thrive and eventually died if not removed. As a result, much of this area was given over to afforestation. It was eventually discovered that

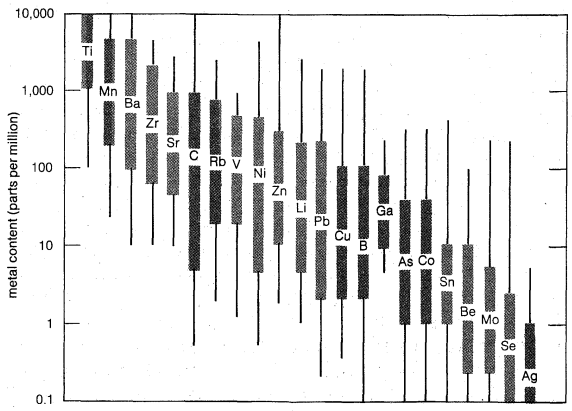


Figure 4: Range of the content of some minor elements in normal soils. Thin lines indicate ranges of more unusual values.

From *Chemistry of the Soil* by Firman Bear. Copyright 1964 by Litton Educational Publishing, Inc.

cobalt, in the amount of a few parts per million, would completely eliminate the disease when applied in fertilizer or administered directly to the animals. The ultimate explanation is the need of animals (but not plants) for vitamin B_{12}, which contains an atom of cobalt in its structure.

Occasionally, an excess of a specific element may have a deleterious effect on plant growth. Most obvious, of course, are the alkaline or saline soils of desert and coastal areas on which only an impoverished vegetation exists. Magnesium-rich soils are notably infertile; such soils develop on areas of ultrabasic igneous rocks consisting largely of olivine, $(Mg,Fe)_2SiO_4$, and the boundaries of these areas can frequently be readily mapped from aerial photographs by the marked change in vegetation. Sometimes plants take up available trace elements in amounts deleterious to animals grazing on them. A well-known example is *Astragalus racemosus* (locoweed), which in some areas of the western U.S. contains sufficient selenium to be poisonous to grazing animals.

The possible correlation between soil geochemistry and the geographical distribution of disease is thus a field of extreme significance which as yet has been insufficiently studied. The problem is a complex one, in large part because of the difficulty in isolating the numerous factors involved.

The hydrosphere. The hydrosphere is the discontinuous shell of water—fresh, salt, and solid—on the surface of the Earth. As such it comprises the oceans and the connecting seas and inlets, the lakes, rivers, and streams, the groundwater that feeds them, and the snow and ice cover of high altitudes and high latitudes. The mass of the ocean waters far outweighs the other parts of the hydrosphere. Goldschmidt estimated that there are 273 litres of water in all its forms for every square centimetre of the Earth's surface made up as follows:

	Litres	Kilograms	Total mass
Seawater	268.4	278.1	1.4×10^{18} tons
Freshwater	0.1	0.1	5.1×10^{14} tons
Continental ice	4.5	4.5	2.3×10^{16} tons

Seawater thus makes up over 98 percent of the total mass of the hydrosphere, and its composition (Table 1) essentially can be taken as giving the average composition of the hydrosphere.

Research during the past century has demonstrated that the composition of seawater is essentially uniform and that the relative proportions of the various ions are practically constant. In the open ocean the salinity (approximately the total weight of dissolved solids per kilogram) averages about 35 parts per thousand, but may rise to 40 parts per thousand in regions such as the Red Sea and the Persian Gulf, where rainfall and inflow are low and evaporation high. Sodium chloride is the dominant compound of the salts in solution and comprises about three-quarters of the whole; the remainder consists largely of chlorides and sulfates of magnesium, calcium, and potassium.

Though many data on minor and trace elements in sea water (Table 1) are available, the interpretation of these data is subject to some uncertainties. The concentrations of these elements are probably more variable than for the major elements and may depend to some degree on the sampling location. This is particularly true for the elements that are utilized by marine organisms. Phosphorus is a good example; it is markedly depleted in surface and near-surface waters by biological activity, but it enriches the deeper parts of the ocean through the dissolution of dead organisms. Silica is brought into the ocean in large amounts in solution in river water, but most of it is soon removed to become the skeletal material of diatoms, radiolaria, and sponges.

Chemical composition of sea water

Seawater also contains dissolved gases in variable amounts. Seawater of normal salinity at 0° C (32° F) in equilibrium with the atmosphere will contain about eight millilitres per litre of dissolved oxygen and 14 millilitres per litre of dissolved nitrogen. Dissolved nitrogen is essentially an inert constituent, but dissolved oxygen plays a fundamental role in the growth and decay of organisms and so varies greatly in concentration from place to place. In stagnant regions that are rich in decaying organic matter the water may be completely depleted in free oxygen, and a considerable concentration of hydrogen sulfide may be present. Much of the Black Sea below a depth of a few hundred metres is in this condition.

Another dissolved gas of prime importance for biological activity is carbon dioxide. The conditions controlling its concentration are quite complex, however, because in solution it can be present as free carbon dioxide, as undissociated carbonic acid, as carbonate ions, and as bicarbonate ions. The concentration of these ions will also be affected by biological activity and by the precipitation or dissolution of calcium carbonate.

The circulation of water through the hydrosphere is controlled in large part by the reservoir effect of the oceans. Evaporation from the ocean surface is precipitated as rainfall. Of that falling on the land, some is directly re-evaporated, some is absorbed into the reservoir of groundwater, and some flows off directly into rivers and streams. The total annual rainfall on the Earth is estimated to be 123×10^{18} grams, and the total annual runoff to the oceans 32×10^{18} grams.

Even the purest rainfall contains some material in solution, not only dissolved gases but also nonvolatile material. Rainfall near seacoasts always contains some sodium chloride and small amounts of other marine salts, the concentration of which falls off generally with distance from the ocean. Rainfall in industrial regions may of course contain a variety of pollutants; in many areas it is essentially a dilute sulfuric acid solution. Such material may also be carried far beyond the place of origin; acid rainfall in the Scandinavian countries probably originates in part from England and Germany.

Elements carried in solution by rivers

The runoff from the land contains additional material in solution, picked up during its circulation through the crustal rocks. River water averages about 120 parts per million dissolved solids, but the range is great, from about 10 parts per million up to several thousand parts per million. Commonly, the range is from 50 to 200 parts per million; contents greater than 200 parts per million are usually the result of human activities or of drainage from soils containing soluble salts, as in desert regions.

With an average content of 120 parts per million dissolved matter, the rivers of the world deliver 3.9×10^9 tons of material in solution to the sea each year. The average concentration of the important constituents (in parts per million) is: bicarbonate, 58.4; sulfate, 11.2; chloride, 7.8; nitrate, 1.0; calcium, 15.0; magnesium, 4.1; sodium, 6.3; potassium, 2.3; iron, 0.67; and silica, 13.1. Although these ten constituents account for most of the dissolved material, many other elements have been detected in river and lake waters. Comparatively little information is available for these elements, but the results of an extensive study of the large rivers of North America is summarized in Figure 5. The figures are in micrograms per litre—that is, in parts per thousand million.

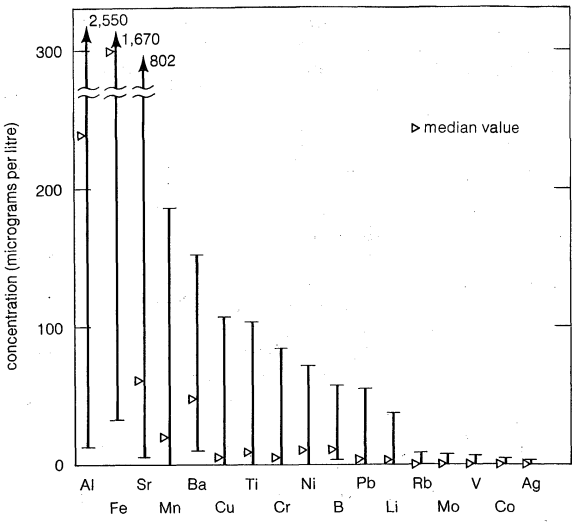


Figure 5: Concentrations of minor and trace elements in river waters of North America.

From W.H. Durum and Joseph Haffty, *Geochimica et Cosmochimica, Acta*, vol. 27 (1963); Pergamon Press

The 3.9×10^9 tons carried annually in solution to the oceans are but a small fraction of the total amount of material in solution in the oceans. Nevertheless, when integrated over the whole of geological time, more than 4×10^9 years, it greatly exceeds the present material in solution. Some of the material, especially sodium chloride, is of course cyclical, being circulated from the oceans to the land as aerosols and incorporated in marine sedimentary rocks and ultimately in large part being returned to the oceans in runoff.

Goldschmidt made an interesting calculation on the geochemical balance in seawater. From the amount and composition of sedimentary rocks he estimated that erosion during geological time had amounted to about 160 kilograms of igneous rock per square centimetre of the Earth's surface. Combining this figure with the amount of seawater per square centimetre, 273 kilograms, he derived a figure of 600 grams of igneous rock eroded per kilogram of seawater. Assuming this 600 grams had gone fully into solution (obviously a gross simplification but a limiting one), he drew up a balance sheet between the amounts of different elements potentially supplied to the oceans and the amounts actually present. Some of these figures are presented in Table 7. Despite the imperfections of the method, the results are certainly significant in a qualitative sense. Some elements—chlorine, bromine, boron, and sulfur—are present in seawater in amounts far in excess of those that can have been derived by erosion. The source of these "superabundant" elements has probably been volcanism and related magmatic activity. Halides, sulfates, and borates are deposited by volcanic gases and carried in solution in hot springs. The relative depletion of fluorine with respect to chlorine in seawater can be ascribed to the precipitation of highly insoluble fluorine-bearing compounds, mainly apatite (calcium fluophosphate). Sodium clearly remains in solution to a much greater extent than potassium; the latter element reacts with sedimentary materials to form insoluble potassium-bearing silicates such as illite and glauconite, which have no sodium-bearing analogs. Calcium is removed from solution much more effectively than strontium, evidently because it is utilized by organisms. Goldschmidt pointed out that many highly poisonous elements, such as arsenic and selenium, have been potentially supplied in dangerous amounts. Their concentration remains very low, however, presumably because of efficient processes of removal as insoluble compounds. Adsorption on colloidal particles of clay and iron oxides is a likely process.

Table 7: Geochemical Balance of Some Elements in Seawater

element	potential amount supplied to oceans (g/ton)	amount present in seawater (g/ton)	percentage in solution
Lithium	39	0.17	0.4
Boron	2	4.5	250
Fluorine	540	1.3	0.2
Sodium	16,980	10,800	64
Magnesium	12,540	1,290	10
Phosphorus	708	0.09	0.01
Sulfur	312	904	290
Chlorine	188	19,400	10,300
Potassium	15,540	392	2.5
Calcium	21,780	411	1.9
Arsenic	3	0.003	0.1
Bromine	0.97	67	6,900
Rubidium	186	0.12	0.06
Strontium	180	8.1	4.6
Iodine	0.18	0.06	33
Cesium	4	0.0003	0.008
Barium	150	0.02	0.01

Steady state of oceanic composition

The geological and geochemical evidence indicates that the ocean waters are, and have been for a long time, in a steady state of essentially unchanging composition. The addition of material by runoff from the land is adjusted by reactions within the ocean waters or between the ocean waters and sedimentary materials whereby the concentrations of the individual elements remain essentially constant. How far back in geological time this steady state has persisted remains something of an open question. The existence of most forms of marine life from the Cambrian to the present indicates a uniformity of marine conditions over the past 600,000,000 years; how far back into the Precambrian this uniformity extended is more difficult to elucidate. The earlier discussion of Precambrian iron formations suggested the possibility of very different atmospheric composition some 2,000,000,000 years ago, and the considerable interdependence of atmospheric and oceanic composition indicates that this might have resulted in marked geochemical differences in the ocean waters.

The atmosphere. The atmosphere is the most homogeneous and thus the most easily studied of the geospheres. Its mass is readily determined from the product of the average height of the mercury barometer in centimetres, the density of mercury (13.6 grams per cubic centimetre), and the area of the Earth (5.1×10^{18} square centimetres). Recent calculations give 51.17×10^{20} grams for its total mass.

The composition is also relatively simple, although a considerable number of gases may be present in small amounts (Table 8). Almost 99 percent consists of oxygen and nitrogen, with argon making up most of the remainder. Carbon dioxide, essential for plant life, is present in an extremely small amount. Some gases not listed in Table 8 may be present as local or even regional pollutants—city dwellers are becoming increasingly aware of oxides of sulfur as atmospheric pollutants, and the scientific study of smog is largely concerned with reactions taking place between hydrocarbons, oxides of nitrogen, oxygen and ozone.

Table 8: Average Composition of the Atmosphere

gas	composition by volume (ppm)*	composition by weight (ppm)*	total mass ($\times 10^{20}$ g)
Nitrogen	780,900	755,100	38.648
Oxygen	209,500	231,500	11.841
Argon	9,300	12,800	0.655
Carbon dioxide	300	460	0.0233
Neon	18	12.5	0.000636
Helium	5.2	0.72	0.000037
Methane	1.5	0.94	0.000043
Krypton	1	2.9	0.000146
Nitrous oxide	0.5	0.8	0.000040
Hydrogen	0.5	0.035	0.000002
Ozone†	0.4	0.7	0.000035
Xenon	0.08	0.36	0.000018

*ppm = parts per million. †Variable, increases with height.

The atmosphere gradually thins out into the vacuum of outer space, and its upper limit can conveniently be

placed at about 600 kilometres. An important zone in the stratosphere is known as the ozonosphere, a diffuse layer characterized by an increase in the concentration of ozone, O_3. This zone is highly important for life on Earth because it absorbs most of the ultraviolet radiation from the Sun; if this penetrated to the Earth's surface it would act as a potent sterilizer, fatal for most forms of life. It also helps to maintain a more uniform surface temperature by reducing the loss of heat by radiation to space—the so-called greenhouse effect.

History of the atmosphere

The geochemical history of the atmosphere has been a complex one. Scientists agree that the present atmosphere is quite different from the original one. It is certainly quite different from those of the other planets. It is reasonable to conclude that this reflects, in part at least, the Earth as the abode of life. The Earth's atmosphere differs from those of its neighbours in the Solar System probably in large part through the action of photosynthesis, a complex biological process which was probably preceded by a lengthy period of organic evolution.

The nature of the Earth's primitive atmosphere is still a subject of some speculation. Some scientists, reasoning by analogy with the larger planets such as Jupiter, have argued for an original atmosphere consisting largely of methane and ammonia. Others have considered that present-day volcanic gases may indicate the nature of the primitive atmosphere, in which case it contained carbon dioxide, possibly carbon monoxide, nitrogen, and water vapour. In either case, free oxygen was absent. If the evolution of the atmosphere is traced backward in time through the geological record, then extensive terrestrial photosynthesis is indicated by an abundance of land plants in Devonian times, about 400,000,000 years ago. Marine photosynthesis, however, is much older, since practically all the major groups of marine organisms were established by the beginning of the Cambrian period, some 570,000,000 years ago. As discussed earlier, the extensive Precambrian iron formations suggest an oxygen-free atmosphere which was terminated about 2,000,000,000 years ago. This evidence has been translated into estimates of oxygen content of the atmosphere of about 1 percent of the present level 2,000,000,000 years ago, about 10 percent of the present level at the beginning of the Cambrian Period, and essentially the present content by Devonian times.

Although it is not yet possible to know the quantitative composition of the primitive atmosphere, the geochemical processes that have operated to modify its composition during geological time can be evaluated. These processes can be summarized as a series of gains and losses. Additions to the atmosphere comprise: (1) gases released by igneous activity; (2) oxygen and hydrogen produced by the photochemical dissociation of water vapour; (3) oxygen produced by photosynthesis; (4) helium produced by the radioactive breakdown of uranium and thorium; and (5) argon produced by the radioactive breakdown of potassium. Atmospheric losses include: (1) oxygen removal by oxidation of ferrous to ferric iron, sulfur compounds to sulfates, hydrogen to water, and similar reactions; (2) carbon dioxide removed by the formation of coal, petroleum, and the death and burial of organisms; (3) carbon dioxide removed by the formation of calcium and magnesium carbonates; (4) nitrogen removed by the formation of oxides of nitrogen in the air and by the action of nitrifying bacteria in the soil; and (5) hydrogen and helium by escape from the Earth's gravitational field.

Photosynthesis has certainly been the most significant process in controlling atmospheric composition during much of geological time. Through this process carbon dioxide and water are converted to carbohydrate, with the accompanying release of oxygen. Much of this carbohydrate is consumed by animals and reconverted to carbon dioxide and water by respiration, and oxidative decay leads to the same result. Some, however, is incorporated into sediments; part may go to form exploitable deposits of coal and petroleum, but most of it remains as disseminated carbonaceous material; the average carbon content of sedimentary rocks is about 0.4 percent.

Quantitatively, more significant amounts of carbon dioxide have been removed from the atmosphere in the form of limestone and dolomite. Most of this removal has been effected by marine organisms, especially algae and corals, but direct inorganic precipitation may occur, especially in warm tropical waters. Judging from the vast deposits of limestone and dolomite throughout the sedimentary record, this process has operated with a remarkable degree of uniformity throughout geological time. Extensive limestone formations are possibly less common in older Precambrian rocks, indicating a slower beginning of carbonate precipitation. It is truly remarkable that so much carbonate rock has been deposited during geological time, with the carbonate being ultimately derived from an atmosphere which may never have contained a much higher concentration of carbon dioxide than is present today. It has been pointed out that reactions like the decomposition of calcium silicate—$CaSiO_3 + CO_2 = CaCO_3 + SiO_2$, in which $CaSiO_3$ is calcium silicate, CO_2 is carbon dioxide, $CaCO_3$ is calcium carbonate, and SiO_2 is silicon dioxide, which tends to go toward the right at ordinary temperatures—will act as buffering mechanisms to keep the carbon dioxide concentration of the atmosphere at a continuously low level.

Removal of carbon dioxide

If the carbon dioxide concentration has remained essentially constant, and yet this compound has been continuously extracted to form carbonates and organic compounds, then clearly a balancing source of "new" carbon dioxide is required. This evidently has been provided by volcanism and other igneous activity. The Earth is steadily being degassed, in the sense that gaseous compounds contained in the mantle are escaping to the surface. The presence of carbon dioxide in the mantle has been demonstrated by the presence of microscopic inclusions of liquid carbon dioxide in the minerals of the peridotite xenoliths (rocks contained within other rocks) brought up in some volcanoes. Along with carbon dioxide, much water and small amounts of other volatiles are being added continuously from sources in the mantle. Ultimately, the hydrosphere, as well as the atmosphere, is the product of the degassing of the Earth's interior.

Of the remaining atmospheric gases, argon presents some intriguing features. Argon is by far the most abundant of the inert gases on Earth, whereas in the universe as a whole it is much less abundant than either helium or neon. In addition, its isotopic composition is quite distinct, consisting almost entirely of argon-40, whereas in the cosmos argon-36 is the most abundant isotope. The reason for these anomalies is that atmospheric argon is almost entirely radiogenic, the product of the decay of the potassium-40 isotope of potassium.

Similarly, the helium in the atmosphere is probably entirely the product of the radioactive decay of uranium and thorium. Actually, the atmosphere contains only about 10 percent of the total amount of helium generated from these sources during geological time. Some of this helium remains occluded in the rocks where it was formed, some has escaped from the upper atmosphere. Helium (and hydrogen), consisting of light atoms, can escape from the gravitational field of the Earth, whereas heavier gases cannot. A minor source of atmospheric oxygen throughout geological time is probably the photochemical decomposition of water vapour in the upper atmosphere, with the subsequent loss of the hydrogen to outer space.

Some oxygen has been removed from the atmosphere by oxidative reactions, of which the most significant has been the conversion of ferrous to ferric iron. In igneous rocks the average ferrous-to-ferric iron ratio (FeO/Fe_2O_3) is greater than unity, whereas in sedimentary rocks the proportion is reversed, ferric iron being dominant over ferrous iron. Other oxidative reactions are the conversion of manganous compounds to manganese dioxide and of hydrogen sulfide to free sulfur and sulfate. Nitrogen is almost inert geochemically, but a little is fixed as oxides of nitrogen by lightning, and somewhat more by the action of nitrifying bacteria in the soil. Most of this nitrogen is ultimately returned to the atmosphere by the decay of the organisms. Oxides of nitrogen formed in the atmosphere are removed in rain as nitrite and ni-

trate. Nitrogen does not accumulate in the soil, however, except perhaps under extremely arid conditions, as in the deserts of northern Chile, the locale of the unique nitrate deposits.

THE GEOCHEMICAL CYCLE

The preceding discussion has shown that at the time of formation of the Earth the chemical elements already had been considerably fractionated: the universe consists almost entirely of hydrogen and helium, probably with less than 1 percent of the heavier elements. The Earth, on the other hand, consists almost entirely of the heavier elements. Helium is a very rare element on Earth, so rare in fact that it was first discovered as an unidentified line in the Sun's spectrum in 1868, some 30 years before it was detected on Earth. Hydrogen is moderately abundant on Earth, largely because it combines with oxygen to form water, whereas helium is an inert element.

Significance of inert gases

The rarity of helium and the other inert gases neon, krypton, and xenon on Earth is good evidence that the Earth formed by the accretion of small solid objects, or planetesimals. (Argon is a special case, since most of the Earth's argon has been formed within the planet by the radioactive decay of potassium.) These planetesimals had no atmosphere, and the atmosphere of the Earth has been derived by the outgassing of combined and occluded gases within these planetesimals. This process has operated throughout geological history and is probably still continuing; volcanic activity not only brings up solid material from the Earth's interior but also large amounts of gases, principally water vapour, carbon monoxide and dioxide, and nitrogen. The oxygen in the present atmosphere is almost entirely the product of photosynthesis, whereby carbon dioxide and water are converted to carbohydrate and free oxygen.

Direct information on the composition of the Earth's crust is available in the form of thousands of analyses of individual rocks, the average of which provides a reasonably precise estimate of the bulk composition. For the mantle and the core the information is indirect and thus much less precise. The origin of the Earth by the accretion of planetesimals is a well-founded hypothesis, however, and meteorites are probably examples of planetesimals that have survived from the preplanetary stage of the solar system. It thus seems likely that the Earth formed by the accretion of solid bodies with the average composition of stony meteorites. The accretion process, however, led to massive segregation of the elements. Much of the iron was reduced to the metallic state and sank to the centre to form the core, carrying with it the major part of the siderophile elements. Lithophile elements, those with a greater affinity for oxygen than iron, combined as oxide compounds, mostly silicates, and provided material for the mantle and crust. Chalcophile elements would tend to form sulfides; however, few sulfides are stable at the high temperatures of the Earth's interior, so the fate of the chalcophile elements during the early history of the Earth is somewhat uncertain.

This primary geochemical differentiation of the Earth can be interpreted in terms of the system iron–magnesium–silicon–oxygen–sulfur, because these five elements make up about 95 percent of the Earth. There was insufficient oxygen to combine with the major metallic elements iron, magnesium, and silicon; because magnesium and silicon have a greater affinity for oxygen than iron, these elements combined completely with oxygen, and the remaining oxygen combined with part of the iron, leaving the remainder as the metal iron and iron sulfide. As indicated above, the metal sank to form the core, carrying with it the major part of the siderophile elements. The iron sulfide phase probably incorporated much of the chalcophile elements; it does not seem to have formed a distinct shell within the Earth and probably remains primarily in disseminated form through the mantle and the core.

This primary geochemical differentiation was largely controlled by two independent factors, the geochemical affinity of the individual elements and the density of the principal phases. It is significant that the density of the individual elements is irrelevant in this context. Uranium and thorium, for example, are very heavy elements; nevertheless, they are concentrated in the crust, not in the core, because of their lithophile character (affinity for oxygen).

Separation of crust from mantle

The separation of crust from mantle has involved rather different geochemical processes than those of the segregation of the core. The core represents the physical separation of a high-density metallic liquid immiscible with the silicate material of the mantle and crust, whereas the separation of crust from mantle has resulted from more subtle geochemical processes. The mantle is made up very largely of the magnesium-iron silicates olivine and pyroxene (or their high-pressure equivalents). As discussed earlier, these minerals act as sorting mechanisms for other elements, accepting those of ionic size comparable to magnesium and iron and rejecting those much larger or smaller. The crust differs from the mantle in being enriched in those elements that do not readily enter the crystal structures of olivine and pyroxene. The common elements include the alkali metals (except lithium), aluminum, and calcium to some degree. These are combined in the crust in the form of the feldspar minerals, and one of the principal differences between crust and mantle is the dominance of feldspar in the former and its absence in the latter.

The differentiation of the crust from the mantle and core permits consideration of the outer part of the Earth as a distinct geochemical system. The distribution and migration of the elements within the crust (and in the hydrosphere and atmosphere) are the results of processes of magmatism, sedimentation, and metamorphism. The fate of an element during magmatic crystallization is primarily a function of its ionic size. If of appropriate size it enters the structure of one or other of the major minerals; otherwise, it tends to remain in the magmatic liquid until its concentration increases sufficiently for it to form specific minerals in which it is a major component. Volatile elements and compounds in the magma (principally water, carbon dioxide, nitrogen, and sulfur compounds) escape and are added to the atmosphere and hydrosphere.

Sedimentary processes lead to further geochemical differentiation. Liquid water is the medium in which these processes largely operate, and the solution chemistry of the individual elements is the key to their geochemical behaviour. The atmosphere and hydrosphere interact importantly with the lithosphere. Water and carbon dioxide are incorporated in sedimentary minerals; soluble ions, especially sodium, are added to the hydrosphere. Processes involving living organisms are intimately associated with sedimentation; indeed, some abundant and widespread sedimentary rocks, such as coal and limestone, are essentially biological in origin. These biological processes thus have had significant geochemical results. Photosynthesis also has been responsible for the present composition of the atmosphere.

Effects of metamorphism

Metamorphism tends in some degree to reverse the geochemical differentiation of the elements produced by igneous and sedimentary processes. In general, metamorphism promotes a more uniform distribution of the elements and tends to erase compositional differences within large rock masses. Unlimited metamorphism would result in a situation in which the whole lithosphere reached a uniform composition; such a trend is exemplified by the comparatively monotonous chemical and mineralogical composition of many extensive areas of metamorphic rocks.

This overall picture of the migration of the elements in the outer part of the Earth is the basis for the concept of the geochemical cycle (Figure 6). The geochemical cycle may be considered as beginning with the crystallization of a magma, which may produce a variety of igneous rocks and which adds volatile materials to the atmosphere and hydrosphere. Weathering decomposes the igneous rocks, and erosion eventually transports the decomposed rock to the oceans, partly as solid matter to form sedimentary rocks and partly as material in solution. Crustal movements may transport sedimentary and igneous rocks to deeper levels in the crust, where the

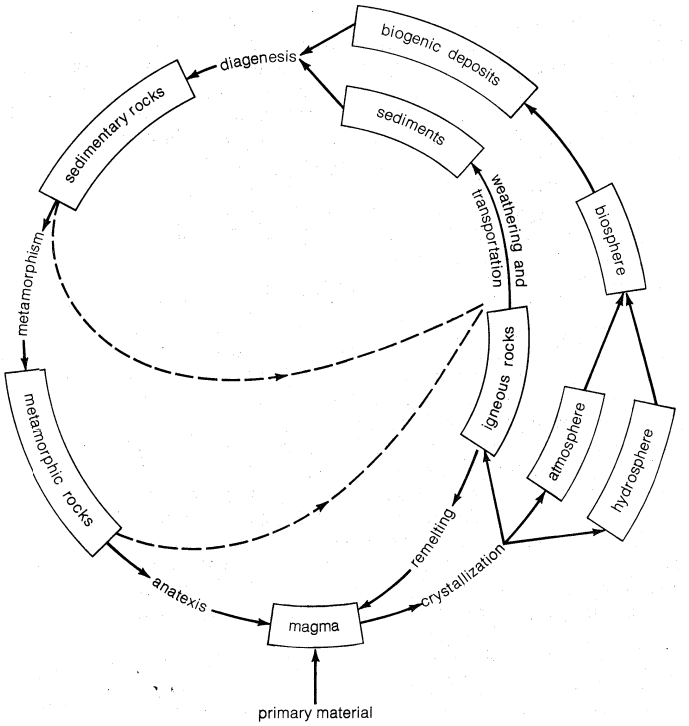


Figure 6: The geochemical cycle.
From B. Mason, *Principles of Geochemistry*, 3rd ed. (1966); John Wiley & Sons, Inc.

combined effects of heat and pressure induce metamorphism. If the temperature rises above the melting point of some of the rock material, magma is regenerated, and the cycle is complete. Such a geochemical cycle seldom goes to completion; at some stage it may be indefinitely halted or its direction reversed. Nor is the cycle a closed one, either materially or energetically. The crust receives "primary" magma generated in the upper mantle, bringing with it additional material and energy in the form of heat. The surface of the Earth receives an increment of meteoritic matter from outer space, insignificant in amount but nevertheless detectable in deep-sea deposits, where the rate of sedimentation is very low.

The geochemical cycle provides a useful conceptual framework for considering the course followed by an individual element in proceeding through the different stages. For a specific element, a complete understanding of its behaviour throughout the cycle is one of the major objectives of geochemical research. An element may tend to concentrate in a specific type of deposit at a particular stage in the cycle, or it may remain dispersed throughout the entire cycle. The geochemical cycles of a few elements have been elucidated in considerable detail, but for many of them knowledge is still fragmentary.

BIBLIOGRAPHY. L.H. AHRENS, *Distribution of the Elements in our Planet* (1965), a brief, up-to-date summary of elemental abundances in the Earth and in its various parts; F.W. CLARKE, *The Data of Geochemistry*, 5th ed., *Bull. U.S. Geol. Surv. 770* (1924), the most extensive and readily available collection of analytical data on geologic materials; M. FLEISCHER (ed.), *The Data of Geochemistry*, 6th ed. *ibid., 440* (1962), an extended, revised edition of Clarke's earlier work; V.M. GOLDSCHMIDT, *Geochemistry* (1954), an authoritative, comprehensive account of the whole field of geochemistry; B.H. MASON, *Principles of Geochemistry*, 3rd ed. (1966), a standard textbook and reference work in this field; K. RANKAMA and T.G. SAHAMA, *Geochemistry* (1950), an authoritative handbook that gives a detailed account of the distribution of each element or of related groups of elements.

(B.H.M.)

Elements, Physiological Concentration of

The physiological concentration of an element may be defined as the accumulation of that element in concentrations greater than those in the surrounding environment by the physiological processes of an organism. Concentration is usually compared, in the case of land organisms, with the content of the soil or of the Earth's crust in general and, in the case of aquatic organisms, with the concentration of elements in the surrounding water. Operating in conjunction with physical and chemical processes, biological processes may concentrate an element as much as 1,000,000 times the average abundance of that element in nature. One example of physiological concentration is illustrated by the distribution of the element carbon, which makes up about 0.03 percent of the Earth's crust. Carbon tends to combine with oxygen, principally producing the gas carbon dioxide, which then rises to the Earth's surface and is emitted in volcanic gases. Plants then concentrate this carbon, through photosynthesis, as a major constituent in their tissues. Some carbon is also concentrated in the production of calcium carbonate skeletons. A great deal of the carbon in the Earth's crust is therefore concentrated by organisms and preserved in their fossilized remains. More than 26 percent of the carbon that has been at the Earth's surface is thought to be preserved as organic material in sediments and sedimentary rocks; 73 percent is preserved in carbonate rocks that have organically produced skeletons as a major constituent.

Elements may concentrate in two basically different forms: as constituents of either organic compounds or mineralized skeletal materials; for example, a clam may concentrate carbon either in its soft tissues or in its calcium carbonate shell. The element concentrated may be a major constituent of the organism, as in the case of carbon, or it may be a trace element, such as iodine in some algae. Many, perhaps most, of the elements are concentrated by some group of organisms, but certain ones of greater importance have been more thoroughly studied and will be emphasized in this article. The reason for the concentration is not always known. The element forms part of a structural component of the organism in some cases, or it may serve a metabolic function.

Effects of physiological concentrations

Physiological concentration of elements is important in the distribution of the element in the Earth's crust, as in the case of carbon, and may also be of economic importance, as, for example, in the formation of coal deposits from carbon initially concentrated by plants. The concentration of elements is also important to human health. Saltwater fish concentrate iodine in their tissue, and man concentrates the iodine in his thyroid gland by eating these fish, thus preventing the development of goitre. On the negative side, fish also concentrate mercury, which is poisonous to man above a certain minimum level; cows concentrate strontium in their milk, including radioactive strontium 90, which may cause damage to persons who drink the milk and concentrate the strontium 90 in their bones. Because some plants concentrate certain metals, such as zinc and lead, one method of prospecting for mineral deposits utilizes chemical analysis of plants growing on the soils overlying the mineral deposit. Geologists have used the concentration of certain elements in fossilized skeletons to study the temperature and chemistry of ancient oceans.

More than 60 elements have been identified in analyses of organic matters, and the selective concentration of elements by many organisms has been established (see Table 1). Among elements consistently present in plants and animals in greater concentrations than in the environment are: carbon, nitrogen, phosphorus, calcium, magnesium, potassium, sodium, chlorine, and sulfur. In general, the most impressive and thoroughly studied and documented examples of physiological concentration of elements are the concentrations by marine organisms that will be discussed in this article. For information on the general distribution of elements in nature, see ELEMENTS, GEOCHEMICAL DISTRIBUTION OF; see also OCEANS AND SEAS and MARINE SEDIMENTS for the chemistry of seawater and sediments. The nature of the marine organisms discussed here is covered in FOSSIL RECORD.

Concentration of major seawater constituents. In terms of the amount of material involved and the permanence of the product, marine invertebrates are the most im-

Table 1: Concentration Factor for Some Metallic Elements in Marine Organisms*

	concentration factor
Titanium	10,000
Vanadium	280,000
Chromium	1,400
Molybdenum	6,000
Manganese	41,000
Iron	86,000†
Cobalt	21,000
Nickel	41,000
Copper	7,500†
Silver	22,000
Gold	1,400
Zinc	32,500
Cadmium	4,500
Gallium	800
Thallium	700
Germanium	7,600
Tin	2,700
Lead	2,600
Arsenic	3,300
Antimony	300
Bismuth	1,000

*With respect to seawater; concentration factor given as dry weight for total organism. †These figures probably are too low.

portant physiologic concentrators of elements. Many marine invertebrates have mineralized skeletons in which various elements have been concentrated from seawater. These skeletons may be fossilized and thus take large quantities of material out of circulation. Even the soft organic tissue may not entirely decompose on death of the organism but may be buried in marine sediment and partially preserved in rocks formed from sediment.

Precipitation of calcium carbonate. Many different groups of organisms concentrate calcium from seawater in the formation of calcium carbonate skeletons (see Table 2). Calcium constitutes about 0.04 percent of seawater and about 40 percent of calcium carbonate skeletons, representing a 1,000-fold enrichment. Calcium carbonate in invertebrate skeletons is precipitated in one of two mineral forms: calcite or aragonite; some organisms form skeletons composed of a combination of the two. The process by which organisms precipitate calcium carbonate skeletons is complex and is still imperfectly understood. As indicated by Table 2, practically every major phylum of organisms has members that form calcium carbonate skeletons. Further, practically all species of many of these phyla, such as the mollusks, brachiopods, and bryozoans, have calcium carbonate skeletons. Some groups of organisms, such as the vertebrates and arthropods, form skeletons of calcium phosphate, thus concentrating calcium in a different mineral form. Phosphates are quantitatively much less important than carbonates as skeletal material among marine invertebrates.

Table 2: Distribution of Skeletal Mineral Types Among Important, Skeletal-Bearing Phyla

	mineral substances concentrated					
	calcite	aragonite	calcite and aragonite	opaline silica	phosphates	other
Algae	X	X		X		*
Protozoa	X	X		X		†
Porifera	X		X	X		
Coelenterata	X	X	X			‡
Bryozoa	X	X	X			
Brachiopoda	X				X	
Annelida	X	X	X			
Mollusca	X	X	X		X	§
Arthropoda	X	X	X		X	
Echinodermata	X					
Hemichordata		X				
Chordata		X			X	

*Brucite? †Iron oxide, celestite. ‡Gypsum. §Fluorite, iron oxides.

In most places the oceans are saturated or even supersaturated (*i.e.*, calcium and carbonate ions are present in solution in sufficient or more than sufficient concentrations to react and produce calcium carbonate, $CaCO_3$) with respect to calcium carbonate, but relatively little inorganic precipitation occurs, practically all calcium being removed in the formation of skeletons. Much of the calcium carbonate skeletal material accumulates in the marine sediment; in fact, the sediment is often composed almost entirely of accumulating skeletal material. Such sediment is formed in many parts of the world in shallow, tropical waters, such as the Bahama Banks, Florida Keys, Persian Gulf, and the Great Barrier Reef of Australia. Deep-sea sediments also often consist largely of the skeletons of free-floating species of a protozoan group, the foraminiferans. Marine sediments tend to be rich in calcium carbonate skeletons in areas remote from sources of land-derived sands and muds. Such sediments may eventually become rocks and form limestone that is elevated by tectonic forces (*i.e.*, structural disturbances of the Earth's crust) to become part of the land. Calcium may thus be removed from further circulation for periods of millions of years. Weathering and erosion eventually return the calcium to solution, and it re-enters the cycle. Much of the calcium in the Earth's crust has been concentrated in the formation of limestone, largely as a result of organic processes.

Abundance of calcium carbonate

Occurrence of magnesium. Although no mineral in which magnesium is the principal metallic constituent is known with certainty in the skeletons of marine organisms, magnesium is common in the skeletons of marine invertebrates. Magnesium is chemically similar to calcium and is readily substituted for it in many minerals, including calcite (magnesium may be substituted for up to 30 percent of the calcium atoms in organically formed calcite). The presence of magnesium in the skeleton can range from barely detectable amounts to approximately 8.5 percent by weight; amounts of about 3–4 percent are common. The highest concentrations of magnesium are found in the calcite secreted by certain groups of calcareous algae, which may contain up to 8.5 percent magnesium by weight. Some of the calcareous algae are thought to contain small amounts of the mineral brucite (magnesium hydroxide) mixed with calcite, but the evidence for this is not considered conclusive. The echinoderms (starfish, sea urchins, crinoids, etc.) are also high in magnesium, containing as much as 4.7 percent. Marine mollusks with aragonite skeletons contain the lowest concentrations of magnesium, as little as 0.04 percent.

Magnesium in calcite

Although magnesium is the fifth most abundant element in seawater, it comprises only about 0.135 percent by weight; thus, the foregoing cases represent concentration by a factor of 50 or more in invertebrate skeletons. In most if not all groups of organisms, the concentration of magnesium is temperature dependent. In the echinoderm class Echinoidea (sea urchins and sand dollars), for example, specimens from cold waters usually contain only about 1.5 percent magnesium, whereas those from warm, tropical waters may have as much as 4.6 percent. On the basis of this relationship, it has been suggested that the magnesium content of fossils be used to determine the temperature of the oceans in the geologic past. The magnesium content of fossils is almost always altered, however, in the fossilization process. When invertebrate skeletons become buried in marine sediment and elevated above sea level, contained magnesium is readily leached out by groundwater. Some geochemists believe that the magnesium may then be further concentrated inorganically in the form of the mineral dolomite (*i.e.*, calcium-magnesium carbonate), one of the most common minerals in sedimentary rocks.

Strontium concentration. Rarely found in the form of a separate strontium mineral, strontium is nevertheless a fairly common constituent of invertebrate skeletons. An exception to the general occurrence of this element is the acantharian radiolarians, which have skeletons composed of the mineral celestite (strontium sulfate). Strontium is also chemically similar to calcium and is readily

substituted for it, especially in skeletons made of the mineral aragonite. Strontium concentrations are highest in aragonite corals and calcareous algae, which may contain as much as 1.0 percent strontium by weight, and probably lowest in calcite mollusks, which may contain only a trace (0.05 percent). The strontium content of seawater is only about 0.0008 percent; thus, organisms with high strontium skeletons concentrate this element by a factor of more than 1,200. In the celestite radiolarians the concentration factor is about 60,000.

Temperature dependent concentration

The strontium concentration in the skeletons of some marine organisms varies with the temperature at which the organism lives, and strontium offers more potential as a method of determining ancient marine temperatures, because it is not as readily altered in fossils as is magnesium. The ratio of strontium to calcium in the high-strontium forms is slightly higher than in seawater, indicating that the organisms concentrate strontium in their skeletons slightly more than they do calcium. According to studies of fossils, the amount of strontium in seawater has probably been nearly constant for at least the last 280,000,000 years; calcium and magnesium also appear to have been constant for at least that long. The strontium balance has probably been maintained largely by organisms removing it from seawater as fast as it is supplied by the weathering of rocks on land.

Concentration of minor seawater constituents. *Silicon.* One of the most abundant elements in the Earth's crust, silicon is present in seawater in very small but quite variable amounts (about 0.0003 percent by weight) and is concentrated by some marine organisms in the form of hydrous silicon dioxide (opaline silica). Diatoms, microscopic algae with cell walls that are usually strongly impregnated with silica, are among the most abundant of organisms on the Earth. In certain places their accumulated skeletons comprise the major portion of thick sediment. Diatomaceous earth, which is composed of diatom skeletons, has important applications as a filter material and as an insulator. Microscopic single-celled radiolarians, although not as abundant as diatoms, are widespread in the seas, and their delicate siliceous skeletons are a major constituent of marine sediments over large areas of the sea floor. Certain varieties of sponges also concentrate silica; the delicate and beautiful Venus's-flower-basket sponge from the deep sea is an especially spectacular example of a "glass sponge." The effect of the removal of silicon from the seawater by these organisms, especially the diatoms, is so great that surface waters of the seas are much depleted in silicon relative to deeper waters, where lack of light prevents growth of diatoms.

Occurrence of phosphate skeletons

Phosphorus. In a few groups of organisms phosphorus is concentrated in skeletons made of calcium phosphate minerals. The vertebrates, including the fish, are the most important group of marine organisms with phosphatic skeletons; the only nonphosphatic portions of the fish skeleton are the otoliths, or ear bones, which are calcium carbonate. Arthropods, especially the crustaceans, are important producers of phosphatic skeletons, and a few species of brachiopods, a minor phylum of shelled marine organisms that was important in the geologic past, also produce phosphatic skeletons. Biological concentration of phosphate in rocks formed from marine sediments is sometimes extensive enough to be economically important as a local source for fertilizer and other commercial products.

In addition to being concentrated in the skeletons of some organisms, phosphorus is important to all plants and animals as an essential constituent of phosphoproteins, nucleoproteins, and phospholipids. Phosphorus is present in seawater in very small and variable amounts (about 0.000009 percent), in part because it is rapidly removed by organisms and thus tends to be depleted from surface waters by organisms and concentrated in deeper waters, where there are fewer forms of life because of the lack of light. Areas where deeper waters come to the surface are likely to be very rich in life, in part because of the abundance of this important nutrient.

Iron. A very minor constituent of seawater (about 0.0000004 percent), iron is concentrated by many organisms as a constituent of certain organic molecules; its presence in the respiratory pigment, hemoglobin, is a well-known example. The skeletons of a few animal groups also contain iron minerals, and an amorphous form of iron oxide is found as the cementing agent between sand grains in the shells of some foraminiferans. Certain snails and chitons make a more spectacular use of iron minerals: as cappings for radular elements with which they scrape food (algae) from hard substrates. The iron minerals used in the radula, including magnetite and goethite, are very hard, an asset to an animal that utilizes an animated file.

Iodine. Extremely small amounts (0.000006 percent) of iodine are present in seawater; in fact, the element was discovered in marine algae long before it was detected in seawater. Concentration of iodine by marine plants is so effective that seaweeds were formerly a source for commercial iodine. Iodine is present in many organisms and ranges to 1 percent of the dry weight in the brown algae, a concentration almost 200,000 times that in the seawater. Among the brown algae, *Laminaria* and *Fucus* are especially rich in iodine. Iodine content is highest in brown algae of the temperate zone of the Northern Hemisphere and lowest in the tropics. The concentration also increases with water depth and has been observed to vary during the year, usually being highest in winter. Many other organisms also concentrate iodine: sponges, alcyonarians, and portions of fish, especially the thyroid gland and liver oil, are especially high in the element. The form and function of concentrated iodine is unknown, but it seems to exist both in organic compounds and as inorganic salts.

Vanadium. Concentration of vanadium by certain tunicates (sea squirts) is perhaps the most spectacular example of the concentration of an element from seawater. Although vanadium is extremely rare in seawater (about 0.0000002 percent, or two parts per 1,000,000,000), it may constitute several percent of the respiratory pigment of certain tunicate species (especially species of the Ascidiidae), a concentration on the order of 10,000,000 times. The bulk dry tissue of some tunicates often contains 0.1–0.2 percent of vanadium. Because the concentration of vanadium in seawater is so low, it was first thought that the tunicates obtained vanadium from some other source, perhaps from marine sediments suspended in seawater, but it has been established that the tunicates do in fact obtain vanadium from seawater. In at least one tunicate species the vanadium content of tropical specimens is considerably higher than in those from cold water, suggesting a temperature effect.

Copper. Another element highly concentrated in invertebrates, especially in the mollusks, is copper. Although it comprises only 0.0000001 percent of seawater, copper may be as high as a few tenths of a percent in some cephalopods (octopuses and squid) and clams (especially oysters). In those organisms, the element is a constituent of the respiratory pigment hemocyanin. The copper content of edible clams (oysters and mussels) has been the subject of especially extensive investigation. So-called green oysters, which are poisonous, are rich in copper. Certain of the arthropods, especially the crustaceans, also use hemocyanin as their respiratory pigment and are thus rich in copper.

Fluorine. Fluorine, another minor seawater constituent, is found in the gizzard plates in a few gastropod species in the form of the mineral fluorite (calcium fluoride).

BIBLIOGRAPHY. The literature on this topic is quite diverse; the following treatments constitute the best general sources of information: A.P. VINOGRADOV, *The Elementary Chemical Composition of Marine Organisms* (1953; orig. pub. in Russian, 1935–44), the most valuable single source of information in this field reviewing all literature through about 1950; H.A. LOWENSTAM, "Biologic Problems Relating to the Composition and Diagenesis of Sediments," in T.W. DONNELLY (ed.), *The Earth Sciences: Problems and Progress in Current Research*, pp. 137–195 (1963), the most thorough modern review of skeletal geochemistry; and E.D. GOLDBERG, "Biogeochemistry of Trace Metals," in the GEOLOGICAL SOCIETY OF

AMERICA, *Treatise on Marine Ecology and Paleoecology*, pp. 345–357 (1957), a review of the interrelationship of the chemistry of marine organisms and seawater.

(J.R.D.)

Elevator

An elevator, or lift, is a car that moves in a vertical shaft to carry passengers or freight among the levels of a multistory building. Most modern elevators are propelled by electric motors, with the aid of a counterweight, through a system of cables and sheaves (pulleys). By opening the way to higher buildings, the elevator played a decisive role in creating the characteristic urban geography of many modern cities, especially in the United States, and promises to fill an indispensable role in future city development.

History. The practice of lifting loads by mechanical means during building operations goes back at least to Roman times; the Roman architect-engineer Vitruvius in the 1st century BC described lifting platforms that used pulleys and capstans, or windlasses, operated by human, animal, or water power. Steam power was applied to such devices in England by 1800. In the early 19th century a hydraulic lift was introduced, in which the platform was attached to a plunger in a cylinder sunk in the ground below the shaft to a depth equal to the shaft's height. Pressure was applied to the fluid in the cylinder by a steam pump. Later a combination of sheaves was used to multiply the car's motion and reduce the depth of the plunger. All these devices employed counterweights to balance the weight of the car, requiring only enough power to raise the load.

Prior to the mid-1850s, these principles were primarily applied to freight hoists. The poor reliability of the ropes (generally hemp) used at that time made such lifting platforms unsatisfactory for passenger use. When an American, Elisha Graves Otis, introduced a safety device in 1853, he gave birth to the passenger elevator. Otis's device, demonstrated at the Crystal Palace Exposition in New York, incorporated a clamping arrangement that gripped the guide rails on which the car moved when tension was released from the hoist rope. The first passenger elevator was put into service in the Haughwout Department Store in New York City in 1857; driven by steam power, it climbed five stories in less than a minute and was a pronounced success.

The first passenger elevator

Improved versions of the steam-driven elevator appeared in the next three decades, but no significant advance took place until the introduction of the electric motor for elevator operation in 1889. This installation, in the Demarest Building in New York City, utilized an electric motor to drive a winding drum in the building's basement. The introduction of electricity led to two further advances: in 1894 push-button controls were introduced, and in 1895 a hoisting apparatus was demonstrated in England that applied the power to the sheave (pulley) at the top of the shaft; the weights of the car and counterweight sufficed to guarantee traction. By removing the limitations imposed by the winding drum, the traction-drive mechanism made possible taller shafts and greater speeds. In 1904 a "gearless" feature was added by attaching the drive sheave directly to the armature of the electric motor, making speed virtually unlimited.

With the safety, speed, and height problems overcome, attention was turned to convenience and economy. In 1915 so-called automatic levelling was introduced in the form of automatic controls at each floor that took over when the operator shut off his manual control within a certain distance from the floor level and guided the car to a precisely positioned stop. Power control of doors was added. With increased building heights, elevator speeds increased to 1,200 feet (365 metres) per minute in such express installations as those for the upper levels of the Empire State Building (1931) and reached 1,800 feet (549 metres) per minute in the John Hancock Center, Chicago, in 1970.

Automatic operation, widely popular in hospitals and apartment buildings because of its economy, was improved by the introduction of collective operation, by which an elevator or group of elevators answered calls in sequence from top to bottom or vice versa. The basic safety feature of all elevator installations was the hoistway door interlock that required the outer (shaft) door to be closed and locked before the car could move. By 1950 automatic group-supervisory systems were in service, eliminating the need for elevator operators and starters.

An early attempt to minimize the sacrifice of floor space in elevator installations in tall buildings was the basis of the idea of the double-deck elevator, first tried in 1932. Each elevator consisted of two cars, one mounted above the other and operating as a unit, serving two floors at each stop. The technique is being increasingly adopted. Automatic double-deck elevators in the Time-Life Building, Chicago, were operating in 1971, and installations in the John Hancock Tower, Boston; the Standard Oil Company (Indiana) building, Chicago; and the Canadian Imperial Bank of Commerce, Toronto, were under construction in 1971.

Double-deck elevators

Modern elevators. Modern elevators are made in a variety of types for many purposes; in addition to ordinary freight and passenger operations they are used in ships, dams, and such specialized structures as rocket launchers. Heavy-lift, rapid-descent elevators are employed in high-rise construction operations. Practically all are electrically propelled, either by cables, sheave and counterweight, by a winding-drum mechanism (still used in many low-rise freight elevators), or by an electro-hydraulic combination. Multiple cables (three or more) increase both the traction surface with the sheave and the safety factor; cable failure is rare.

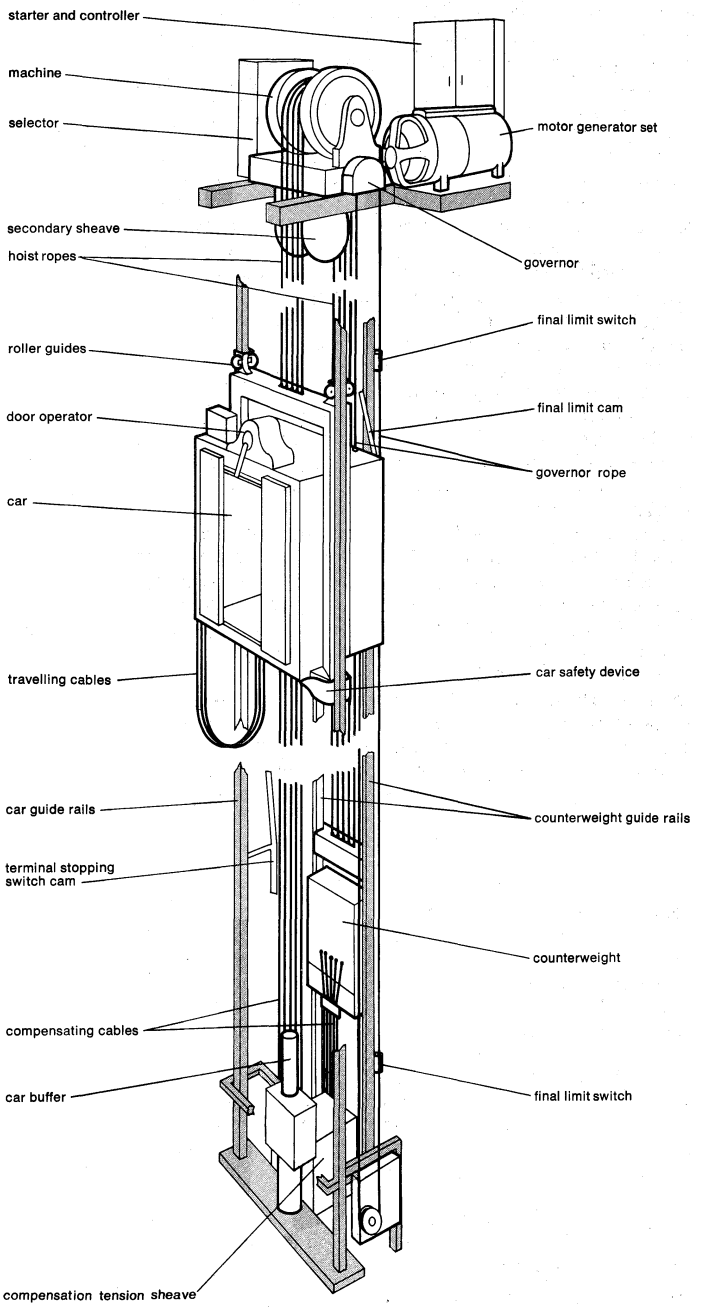
The drive motor is usually alternating current for slower speeds and direct current for higher speeds. With the direct-current motor, the speed is varied by varying the field strength of a direct-current generator, and direct connection of the armature of the generator with the armature of the drive motor. For high-speed operations, a gearless arrangement is used, usually with the cables wrapped twice around the sheave. The traction elevator may have an unlimited rise; however, rises exceeding 100 feet require compensating ropes—*i.e.*, ropes from the bottom of the car to the bottom of the counterweight; as the car rises, the compensating rope weight is transferred to the car, and as it descends, more is transferred to the counterweight, keeping the load on the drive machine nearly constant (see illustration).

Hydraulic elevators. Hydraulic cylinders and plungers are used for low-rise passenger elevators, and for heavy duty freight elevators. The plunger pushes the platform from below by the action of pressurized oil in the cylinder. A high-speed electric pump develops the pressure needed to raise the elevator; the car is lowered by the action of electrically operated valves which release the oil into a storage tank. Specialized types of hydraulic cylinder and plunger arrangements, including horizontally placed elements, are used for unusual applications. For instance, the roped, or "geared," type of hydraulic elevator common around 1900, with plunger and cylinder fitted with sheaves at each end, is employed on aircraft-carrier elevators to lift heavy loads short distances. As pressure is applied to the plunger, the distance between the sheaves increases and the ropes wrapped around the sheaves pull up the elevator.

Safety devices. Elevators lifted by hoisting ropes are required to have platform "safeties," devices designed to clamp onto the steel guide rails upon activation, quickly braking the elevator to a halt. The safety, usually mounted below the car platform, is actuated by a speed governor through a rope. The rope pulls the safety to the on position in the event of governor overspeeding and continued downward motion of the car. The device first cuts off elevator power; if overspeed continues, it applies the safety brake.

Automatic control system. Most modern elevators are automatic, using various control systems to operate elevators individually or in groups.

The earliest automatic control system, single-automatic-push-button, gives a rider exclusive use of the car for a



Electric passenger elevator for automatic service.
By courtesy of Otis Elevator Co.

trip. It is used in small apartment buildings and for freight elevators.

Collective operation is popular for use with a single elevator in a building. The car answers all calls in one direction in sequence and then reverses and answers all calls in the opposite direction. It is used in larger apartments, hospitals, and small office buildings. A variation, called two-car or duplex collective, permits two cars to operate together and share calls between them.

Group-automatic operation

Group-automatic operation controls two or more cars as a group, keeping them spaced within a specified operating interval. Group-automatic operation is used if traffic is heavy and two or more elevators are operating as in hospitals, department stores, and offices.

Doors. Separate outer doors and car doors are essential parts of modern elevator systems. The two usually employ the same type of operation—*e.g.*, centre-opening, two-leaf, single-slide. Doors are opened and closed by an electric motor on the car. Door speed in closing is regulated to avoid injury to persons caught in the closure. A sensor electrically reverses the door if it strikes an object in closing. Photoelectric controls and electronic proximity devices are also employed to control door reversal. Hoistway doors are designed so that they are always closed before the elevator can operate.

For freight elevators, vertically sliding, biparting doors are common. Such doors consist of an upper and lower leaf, mechanically linked so that the bottom half drops to floor level while the top half rises above the cab roof. A protective inner gate is often required.

Communication systems. In isolated locations, especially in private residences, a telephone to an outside exchange is often required by law. In many buildings, elevators have intercommunication systems in case of mechanical failures. Alarm buttons, emergency lighting, and emergency power are often provided.

Automatic freight handling. Automatic loading and unloading devices have been incorporated into modern freight elevators. A call button activates the automatic pickup; the elevator arrives, the load is pulled into the car, the car moves to the proper floor, and the load is discharged.

Sky lobbies. At the two World Trade Center Tower Buildings in New York City, a new technique has been introduced for minimizing the space required for elevators. Each 110-story building is divided into three layers. The lower layer, the first 41 floors of the tower, has four banks of six elevators each; similarly, the middle and upper layers each have their own four banks of six elevators. At the 44th and 78th floors there is a "sky lobby," connected to the main floor by 11 high-speed shuttle elevators.

Observation cars. Transparent walled elevators rising on the outside of buildings and observation towers have recently grown popular; the original was probably that installed in the Eiffel Tower, Paris (1889). Such observation elevators have glass-enclosed cabs affording passengers a view as they ride.

BIBLIOGRAPHY. RODNEY R. ADLER, *Vertical Transportation for Buildings* (1970), a work written primarily for architects and others who are not elevator specialists; F.A. ANNETT, *Elevators: Electric and Electrohydraulic Elevators, Escalators, Moving Sidewalks and Ramps*, 3rd ed. (1960), a fairly detailed description of equipment for electric and hydraulic elevators; J.E. BECHDOLT, *Going Up: The Story of Vertical Transportation* (1948), an illustrated history of various vintage installations; R.S. PHILLIPS, *Electric Lifts*, 5th ed. (1966), an authoritative review of modern elevator and escalator practices in terms of British application; G.R. STRAKOSCH, *Vertical Transportation: Elevators and Escalators* (1967), a comprehensive and authoritative review of elevator and escalator application principles and architectural planning requirements; H. THIEMANN (ed.), *Aufzüge: Betrieb und Wartung* (1963), a guide to German and European elevator practice encompassing application and planning requirements; R.M. VOGEL, *Elevator Systems of the Eiffel Tower, Smithsonian Institution Bulletin* 228 (1961). See also *Elevator World* (monthly), a current review of developments in all forms of vertical transportation.

(G.R.S.)

Elijah ben Solomon

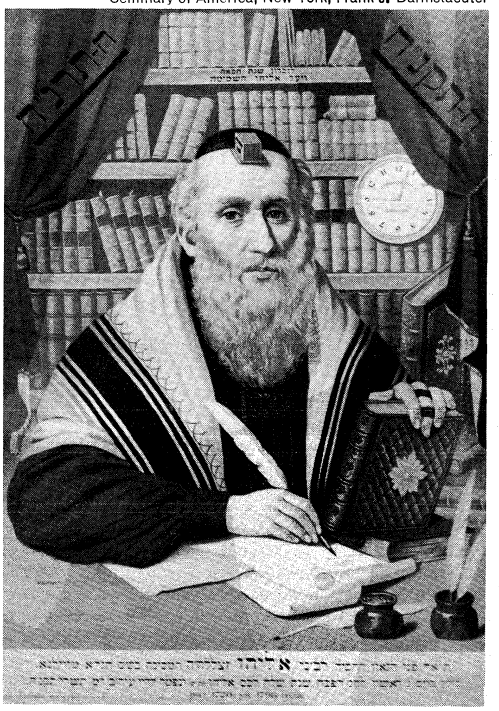
The status of Elijah ben Solomon Zalman, the Vilna *gaon* ("Excellency"), as the outstanding authority in the religious and cultural life of Jewry in Lithuania in the 18th century derived from his personal qualities of character and intellect, not from any institutionally defined leadership role. Indeed, he never held a rabbinic office or title, but his way of life as well as his achievement epitomize the role of the saintly scholar in Jewish traditional culture. His personality—the historical as well as the legendary one—and the religious and social tenets impressed by it shaped the peculiarly "Litvak" (Lithuanian–Jewish) attitude to life and culture of eastern European Jewry and later of their descendants in the West. They also influenced forms of Jewish culture that were opposed in principle to much that Elijah taught, such as Hasidism (18th-century pietistic mysticism) and Haskala (the 18th–19th-century Jewish Enlightenment).

Elijah was born April 23, 1720, at Sielec, Lithuania (now Selets, Belorussian S.S.R.), into a long line of scholars. He became famous in early childhood for

prodigious feats of rapid learning and an unusually retentive memory, intellectual gifts he retained to his last day. Married at approximately 18, he settled in Vilna (Vilnius, Wilno), leaving it only for the penitential *galut* ("exile") journey through Jewish communities in Poland, Lithuania, and Germany and returning in 1748. The rest of his life, Elijah was a central public figure in Vilna—the cultural centre of eastern European Jewry—leading the Jewish community through a combination of a life devoted to study, prayer, vision, and meditation and of teaching a chosen circle of devoted pupils. The Vilna community supported him with a stipend that came largely from funds bestowed by one of Elijah's ancestors for the support of scholars from his family. It added its own funds and also gave Elijah a rent-free apartment. Thus, the community gave him a monthly payment higher than the salary of the communal rabbi, though he had no explicit social function; it did so because of his exemplary life and voluntary leadership.

Leader of the Jewish community in Vilna

Picture from the photographic archive of the Jewish Theological Seminary of America, New York, Frank J. Darmstaedter



Elijah ben Solomon, engraving by an unknown artist.

Elijah's view of man and his struggle for self-realization throughout history was based on the assumption that for every individual there exists a peculiar path, "according to the specific roots of his soul and the particular nature of his body." In the first era of history, prophets revealed to each man his own path. In the following era, when prophecy had ended, one had to turn inward and discover one's way for oneself. Elijah held that, theoretically, this could be still done even in the present era of decline; but temptation and deception are ripe and active. Spiritual safety lies, therefore, in following the uniform Torah—the divine guidance or law.

Individual strength and guiding authority marked Elijah's spiritual life. He steadily and systematically restrained his extraordinarily quick perception and compelled himself to constant analysis and repetition, an intellectual self-discipline that he called "grinding." He rejected his own active dream life, which was full of scholarly insights and visions. Indeed, he resented the knowledge imparted to him in sleep, for he sought to worship through conscious intellectual labour; gifts from on high—even through the intermediary of heavenly angels and heralds—were seen as misleading. His many visions were even more suspect to him, for he feared evil forces that might have lurked within them.

Elijah's vast knowledge in all fields of traditional Jewish learning in Talmud and Midrash (the traditional interpretative and commentative literature) and in biblical exegesis, as well as in mystical literature and lore, was combined with a deep interest in and knowledge of philosophy, grammar, and folk medicine. He studied day and night not only as the best way to worship God but also as the true method of exercising proper leadership, which he believed should be the emanation of a scholar overflowing with learning. Elijah's metaphor for this is that of a large central vessel filling itself up to overflowing into smaller vessels—the pupils—who in their turn overflow into groups around them; thus does the central teacher influence the whole people.

Elijah put this theory in practice and drew a small circle of very able and influential pupils to him, most of whom were already rabbis or preachers. Outstanding among them were Ḥayyim ben Issac, rabbi of Volozhin, the founder of a new type of Lithuanian yeshiva (Talmudic academy) and Elijah's successor as leader of Lithuanian Jewry; and Jacob ben Wolf Kranz, an itinerant preacher of great renown, who created a new school and style of preaching.

Elijah usually avoided public activities and engaged actively in certain causes only as an extraordinary form of his unceasing leadership and teaching from his study. The most conspicuous and sustained of his public activities was his implacable fight against Ḥasidism, which lasted from 1772 until his last day. He feared and despised it as an "idolatrous," superstitious, vulgar, and antischolarly movement and participated in a manifesto excommunicating its adherents. Ḥasidic admiration for miracles and visions, adulation of their group leader (the *tzaddiq*), and striving for collective joy, with the aid of alcoholic liquor and dancing, were to him the propagation of all he rejected for himself. Moreover, Ḥasidic stress on prayer was to him an attack on the supremacy of Torah study, which he held sacred.

Opposition to Ḥasidism

Elijah died at Vilna on Oct. 9, 1797. As a result of his concentrated devotion to incessant, wakeful learning and direct, oral teaching, as well as his concern with minute textual criticism and emendation and the urgent tempo of his intellectual work, all that is extant from him are notes of his teachings kept by his pupils or short manuscript notes of his own written in the margins of basic texts. These have been combined and edited into more than 40 works now in print in all fields of his interest, yet some material still remains in manuscript. Lately, there has been ascribed to him a prophetic work, *Kol ha-tor* ("Voice of the Turtledove") foretelling the success of Zionism, but it is, in its main elements, probably a modern forgery.

Elijah's main contribution to posterity was to provide an example of the intellectual love of God in the traditional mode at the threshold of Jewish modernity.

BIBLIOGRAPHY. S. SCHECHTER, "Rabbi Elijah Wilna, Gaon," in *Studies in Judaism*, First Series (1911), L. GINZBERG, "The Gaon, Rabbi Elijah Wilna," in *Students, Scholars, and Saints* (1928, reprinted 1958), the best accounts available in English; H.H. BEN-SASSON, "The Personality of Elijah, Gaon of Wilna, and His Historical Influence," in *Zion*, 31:39–86, 197–216 (1966), an attempt to understand the man through an analysis of his works (in Hebrew, with English summaries).

(H.H.B.-S.)

Elimination

Biologically, elimination may be considered to encompass all of the various mechanisms and processes by which life forms dispose of or throw off waste products, toxic substances, and dead portions of the organism. In a more broad context, elimination may be thought to include all waste-disposal processes in nature, including those of entire animal and plant populations as well as of individuals.

GENERAL FEATURES

Definitions. Four terms are commonly associated with waste-disposal processes and are often used interchangeably, though not always correctly, viz., excretion, secretion, egestion, and elimination.

Waste disposal processes

Excretion. Excretion is a general term referring to the separation and throwing off of waste materials or toxic substances from the cells and tissues of a plant or animal.

Secretion. The separation, elaboration, and elimination of certain products arising from cellular functions in multicellular organisms is called secretion. Though these substances may be a waste product of the cell producing them, they are frequently useful to other cells of the organism. Examples of secretions are the digestive enzymes produced by intestinal and pancreatic tissue cells of vertebrate animals, the hormones synthesized by specialized glandular cells of plants and animals, and sweat secreted by glandular cells in the skins of some mammals. Secretion implies that the chemical compounds being secreted were synthesized by specialized cells and that they are of functional value to the organism. The disposal of common waste products should not, therefore, be considered to be of a secretory nature.

Egestion. Egestion is the act of excreting unusable or undigested material from a cell, as in the case of single-celled organisms, or from the digestive tract of multicellular animals.

Elimination. As defined above, elimination broadly defines the mechanisms of waste disposal by living systems at all levels of complexity. The term may be used interchangeably with excretion.

Biological significance of elimination. Waste disposal by single-celled and multicellular organisms is vital to their well-being and to the continuance of life. Animals must take in (ingest) energy-containing chemical compounds, extract a portion of the energy to power their life processes, and dispose of the unusable material or by-products formed during the energy-extraction process. An analogous series of events occurs in an internal combustion engine. Fuel, containing energy, is taken into the engine, where it is burned, and a portion of the energy released is used to move the pistons. As in living cells, a portion of the energy-containing material (fuel) not utilized in the engine is exhausted in the form of carbon monoxide, carbon dioxide, and other by-products of combustion. Blockage of the exhaust system in an engine results in loss of efficiency and eventual total breakdown. Similarly, the rate of waste disposal in biological systems can and does provide a means of controlling the metabolic rate. Complete blockage of waste-disposal mechanisms in living systems is as effective in destroying vital functions as the cutting off of food, oxygen, or water from the system. In addition, some substances produced as metabolic by-products are toxic in themselves and must be removed from living cells at a rate equal to that at which they are produced by those cells. Thus, the excretion of waste products from living cells must occur continually in order to ensure the normal progression of vital chemical events.

Importance of elimination in living systems

Waste and poisonous substances produced by the metabolic activities of plant and animal communities must, in a similar manner, be removed or detoxified if community well-being is to be preserved. Collective wastes of individual organisms composing a community, if allowed to accumulate to any marked degree, will eventually destroy the lives of all the community members.

The biosphere, composed of all individuals and communities of life forms and their environments on the Earth, is equally sensitive to the effects of waste and poison accumulation. A continual buildup of substances harmful to life forms can only result in the eventual destruction of most or all of the presently existing species of plants and animals.

Types of wastes. Waste products may be categorized as metabolic or nonmetabolic. The difference lies in whether the substances in question are produced by the chemical processes of a living cell or are merely passed through the digestive tract of an organism without actually entering into its life processes.

The nonmetabolic wastes are mainly materials that, by virtue of their chemical makeup, are indigestible or unusable by an organism. In addition, nonmetabolic wastes include any substances that are absorbed, ingested, or otherwise taken into a living system in excess of the needs and storage capabilities of the organism. These substances include digestible (metabolizable) as well as indigestible materials, and they may be excreted almost immediately, even though they are often usable as food.

Metabolic and nonmetabolic wastes

Metabolic wastes may be separated into gases, liquids, solids, and heat. Heat, though usually not classified as a waste product, should be classified as such because it is a by-product of metabolic activity and must be eliminated to avoid harmful elevation of body temperatures in warm-blooded animals.

Gaseous wastes. Oxygen produced during photosynthetic reactions in green plants and certain bacteria may be considered to be a waste, or at least a by-product, requiring removal. Carbon dioxide is produced by all animals and by green plants in darkness. Nitrogen gas is produced by denitrifying sulfur bacteria (*Thiobacillus*), and ammonia is excreted by decay-causing bacteria and by most invertebrate and vertebrate animals.

Liquid wastes. The sole liquid waste produced as a metabolic by-product by all animals and photosynthetic plants in darkness is water.

Solid wastes. Several important kinds of materials may be classified as solid wastes. Among them are nitrogenous wastes, by-products of protein and amino acid metabolism by animals; nitrite and nitrate compounds produced by nitrifying bacteria; and sulfur and sulfates resulting from the metabolic activities of sulfur bacteria. Numerous other substances also enter into solid wastes to be disposed of by organisms. Iron compounds in insoluble form are secreted by iron bacteria that have utilized soluble iron compounds. Various resins, fats, waxes, and complex organic chemicals are exuded from certain plants—as in the latex from rubber trees and milkweeds. Organic pigments from the breakdown of biological pigments, such as hemoglobin in vertebrates, become components of solid waste. Inorganic salts, including molecules and ions such as carbonates, bicarbonates, and phosphates resulting from life sustaining chemical reactions, eventually may become solid waste products.

METHODS OF WASTE DISPOSAL

Disposal of metabolic and nonmetabolic wastes involves both active and passive mechanisms. In general, gaseous wastes are eliminated principally without the direct expenditure of energy on the part of the living system. The solid and liquid waste-disposal mechanisms utilized by higher animals are active (energy consuming) systems that act to separate waste materials from vital substances prior to excretion. Methods of disposal may be classified into specific and nonspecific systems.

Specific elimination mechanisms. Three pathways exist in this context: (1) the alimentary canal, (2) the respiratory system, and (3) the kidney.

Alimentary canal. The alimentary canal is a pathway utilized almost exclusively for the elimination of solid wastes of an indigestible nature, and the act of elimination by this means is termed egestion. Materials disposed of in this manner cannot be said to have entered the tissues of the animal but are the residue of enzymatic and absorptive activities occurring in the digestive tract. True metabolic wastes are excreted by means of the flow of bile from the liver into the intestine. The destruction of cells, in animals, produces bile pigments—residues of hemoglobin and other pigments—which may be considered to be the principal metabolic wastes eliminated via the alimentary canal. Waste disposal in this manner requires little energy expenditure other than that employed in the peristaltic contractions of muscle in the walls of the tract that act to push material along the length of the tube.

Respiratory system. This pathway is concerned principally with the gaseous waste products of metabolism (carbon dioxide and ammonia), which move to the external environment by means of diffusion from the cells of origin. In animals, transport is by means of the circulatory system when present or simply by diffusion through the cell membranes of lower animals. A few multicellular, aquatic animals lose carbon dioxide to the surrounding water by way of diffusion through the thin

vascular membranes of their general body surface. In most higher animals, however, the skin is too hard or thick and nonvascular to function effectively in gas disposal. In these animals, gills and lungs—aggregations of thin, moist, vascular membranes—have evolved. Membranes of the gills of aquatic animals and the lungs of terrestrial forms are provided with large surface areas for the diffusion of waste gases from the circulatory system to the outside environment. Because carbon dioxide is soluble in the body water, it can easily diffuse into the circulatory system, in solution, from the cells of origin.

Carbon dioxide elimination

Transport and excretion of carbon dioxide requires little energy as it diffuses along concentration gradients from cells to the circulation and finally to the outside environment.

Because more carbon dioxide (CO_2) is produced by metabolic activity than can be carried in the circulation system in the form of dissolved CO_2, the major portion is transported to the gills and lungs as bicarbonate (HCO_3^-), via two chemical reactions:

$$CO_2 + H_2O \leftrightharpoons H_2CO_3 \rightleftarrows H^+ + HCO_3^-.$$

Thus, carbon dioxide reacts with water, producing carbonic acid, which in turn dissociates to produce a hydrogen ion and a bicarbonate ion. In the lungs or gills, the reverse of these reactions occurs, and carbon dioxide diffuses out of the body into the outside environment. Certain aquatic animals are capable of eliminating gaseous ammonia—derived from protein breakdown—by way of specialized cells in their gill tissues.

Salt secretion by way of specialized cells in the gills occurs in marine vertebrates whose bodies constantly absorb salt through thin membranes of their oral, respiratory, and body surfaces.

The kidney. Kidneys have evolved in multicellular animals as a highly sophisticated channel for waste disposal, and they function to regulate the levels of water, salts, and organic materials in the bodies of higher animals. Materials eliminated via the kidney include nitrogenous waste products (ammonia, uric acid, urea, creatine, creatinine, and amino acids), excess quantities of salts and water that may be taken into the body, and various other organic materials produced by life-sustaining chemical reactions. Functionally, the kidney is a microfilter that initially removes dissolved as well as some suspended materials from the circulation system, along with large quantities of water. These substances are differentially reabsorbed, before being excreted as urine, to a degree that varies considerably throughout the animal kingdom. For example, animals that absorb large quantities of water into their bodies (such as freshwater fishes) excrete copious quantities of water in their urine. The reverse is true of many desert animals, which must conserve water and therefore produce a thick, semisolid urine. The kidney, in its various stages of evolution, functions at the expense of considerable metabolic energy and cannot be considered to be a passive system.

Nonspecific mechanisms of waste disposal. A multitude of disposal mechanisms exist throughout the plant and animal kingdoms for the disposal of excess plant and animal material. Among plants, the shedding and dropping of bark, leaves, and twigs might, in a broad sense, be said to represent disposal mechanisms. Certain plants, in addition, secrete or exude resins, sap, and other substances that accumulate in excessive quantities within the plant.

Elimination of particulate wastes from blood stream

Specialized, mobile, amoeba-like cells exist in the blood and tissues of animals and engulf particulate wastes resulting from the disintegration of dead cells or the intake of foreign particles into the bodies of animals. Waste matter thus stored inside these small cells is removed from contact with the organism or its metabolism and may be considered to be eliminated whether or not the material is ever actually eliminated from the body of the organism during its normal life cycle.

Toxic substances are produced by normal metabolic activities. Though some of these poisons are eliminated in their original chemical form, others, such as some nitrogenous compounds, are altered biochemically to less toxic compounds. In this manner, more of the original waste may be safely stored, or permitted to accumulate without harmful effects to the organism, until they can be eliminated. In addition, toxic chemicals that are inadvertently ingested or produced by bacterial action (infection) are frequently converted to nontoxic forms by enzymatic and antibody (immune) reactions. Such materials can then safely be eliminated with other wastes along normal pathways of excretion.

Heat is eliminated from the bodies of animals by conduction to the external surface of the organism. In animals possessing a circulatory system, heat travels in the fluid of the circulation system from the deeper portions of the body to the surface. At the body surface, heat is lost by physical processes of convection, radiation, conduction, and evaporation of sweat.

COMPARATIVE OVERVIEW OF ELIMINATION SCHEMES

Protista. No specialized elimination mechanisms are present in algae, fungi, protozoans, and slime molds, the main groups of protists. Metabolic wastes (carbon dioxide, water, oxygen, and nitrogenous compounds) diffuse through the cell membranes of these single-celled organisms into the outside environment. Particulate wastes pass from the bodies of certain protozoans to the exterior by way of small openings in the body surface—anal pores and other cell openings.

Elimination in protists is carried out passively and therefore requires little or no expenditure of metabolic energy on the part of the organism.

Plants. Plants are not generally considered to possess special mechanisms of elimination. Photosynthetic activities of green plants, in the presence of light, produce oxygen, which diffuses out through openings in the leaves (stomata) or through the cell walls of roots and other plant structures. Excess water passes to the exterior via similar routes and is eliminated by processes of guttation (droplet exudation) and transpiration (evaporation of water from plant surfaces).

Green plants in darkness or plants that do not contain chlorophyll produce carbon dioxide and water as respiratory waste products. Carbon dioxide is secreted in the same manner as oxygen via diffusion through stomata and cell walls. Materials that are exuded by some plants —resins, saps, latexes, etc.—are forced from the interior of the plant by hydrostatic pressures inside the plant and by absorptive forces of plant cells. These forces are passive in nature, and exudation requires no energy expenditure on the part of the plant.

Animals. Diverse mechanisms have evolved that enable the various animal species to inhabit a wide range of environments. In animals whose bodies consist of a single layer of cells, waste disposal is accomplished principally by diffusion from the site of waste production to the outside environment. This method is very efficient when the distances over which wastes diffuse are relatively short, when there is a high surface area to volume relationship, and when the rate of waste production is relatively low. In more complex animals, however, waste elimination by diffusion through the body wall to the exterior is less efficient because individual cells are further removed from the exterior surface of the organism. The presence of specialized mechanisms of elimination in higher animals enables wastes to be rapidly transported to the exterior surface of the body (see EXCRETION AND EXCRETORY SYSTEMS).

Efficiency of the diffusion mechanism of waste disposal

Sponges. Phylogenetically, the sponges (phylum Porifera) are the simplest of animals, being multicellular and composed of specialized cells in a single layer for the maintenance of life processes. Elimination in these aquatic animals proceeds by diffusion of gaseous wastes into the surrounding water and by the ejection of solid wastes and indigestible material from the digestive cells into the streams of water that constantly flow through the animal.

Coelenterates. The jellyfishes, coral animals, ctenophores, and comb jellies possess a rudimentary canal-like cavity in their two-layered bodies for the ingestion, digestion, and egestion of food and wastes. Gaseous wastes are

eliminated by diffusion, and solid wastes in dissolved or undissolved form pass out through an opening in the body wall that serves the dual purposes of food intake and waste elimination.

Flatworms. Flatworm bodies consist of three layers of cells, and in this aquatic group elimination is similar to that of the less complex animals. Food and solid wastes enter and leave through a common opening in the well-developed digestive tract, which consists of a mouth, pharynx, and gastrovascular cavity.

Nemertine worms. The digestive and excretory systems of the aquatic proboscis worms is more efficient than that of lower animals in that a well-defined mouth, intestine, and excretory opening (anus) permit the one-way flow of food and waste through the animal. Egested food and nitrogenous wastes, which are secreted into the intestine, are passed along it to the anus by peristaltic waves of the smooth muscle lining the intestinal walls. The efficiency of waste elimination is increased by the presence of a well-defined circulatory system, which enhances the carriage of wastes to the intestine.

Nematodes. An additional excretory structure has evolved in the roundworms. Excretory canals located on both sides of the intestine facilitate waste disposal by carriage of material to an excretory pore in the body wall.

Other invertebrates. Increasing structural complexity is accompanied, in invertebrates, by more efficient waste-disposal mechanisms. In the phylum Mollusca (clams, snails, oysters, mollusks, octopuses, and squids), gills are present, adding another more efficient channel for waste disposal. A heart increases the rate of flow in the circulatory system and speeds the transport of wastes to the gills. An excretory, kidney-like organ acts to remove metabolic wastes from the circulation and body fluid prior to excretion. All basic mechanisms of excretion are thus present in relatively simple animals. As invertebrates become more specialized and complex, as in the arthropods (insects, crabs and other joint-legged animals) and annelids (segmented worms), adaptations exist whereby excretion is modified for purposes of survival in nonaquatic environments.

Vertebrates. Though the wastes produced by vertebrates differ little qualitatively from those of higher invertebrates, increased structural complexity and body size, in combination with environmental adaptations, require more specific waste-disposal mechanisms in order to maintain a constant internal environment. The presence of highly efficient, water-retaining kidneys, for example, permits vertebrates to inhabit arid, hot regions of the Earth. It seems proper, within the vertebrate group, to consider elimination schemes as variations of mechanisms common to all higher animals but which enable animals to inhabit widely diversified environments.

PROBLEMS OF WASTE DISPOSAL IN THE BIOSPHERE

Environmental pollution

Man is unique among living things in that his activities result in the production of waste materials (pollutants) that, by virtue of their chemical structure, are poisonous to all living things, including himself (see BIOSPHERE; CONSERVATION OF NATURAL RESOURCES; and POLLUTION CONTROL). While the metabolic and nonmetabolic wastes produced by most living things are used in the maintenance of various ecological cycles of the biosphere, substances such as insecticides, waste metals, strong acids and alkalies, atmospheric poisons, etc., act to unbalance and interrupt natural cycles. The addition of such materials to the biosphere poses a serious threat to all life on Earth.

Complex chemical cycles in the natural world act to utilize the waste products of all living systems. Biological wastes of one life group become the energy sources of other groups. Disposal of water-borne wastes on a biosphere level involves such processes as the filtering of water through fine soil, the settling out of particulate matter in streams and lakes, and the destruction of wastes by lower life forms (bacteria and fungi). Carbon dioxide, produced by the oxidation of organic molecules, diffuses into the atmosphere, where it may dissolve in the surface waters of the world or be utilized by green plants during photosynthesis. The nitrogenous wastes of animals are utilized directly by bacteria and are eventually recycled by way of plants to become the proteins and amino acids consumed by animals. Thus, all living things, on a broad scale, are dependent on one another and on natural biogeochemical cycles.

BIBLIOGRAPHY. P.B. WEISZ, *The Science of Biology*, 3rd ed. (1967), a comprehensive general text emphasizing molecular biology, ecology, and morphology; R.M. BUCHSBAUM, *Animals Without Backbones*, 2nd ed. (1948), an elementary, profusely illustrated account of invertebrate animals; A.S. ROMER, *The Vertebrate Body*, 4th ed. (1970), a general history of the vertebrate body emphasizing comparative aspects of structure and function; C.L. PROSSER and F.A. BROWN, *Comparative Animal Physiology*, 2nd ed. (1961), a college level text on the comparative aspects of functional systems in animals; *Scientific American*, vol. 223 (1970), an issue devoted to the biosphere and cyclic activities of life on Earth.

(F.C.Ke.)

Eliot, George

One of the great English novelists of the Victorian Age, George Eliot (the pen name of Mary Anne Evans) developed the method of psychological analysis characteristic of modern fiction. She was born at Chilvers Coton, Warwickshire, November 22, 1819, on an estate, for the owner of which her father was agent. Her command of England's Midland dialects is not strange; she heard little else until she went as a boarder to Mrs. Wallington's School at Nuneaton (1828–32). There she came under the influence of Maria Lewis, the principal governess, who, along with a formal Johnsonian diction, inculcated a strong evangelical piety in the young girl. At her last school (1832–35), conducted by the daughters of the Baptist minister at Coventry, her religious ardour increased. She dressed severely and engaged earnestly in good works. The school gave her a reading knowledge of French and Italian, and after her mother's death had compelled her to return home to keep house for her father, he let her have lessons in Latin and German. In 1841, when her father turned his business over to his son Isaac, she moved with him to Coventry.

Education

By courtesy of the National Portrait Gallery, London

George Eliot, chalk drawing by F.W. Burton, 1865. In the National Portrait Gallery, London.

There she became acquainted with a prosperous ribbon manufacturer, Charles Bray (1811–84), a self-taught freethinker who campaigned boldly for secular education, trade unions, extension of the franchise, and other radical causes. Mrs. Bray came of an earnest Unitarian family; her sister, Sara Hennell, was immersed in theological and philosophical speculations, and her brother, Charles Hennell, was the author of *An Inquiry Concerning the Origin of Christianity* (1838), a book that precipitated Miss Evans' break from orthodoxy that had been

long in preparation. Her favourite Scott's novels had early shown her that papists might have noble characters, and she knew many Dissenters who lived better lives than many churchmen. Various books on the relation between the Bible and science had instilled in her keen mind the very doubts they were written to dispel. In 1842 she told her father that she could no longer go to church. The ensuing storm raged for several months before they reached a compromise, leaving her free to think what she pleased so long as she appeared respectably at church, and she lived with him until his death in 1849.

Early literary endeavours

The Brays and the Hennells quickly drew her from extreme provincialism, introducing her to many ideas in violent antithesis with her Tory father's religious and political views. When Charles Hennell married in 1843, she took over from his wife the translating of D.F. Strauss's *Das Leben Jesu kritisch bearbeitet*, which, published anonymously as *The Life of Jesus Critically Examined* (3 vol., 1846), had a profound influence on English rationalism. After the wedding Mrs. Hennell's father, R.H. Brabant, invited Miss Evans to visit at Devizes. A rather silly man, he had worked for years on a book (never completed) that was to dispose forever of the supernatural elements in religion. They read German and Greek together and discussed theology on long walks; soon Mrs. Brabant became jealous of their intimacy, and, before the term of her visit, Mary Anne was forced to leave. Mrs. Hennell felt that her father had acted ungenerously. Out of the humiliation of this episode George Eliot drew the horrible vividness of Mr. Casaubon in *Middlemarch*.

She spent the winter of 1849–50 at Geneva, reading widely, living in the family of François d'Albert Durade, who painted a portrait of her. Like those by Mrs. Bray (1842) and Sir Frederic Burton (1865), all in the National Portrait Gallery, it shows her with light brown hair, gray-blue eyes, and a very fair complexion. Returning to Coventry, she spent the rest of 1850 with the Brays, pondering how to live on the £100 a year left by her father. After John Chapman, the publisher of *The Life of Jesus Critically Examined*, got her a chance to review R.W. Mackay's *The Progress of the Intellect* in *The Westminster Review* (January 1851), she decided to settle in London as a free-lance writer, and in January 1851 she went to board with the Chapmans at 142, Strand.

Soon after her arrival Mrs. Chapman and the children's governess, who was also Chapman's mistress, became jealous of Marian, as she now signed her name, and after ten weeks she returned to Coventry in tears. Doubtless her feelings were strongly attracted to the magnetic Chapman, whose diary supplies these details, but there is no evidence that she was ever his mistress. A few months later he bought *The Westminster Review*, and Marian, contrite at the domestic complications she had unwittingly caused, returned to London. For three years, until 1854, she served as subeditor of *The Westminster*, which under her influence enjoyed its most brilliant run since the days of John Stuart Mill. At the Chapmans' evening parties she met many notable literary figures in an atmosphere of political and religious radicalism. Across the Strand lived the subeditor of *The Economist*, Herbert Spencer, whose *Social Statics* (1851) Chapman had just published. Marian shared many of Spencer's interests and saw so much of him that it was soon rumoured that they were engaged. Though he did not become her husband, he introduced her to the two men who did.

Relationship with George Henry Lewes

George Henry Lewes was the most versatile of Victorian journalists. In 1841 he had married Agnes Jervis, by whom he had four sons. In 1850 Lewes and a friend, the journalist Thornton Leigh Hunt, founded a radical weekly called *The Leader*, for which he wrote the literary and theatrical sections. In April 1850, two weeks after the first number appeared, Agnes Lewes gave birth to a son whose father was Thornton Hunt. Lewes, being a man of liberal views, had the child registered as Edmund Lewes and remained on friendly terms with his wife and Hunt. But after she bore Hunt a second child in October 1851, Lewes ceased to regard her as his wife, though, having condoned the adultery, he was precluded from suing for divorce. At this moment of dejection, his home hopelessly broken, he met Marian Evans. They consulted about articles and went to plays and operas that Lewes reviewed for *The Leader*. Convinced that his break with Agnes was irrevocable, Marian determined to live openly with Lewes as his wife. In July 1854, after the publication of her translation of Feuerbach's *Essence of Christianity*, they went to Germany together. In all but the legal form it was a marriage, and it continued happily until Lewes' death in 1878. "Women who are content with light and easily broken ties," she told Mrs. Bray, "do *not* act as I have done. They obtain what they desire and are still invited to dinner."

At Weimar and Berlin she wrote some of her best essays for *The Westminster* and translated Spinoza's *Ethics* (still unpublished), while Lewes worked on his pioneer life of Goethe. By his pen alone he had to support his three surviving sons at school in Switzerland as well as Agnes, whom he gave £100 a year, which was continued until her death in 1902. She had four children by Hunt, the last born in 1857, all registered under Lewes' name. The few friends who knew the facts agreed that toward Agnes his conduct was more than generous, but there was a good deal of malicious gossip about the "strong-minded woman" who had "run off with" her husband. Marian's deepest regret was that her act isolated her from her family in Warwickshire. She turned to early memories and, encouraged by Lewes, wrote a story about an episode in Chilvers Coton parish during her childhood. Published in *Blackwood's Magazine* (1857) as "The Sad Fortunes of the Reverend Amos Barton," it made an instant success. Two more tales, "Mr. Gilfil's Love-Story" and "Janet's Repentance," also based on local events, appeared serially in the same year, and Blackwood republished all three as *Scenes of Clerical Life* (2 vol., 1858) under the pseudonym George Eliot.

Adam Bede (3 vol., 1859), her first long novel, she described as "a country story—full of the breath of cows and the scent of hay." Its masterly realism—"the faithful representing of commonplace things"—brought to English fiction the same truthful observation of minute detail that Ruskin was commending in the Pre-Raphaelites. The book is rich in humour. The germ of the plot was an anecdote her Methodist aunt told of visiting a girl condemned for child murder. The dialect of the Bedes she had heard her Derbyshire uncles speaking with her father, some of whose early experiences she assigned to Adam. But what was new in English fiction was the combination of deep human sympathy and rigorous moral judgment. *Adam Bede* went through eight printings within a year, and Blackwood doubled the £800 paid for it and returned the copyright.

Adam Bede

In *The Mill on the Floss* (3 vol., 1860) she returned again to the scenes of her early life. The first half of the book, with its remarkable portrayal of childhood, is irresistibly appealing, and throughout there are scenes that reach a new level of psychological subtlety.

At this time historical novels were in vogue, and when Lewes during their visit to Florence in 1860 suggested Savonarola as a good subject, George Eliot grasped it enthusiastically and began to plan *Romola* (1862–63). First, however, she wrote *Silas Marner* (1861), which had thrust itself between her and the Italian material. Its brevity and perfection of form made this story of the weaver whose lost gold is replaced by a strayed child the best known of her books, though it has suffered unfairly from being forced on generations of school children. *Romola* was planned as a serial for *Blackwood's*, until an offer of £10,000 from *The Cornhill Magazine* induced George Eliot to desert her old publisher; but rather than divide the book into the 16 installments the editor wanted, she accepted £3,000 less, an evidence of artistic integrity few writers would have shown. Details of Florentine history, setting, costume, and dialogue were scrupulously studied at the British

Museum and during a second trip to Italy in 1861. It was published in 14 parts between July 1862 and August 1863. Though the book lacks the spontaneity of the English stories, it has been unduly disparaged.

George Eliot's next two novels are laid in England at the time of agitation for passage of the Reform Bill. In *Felix Holt, the Radical* (3 vol., 1866) she drew the election riot from recollection of one she saw at Nuneaton in December 1832. The initial impulse of the book was not the political theme but the tragic character of Mrs. Transome, one of her greatest triumphs. The intricate plot popular taste then demanded now tells against the novel. *Middlemarch* (8 parts, 1871–72) is by general consent George Eliot's masterpiece. Under her hand the novel had developed from a mere entertainment into a highly intellectual form of art. Every class of Middlemarch society is depicted from the landed gentry and clergy to the manufacturers and professional men, the shopkeepers, publicans, farmers, and labourers. Several strands of plot are interwoven to reinforce each other by contrast and parallel. Yet the story depends not on close-knit intrigue but on showing the incalculably diffusive effect of the unhistoric acts of those who "lived faithfully a hidden life and rest in unvisited tombs."

Daniel Deronda (8 parts, 1876), in which George Eliot comes nearest the contemporary scene, is built on the contrast between Mirah Cohen, a poor Jewish girl, and the upper class Gwendolen Harleth, who marries for money and regrets it. The less convincingly realized hero, Daniel, after discovering that he is Jewish, marries Mirah and departs for Palestine to establish a home for his nation. The picture of the Cohen family evoked grateful praise from Jewish readers. But the best part of *Daniel Deronda* is the keen analysis of Gwendolen's character, which seems to many critics the peak of George Eliot's achievement.

In 1863 the Leweses bought the Priory, 21, North Bank, Regent's Park, where their Sunday afternoons became a brilliant feature of Victorian life. There on November 30, 1878, Lewes died. For nearly 25 years he had fostered her genius and managed all the practical details of life, which now fell upon her. Most of all she missed the encouragement that alone made it possible for her to write. For months she saw no one but his son Charles Lee Lewes; she devoted herself to completing the last volume of his *Problems of Life and Mind* (1873–79) and founded the George Henry Lewes Studentship in Physiology at Cambridge. For some years her investments had been in the hands of John Walter Cross (1840–1924), a banker introduced to the Leweses by Herbert Spencer. Cross's mother had died a week after Lewes. Drawn by sympathy and the need for advice, George Eliot soon began to lean on him for affection too. On May 6, 1880, they were married in St. George's, Hanover Square. Cross was 40; she was in her 61st year. After a wedding trip in Italy they returned to her country house at Witley before moving to 4, Cheyne Walk, Chelsea, where on December 22, 1880, George Eliot died. She was buried at Highgate Cemetery.

MAJOR WORKS

NOVELS: *Scenes of Clerical Life*, 2 vol. (1858; "The Sad Fortunes of the Reverend Amos Barton," "Mr. Gilfil's Love-Story," and "Janet's Repentance," all first published in *Blackwood's Magazine* in 1857); *Adam Bede*, 3 vol. (1859); *The Mill on the Floss*, 3 vol. (1860); *Silas Marner: The Weaver of Raveloe* (1861); *Romola*, 3 vol. (1863; first published in *Cornhill Magazine*, 1862–63); *Felix Holt, the Radical*, 3 vol. (1866); *Middlemarch: A Study of Provincial Life*, 4 vol. (1871–72); *Daniel Deronda*, 4 vol. (1876).

OTHER WORKS: *The Spanish Gypsy* (1868), drama in blank verse; *The Legend of Jubal and Other Poems* (1874), which ends with, "Oh, May I Join the Choir Invisible"; *Impressions of Theophrastus Such* (1879), a collection of moral essays; *Essays and Leaves from a Note-Book*, ed. by C.L. Lewes (1884), mostly reprinted from journals.

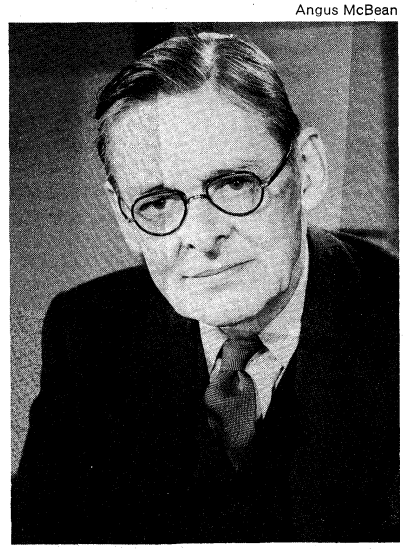
TRANSLATIONS: D.F. Strauss, *The Life of Jesus Critically Examined*, 3 vol. (1846; trans. from the German begun by Mrs. Charles Hennell and completed anonymously by George Eliot); L. Feuerbach, *The Essence of Christianity*, trans. from the German (1854); and B. Spinoza, *Ethics* (unpublished).

BIBLIOGRAPHY. GORDON S. HAIGHT, *George Eliot: A Biography* (1968), the definitive life, includes materials omitted by J.W. CROSS in *George Eliot's Life As Related in Her Letters and Journals*, 3 vol. (1885), rev. ed. with additions, 1 vol. (1887). *The George Eliot Letters*, ed. by GORDON S. HAIGHT, 7 vol. (1954–55), is a basic source of information; Haight's *George Eliot and John Chapman, with Chapman's Diaries* (1940), gives valuable details of her early years in London; and *The Essays of George Eliot*, ed. by THOMAS PINNEY (1963), collects her most important journalistic writings. JOAN BENNETT, *George Eliot: Her Mind and Art* (1948), has sound criticism of the novels. F.R. LEAVIS in *The Great Tradition* (1948; new ed., 1962), ranking George Eliot among the four great English novelists, proclaims vigorous opinions, some of which he later revised. One of the subtlest general studies is BARBARA HARDY, *The Novels of George Eliot: A Study in Form* (1959; reprinted with corrections, 1967). JEROME THALE, *The Novels of George Eliot* (1959); and BERNARD J. PARIS, *Experiments in Life: George Eliot's Quest for Values* (1965), both concentrate on the moral and philosophical ideas in her work. U.C. KNOEPFLMACHER, *George Eliot's Early Novels: The Limits of Realism* (1968), studies her compromise in the search for form between her vision of the real world and the ideal world of her moral beliefs. HENRY AUSTER, *Local Habitations: Regionalism in the Early Novels of George Eliot* (1970), less limited than its title suggests, offers much perceptive comment and includes a well-selected bibliography; MARGHANITA LASKI, *George Eliot and Her World* (1973), is richly illustrated. (G.S.Ha.)

Eliot, T.S.

As poet, dramatist, literary critic, and editor, T.S. Eliot exercised a strong influence on 20th-century culture in the period between World Wars I and II; his experiments in diction, style, and versification revitalized English poetry, and in a series of critical essays he shattered old orthodoxies and erected new. He pronounced on social and cultural questions, and as director of the publishers Faber & Faber Ltd. he was a generous and discriminating patron of young poets. The publication of *Four Quartets*, during World War II, led to his recognition as the greatest living English poet and man of letters with the award in 1948 of the Order of Merit and the Nobel Prize for Literature.

Angus McBean



T.S. Eliot, 1954.

Family background and early years. Thomas Stearns Eliot was born in St. Louis, Missouri, on September 26, 1888. His first American paternal ancestor, Andrew Eliot, a shoemaker of East Coker, Somerset, arrived in Boston in 1670. The poet's grandfather, the Rev. William Greenleaf Eliot, migrated from Boston to St. Louis immediately after his graduation from Harvard in 1834. A fervent Unionist, Unitarian, and anti-slavery man in a slave state, he was a founder of Washington University, which would have been named for him but for his objection; he became its chancellor in 1872. The poet's father, Henry Ware Eliot, broke with family tradition

and went into business; his mother, Charlotte Champe Stearns, was a prolific poet of some reputation, author of a life of her father-in-law and of a poetic drama on the martyrdom of the 15th-century Italian religious reformer Savonarola, to which her son wrote an introduction in 1926. At the time of the poet's birth, the Eliots had been in Missouri 54 years; yet they retained, against Southern and Western influences, the New England political and theological culture. Late in life the poet said that he was glad to have been born in St. Louis instead of Boston or New York or London. He was, nevertheless, a New Englander to the end.

Early influences

Eliot's family background is important for an understanding of his career, for he was allowed the widest education available in his time, with no influence from his father to be "practical" and to go into business. From Smith Academy in St. Louis he went to Milton, in Massachusetts; from Milton he entered Harvard in 1906; he was graduated B.A. in 1909, after three instead of the usual four years. The men who influenced him at Harvard were George Santayana, the philosopher and poet, and the critic Irving Babbitt. From Babbitt he derived an anti-Romantic attitude that, amplified by his later reading of British philosophers F.H. Bradley and T.E. Hulme, lasted through his life. In the academic year 1909–10 he was an assistant in philosophy at Harvard.

The year 1910–11 was crucial for the man who was to become one of two or three great poets of the first half of the 20th century. He spent the year in France, attending Henri Bergson's lectures in philosophy at the Sorbonne and reading poetry with Alain-Fournier, whose "lessons" gave Eliot a mastery not only of the French language but of Symbolist poetry from Charles Baudelaire through Jules Laforgue to Stéphane Mallarmé. In 1914 he first met Ezra Pound. In 1946 he was to publish in *Poetry: A Magazine of Verse* a defense of Pound's literary career (ignoring Pound's politics) in which he described the arid literary milieu of young American poets in the decade before World War I. Both Eliot and Pound turned to poets in other languages—Pound to the troubadours and the early Renaissance poets of Italy, Eliot to Laforgue and Dante, who even more than John Webster and John Donne helped Eliot to find his own style. From 1911 to 1914 he was back at Harvard reading Indian philosophy and studying Sanskrit under the distinguished scholar Charles Lanman. In 1913 he read Bradley's *Appearance and Reality;* by 1916 he had finished, in Europe, a dissertation entitled *Knowledge and Experience in the Philosophy of F.H. Bradley* (not published until 1964); but World War I had intervened, and he never returned to Harvard to take the final oral examination for the Ph.D. degree.

Early publications. Eliot was to pursue four careers: editor, dramatist, literary critic, and philosophical poet. He was probably the most erudite poet of his time in the English language. His undergraduate poems were "literary" and conventional. His first important publication, and the first masterpiece of "modernism" in English, was "The Love Song of J. Alfred Prufrock." Nothing like the first three lines of "Prufrock" had previously appeared in English poetry:

"The Love Song of J. Alfred Prufrock"

> Let us go then, you and I,
> When the evening is spread out against the sky
> Like a patient etherized upon a table. . . .

Eliot had given a copy of the poem to his friend Conrad Aiken, who passed it on to Ezra Pound, who in turn sent it to Harriet Monroe, editor of *Poetry*, where it languished more than a year because Miss Monroe was not sure it was poetry. Although Pound had printed privately a small book, *A lume spento*, as early as 1908, "Prufrock" was the first poem by either of these literary revolutionists to go beyond experiment to achieved perfection. It represented a break with the immediate past as radical as that of Samuel Taylor Coleridge and William Wordsworth in *Lyrical Ballads* (1798). From the appearance of Eliot's first volume, *Prufrock and Other Observations*, in 1917, one may conveniently date the maturity of the 20th-century poetic revolution; for, in addition to the title poem, the book contained "Preludes" and "Portrait of a Lady"—both mature works; and the revolution had come of age. The significance of the revolution is still disputed, but the striking similarity to the Romantic revolution of Coleridge and Wordsworth is obvious: Eliot and Pound, like their 18th-century counterparts, set about reforming poetic diction. Whereas Wordsworth thought he was going back to the "real language of men," Eliot sought a diction that might be spoken by an educated person, "neither pedantic nor vulgar." (He wrote of Swinburne: "For the first time in verse of such eminence, sound exists without sense.")

For a year he taught French and Latin at the Highgate School; in 1917 he began his brief career as a bank clerk in Lloyds Bank Ltd. During this period he was also a prolific reviewer and essayist in both literary criticism and technical philosophy. In 1919 he published *Poems*, which contained the poem "Gerontion," a meditative interior monologue in blank verse: nothing like this poem had appeared in English. It was with the publication in 1922 of *The Waste Land*, however, that Eliot achieved an international reputation.

International recognition with *The Waste Land*

This poem, in five parts, proceeds on a principle of "rhetorical discontinuity" that reflects the fragmented experience of the 20th-century sensibility of the great modern cities of the West. For Eliot is the poet of the secularized city, the decay of *urbs aeterna* (the "eternal city"). This is the ultimate theme of *The Waste Land*, concretized by the constant rhetorical shifts, the juxtapositions of contrasting styles. But it is not, as his early critics supposed, a simple contrast of the heroic past with the degraded present; it is rather a timeless, simultaneous awareness of moral grandeur and moral evil. The original manuscript of about 800 lines was cut down to 433 at the suggestion of Ezra Pound. *The Waste Land* is not Eliot's greatest poem, though it is his most famous.

"Programmatic criticism"

Eliot has said that the poet-critic must inevitably write "programmatic criticism"—that is, criticism that expresses the poet's own interests as a poet, quite different from historical scholarship, which stops at placing the poet in his background. Consciously intended or not, this kind of criticism creates an atmosphere in which the poet-critic's own poetry will be better understood and appreciated than if it had to appear in a literary milieu dominated by the standards of the preceding age. Eliot's essay "Tradition and the Individual Talent," appearing in his first critical volume, *The Sacred Wood* (1920), is historically analogous to and as historically important as Wordsworth's "Preface." Tradition, according to Eliot, must be earned by the poet; it is not a repetition of the work of the immediate past ("novelty is better than repetition," he said); it comprises the whole of European literature from Homer to the present. The poet writing in English may therefore make his own tradition by using materials from any past period, in any language. This point of view is "programmatic" in the sense that it disposes the reader to accept the revolutionary novelty of Eliot's polyglot quotations and serious parodies of other poets' styles in *The Waste Land*. Also in *The Sacred Wood*, "Hamlet and His Problems" sets forth Eliot's theory of the objective correlative:

> The only way of expressing emotion in the form of art is by finding an "objective correlative"; in other words, a set of objects, a situation, a chain of events which shall be the formula for that particular emotion; such that, when the external facts, which must terminate in sensory experience, are given, the emotion is immediately evoked.

Eliot adapted the phrase from either Santayana or from the U.S. painter and author Washington Allston and placed it in the context of his impersonal theory of poetry; it thus had an immense influence toward correcting the vagueness of late Victorian rhetoric by insisting on a correspondence of word and object. Two other essays, first published the year after *The Sacred Wood*, almost complete the Eliot critical canon: "The Metaphysical Poets" and "Andrew Marvell," published in *Selected Essays, 1917–32* (1932). In these essays he effects a new historical perspective on the hierarchy of English poetry, putting at the top the Metaphysical poets

of the 17th century and lowering poets of the 18th and 19th centuries. Eliot's second famous phrase appears here: "dissociation of sensibility," invented to explain the change that came over English poetry after Donne and Marvell. This change seems to him to consist in a loss of the union of thought and feeling. The phrase has been attacked; yet the historical fact that gave rise to it cannot be denied; and with the poetry of Eliot and Pound it had a strong influence in reviving interest in certain 17th-century poets. The first, or programmatic, phase of Eliot's criticism ended with *The Use of Poetry and the Use of Criticism* (1933), his Charles Eliot Norton lectures at Harvard. Shortly before this his interests had broadened into theology and sociology; three short books, or long essays, were the result: *Thoughts After Lambeth* (1931), *The Idea of a Christian Society* (1939), and *Notes Towards the Definition of Culture* (1948). These book-essays, along with his *Dante* (1929), an indubitable masterpiece, broadened the base of literature into theology and philosophy: whether a work is poetry must be decided by literary standards; whether it is great poetry must be decided by standards higher than the literary.

Essays on theology and sociology

The criticism and the poetry are so interwoven, so much the product of a unified mind and sensibility, that it is difficult to discuss them separately. The great essay on Dante appeared two years after Eliot had become an Anglo-Catholic (1927); in that year he also became a British subject. The first long poem after his conversion was "Ash Wednesday" (1930), a religious meditation in a style entirely different from that of any of the earlier poems. Indeed, Eliot never repeated himself, though every poem is unmistakably his. "Ash Wednesday" was not well received in an era that held that poetry, though autonomous, is strictly secular in its outlook; it was misinterpreted, typically by the critic Edmund Wilson, as an expression of personal disillusion.

Later career and influence. Eliot's masterpiece is *Four Quartets*, taken as a single work, though each "quartet" is a complete poem. The four parts, published at intervals between 1935 and 1941, were issued as a book in 1943. This led to the award to Eliot, in 1948, of the Nobel Prize for Literature. Perhaps the high-water mark of poetry in English of the 20th century is the passage in "Little Gidding" in which the poet meets a "compound ghost," a figure composite of two of his masters: William Butler Yeats and Mallarmé. The scene takes place at dawn in London after a night on duty at an air-raid post during an air-attack; the master speaks in conclusion:

> From wrong to wrong the exasperated spirit
> Proceeds, unless restored by that refining fire
> Where you must move in measure, like a dancer.
> The day was breaking. In the disfigured street
> He left me, with a kind of valediction,
> And faded on the blowing of the horn.

The passage is 72 lines, in modified terza rima; the diction is as near to that of Dante as is possible in English; and it is a fine example of Eliot's belief that a poet can be entirely original when he is closest to his models.

Dramatic works

The plays, which begin with *Sweeney Agonistes* (published 1926; first performed in 1934) and end with *The Elder Statesman* (first performed 1958; published 1959), are, with the exception of *Murder in the Cathedral* (published and performed 1935), inferior to the lyric and meditative poetry. Eliot's belief that even secular drama attracts people who unconsciously seek a religion, that in fact all drama has a religious structure of progress toward self-knowledge, reconciliation, or purification, led him to put drama above all other forms of poetry. All his plays are in a blank verse of his own invention, in which the metrical effect is not apprehended apart from the sense; thus he brought "poetic drama" back to the popular stage. *The Family Reunion* (1939) and *Murder in the Cathedral* are Christian tragedies, the former a tragedy of revenge, the latter of the sin of pride. The other plays are comedies in which the plots are derived from Greek drama. The most successful of these plays is *The Cocktail Party* (first performed 1949; published 1950), based upon the *Alcestis* of Euripides.

Eliot's career as editor was ancillary to his main interests, but his quarterly review, *The Criterion* (1922–39), was the most distinguished international critical journal of the period. He was a "director," or working editor, of the publishing firm of Faber & Faber Ltd. from the early 1920s until his death.

Editor of *The Criterion*

Eliot's posthumous volume of criticism, *To Criticize the Critic, and Other Writings* (1965), contains two essays of great value: the title essay is a review and a severe but not falsely modest evaluation of his literary criticism; "American Literature and the American Language" deals with the chauvinist myth of American English as a separate language and with equal conviction treats British and American literature as branches of one literature. The first half of the 20th century will probably be known to historians as the age of Eliot, as the mid-18th century was the age of Johnson. But Samuel Johnson came at the end of the Neo-classical period and summed it up; whereas Eliot, as critic and poet, created his age. He always kept his private life rigorously in the background. In 1915 he married Vivian Haigh-Wood; after 1933 she was mentally ill, and they lived apart; she died in 1947. In January 1957 he married Miss Valerie Fletcher, with whom he lived happily until his death on January 4, 1965, in London.

MAJOR WORKS

POETRY: *Prufrock and Other Observations* (1917); *Poems* (1919); *Ara Vos Prec* (1920), incorporates *Prufrock* and *Poems* with additional verse, including "Gerontion"; *The Waste Land* (1922); *Poems, 1909–25* (1925), incorporates *Ara Vos Prec*, *The Waste Land*, and "The Hollow Men"; *Journey of the Magi* (1927); "Ash Wednesday" (1930); *Collected Poems, 1909–35* (1936), incorporates *Poems, 1909–25*, "Ash Wednesday," chorus from *The Rock*, "Burnt Norton," and others; *Old Possum's Book of Practical Cats* (1939); *Four Quartets* (1943), each of the four issued as a separate pamphlet: "Burnt Noron" (1941); "East Coker" (1940); "The Dry Salvages" (1941); "Little Gidding" (1942); *Collected Poems, 1909–62* (1963).

CRITICISM: *Ezra Pound: His Metric and Poetry* (1917); *The Sacred Wood: Essays on Poetry and Criticism* (1920); *Homage to John Dryden* (1924), contains "John Dryden," "The Metaphysical Poets," and "Andrew Marvel"; *For Lancelot Andrewes: Essays on Style and Order* (1928); *Dante* (1929); *John Dryden: The Poet, the Dramatist, the Critic* (1932); *Selected Essays, 1917–32* (1932, rev. ed. 1951); *The Use of Poetry and the Use of Criticism* (1933); *Elizabethan Essays* (1934); *Essays Ancient and Modern* (1936); *The Music of Poetry* (1942); *The Classics and the Man of Letters* (1942); *What Is a Classic?* (1945); *Milton* (1947); *Poetry and Drama* (1951); *The Three Voices of Poetry* (1954); *On Poetry and Poets* (1957); *To Criticize the Critic and Other Writings* (1965).

OTHER PROSE WORK (sociology and church politics): *Thoughts After Lambeth* (1931); *After Strange Gods—A Primer of Modern Heresy* (1934); *The Idea of a Christian Society* (1939); *Reunion by Destruction* (1943); *Notes Towards the Definition of Culture* (1948).

TRANSLATION: *Anabasis—A Poem by St. J. Perse with a Translation by T. S. Eliot* (1930).

PLAYS: *Sweeney Agonistes: Fragments of an Aristophanic Melodrama* (pub. in *Criterion*, 1926; pub. separately 1932); *The Rock* (1934); *Murder in the Cathedral* (1935; film script, 1951); *The Family Reunion* (1939); *The Cocktail Party* (1950); *The Confidential Clerk* (1954); *The Elder Statesman* (1959).

BIBLIOGRAPHY. Three distinguished works on Eliot are ELIZABETH DREW, *T.S. Eliot: The Design of His Poetry* (1949), a Jungian interpretation of Eliot's symbolism; F.O. MATTHIESSEN, *The Achievement of T.S. Eliot*, 3rd ed. (1958), with a supplementary chapter by C.L. BARBER covering the works written after Matthiessen's death in 1950; and HELEN GARDNER, *The Art of T.S. Eliot* (1949), a brilliantly perceptive study that, however, could not include criticism of the later plays. Other valuable long studies are HUGH KENNER, *The Invisible Poet* (1959); GROVER C. SMITH, *T.S. Eliot's Poetry and Plays* (1956); and LEONARD UNGER, *T.S. Eliot: Moments and Patterns* (1967), an essay on the influence of Joseph Conrad. Studies of the plays are DAVID JONES, *The Plays of T.S. Eliot* (1960); and E. MARTIN BROWNE, *The Making of T.S. Eliot's Plays* (1969), a work indispensable to the understanding of Eliot's methods of dramatic composition, since Browne produced and directed all the plays and was to that extent a collaborator. The first symposium on Eliot, and still valuable,

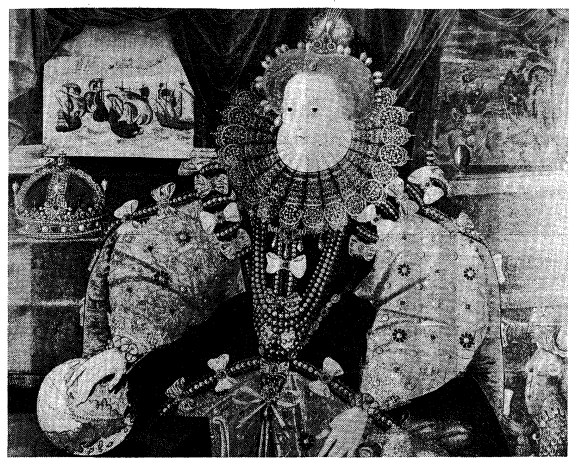
is LEONARD UNGER (ed.), *T.S. Eliot: A Selected Critique* (1948). Others are B. RAJAN (ed.), *T.S. Eliot: A Study of His Writings by Several Hands* (1948); TAMBIMUTTU and RICHARD MARCH (eds.), *T.S. Eliot: A Symposium* (1948); NEVILLE BRAYBROOKE (ed.), *T.S. Eliot: A Symposium for His Seventieth Birthday* (1958); and HUGH KENNER (ed.), *T.S. Eliot: A Collection of Critical Essays* (1962). These symposia represent cumulative insights of permanent value. The latest symposium, ALLEN TATE (ed.), *T.S. Eliot: The Man and His Work* (1966), published a year after Eliot's death, is retrospective in emphasis, including personal reminiscence and criticism by I.A. Richards, Herbert Read, Conrad Aiken, F.V. Morley, Bonamy Dobrée, Helen Gardner, John Crowe Ransom, G. Wilson Knight, E. Martin Browne, Austin Warren, and 16 other critics. CLEANTH BROOKS, *Modern Poetry and the Tradition* (1939), contains the most perceptive analytical commentary on *The Waste Land*. CONRAD AIKEN, *Collected Criticism* (1968), contains a review of *The Waste Land*, written three months after the poem appeared, as valid today as it was in 1923. See also VALERIE ELIOT (ed.), *The Waste Land: A Facsimile and Transcript of the Original Drafts Including the Annotations of Ezra Pound* (1971). Two general surveys of Eliot's career—including his family background, education, public life, philosophy, and social and political ideas—contain as much relevant information as one may reasonably expect to find until Mrs. Eliot authorizes an official biography: HERBERT HOWARTH, *Notes on Some Figures Behind T.S. Eliot* (1964); and RUSSELL KIRK, *Eliot and His Age* (1971).

(A.Ta.)

Elizabeth I of England

One of the greatest of all English sovereigns, regarded with admiration and pride by English people not only in her own time but in later ages, Queen Elizabeth I, who reigned from 1558 to 1603, owed her outstanding success to her own prudence and statecraft and also to her shrewd judgment of character and her unrivalled capacity to inspire loyalty and devoted service. The daughter of King Henry VIII by his second wife, Anne Boleyn, she was the third of Henry's children to ascend the throne, following the brief reigns of her sickly brother, Edward VI, and her elder sister, Mary I. After Mary's attempt to restore Catholicism in England and, even more unpopular, her marriage with Philip II of Spain, the accession of Elizabeth—who had always been connected with the Protestant cause and who, an attractive young woman of 25, contrasted sharply with her sad, middle-aged sister—was inevitably greeted with relief and delight. It was the queen's achievement that she converted this ephemeral goodwill into a constant loyalty and awe, firmly assuming the reins of power and impressing observers as a monarch obeyed as absolutely as her father had been. Even her indecisiveness and other marks on her character, perhaps deriving from the terrifying years of her childhood, were turned to advantage in the service of the state.

Early life. Born at Greenwich Palace on September 7, 1533, and brought up mainly at Hatfield House in Hertfordshire, Elizabeth had an exacting childhood, full of reverses of fortune and of dangers from political and religious intrigue. Only her extreme prudence saved her from being held an accomplice in the alleged treason of Thomas, Lord Seymour, who, in the reign of her young brother, Edward VI, was executed on the accusation of planning to marry her and overthrow the government of his brother, the Lord Protector. Lord Seymour was the husband of Elizabeth's stepmother, Catherine Parr (Henry VIII's last wife, who had survived him), in whose house Elizabeth was then living. In the reign of her sister, Mary, she suffered an even more traumatic and nerve-wracking ordeal when she was imprisoned in the Tower of London during attempts to prove her cognizant of Sir Thomas Wyat the Younger's rebellion of 1554; her position as next in the succession and her connection from birth with the Protestant cause made her a constant threat to her Catholic sister, Mary, and to the ambitious plans of Mary's husband, King Philip of Spain. Only extreme circumspection saved her in these crises, and the habit of giving what she later called "an answer answerless" characterized almost all of her political and personal affairs for the rest of her life.



Elizabeth I, the Armada portrait by Gower (d. 1596). In Woburn Abbey, Bedfordshire.
From the Woburn Abbey Collection, by kind permission of His Grace, the Duke of Bedford

Reign. *The years of consolidation.* Elizabeth came to the throne upon the death, on November 17, 1558, of her sister, Mary. The first 12 years of the reign established Elizabeth's value with the majority of the nation, Protestant and Catholic alike. Her religious settlement of 1559, though enforcing the Protestant religion by law and declaring the celebration of the mass illegal, was for the time exceptionally tolerant. Catholics could opt out of the obligatory monthly attendance at the parish church on payment of a moderate fine, and their private celebration of mass was not interfered with, except when the government suspected treason. But in 1570 the compromise that most of them had found bearable was abruptly shattered by the interdict of Pope Pius V against "Elizabeth, the Pretended Queen of England," releasing all her subjects from their allegiance to her. English Catholics were thus, inevitably, but for the most part with cruel injustice, placed under the suspicion of being actual or potential traitors. The penal laws against them increased in severity in direct proportion to the action of Catholic powers on the Continent, reaching their height when the Jesuit missionaries began to be dispatched to England, from 1580 onward.

The religious settlement of 1559

The "Virgin Queen." In view of her keen enjoyment of men's society and the role some of her never-consummated marriage plans played in England's foreign policy, Elizabeth's celebrated virginity requires a brief consideration. Her friendship with several men was couched in a romantic idiom: they included Sir Christopher Hatton, lord chancellor from 1587 to 1591; Sir Walter Raleigh, courtier, historian, and adventurer; the young and impulsive Earl of Essex; and above all Lord Robert Dudley, whom she created earl of Leicester in 1564. Her famous love affair with Dudley, reaching its spectacular height when he entertained her at Kenilworth Castle in 1575, survived even her discovery of his secret marriage to her cousin, Lettice Knollys, the widowed countess of Essex, in 1579. Her devotion to him lasted until his death in 1588, the news of which caused the queen to shut herself up alone for so long that at last Lord Burghley, the Lord High Treasurer and for 30 years Elizabeth's most trusted servant, had her door broken open. Her obvious susceptibility, combined with her refusal to marry, gave rise to rumours that she was unchaste and had borne bastards or that she had some physical disability that made a sexual union impossible. The latter theory is dismissed by Lord Burghley's memorandum on the Queen's marriage negotiations with the French king's brother, the duc d'Alençon, in 1579: "Her Majesty . . . has no sickness, nor lack of natural functions in those things that properly belong to the procreation of children." But Leicester's words to the French ambassador in August 1566, that he had known the Queen since she was a child of eight and from that time she had always said, "I will never marry," suggest an explanation. When Elizabeth was two years and

eight months old, her father executed her mother, Queen Anne Boleyn; when she was eight years and five months old, he executed her cousin and stepmother, Queen Catherine Howard. These two events in the life of so young a child may have produced a psychological condition never outgrown. It is often said that Elizabeth would have liked to marry Leicester but considered it impolitic because he was unpopular and his first wife, Amy Robsart, had died suddenly in suspicious circumstances in 1560. But by 1566 she could have done so, for the nation had become so anxious to see the queen married and a mother—an heir being their only safeguard against civil war and invasion in case of her death—that Parliament besought her to marry anyone she liked. Negotiations for her hand played a very important part in Elizabeth's foreign policy, from her accession to the final breaking-off in 1583, when she was 49, of the courtship of the duc d'Alençon. These negotiations had been of great use in detaching Alençon from the sphere of his brother, King Henry III of France, and in ensuring that when Alençon entered on a campaign in the Netherlands injurious to Spanish interests, he would be doing so not to increase French influence in that area but as an ally of the English crown.

Affection for Leicester

The controversy over Mary, Queen of Scots. After the religious settlement of 1559, the inextricable confusion of religion and politics in the 16th century was again exemplified in Elizabeth's greatest domestic peril. Mary Stuart, hereditary queen of Scotland, was also, by descent from Henry VIII's sister Margaret, heiress presumptive to the English throne. Those who thought unlawful Henry VIII's second marriage—that to Anne Boleyn after he had his marriage with Catherine of Aragon annulled—regarded her as queen of England already. This was Mary's own view, and after she returned to Scotland in 1561 on the death of her husband, the young French king Francis II, she never ceased to plan her acquisition of the English throne. She believed that if she appeared on English soil, the English Catholics would rise to welcome her and dethrone Elizabeth. Though this proved a delusion, she exploited the possibility of gaining powerful assistance elsewhere.

When Mary's reverses in Scotland finally drove her to take refuge in England in 1568, the Scots refused to have her back on any terms, and it was too dangerous to let her go abroad, where she would obviously collect a French or Spanish army and try to invade England. There was, therefore, no alternative but to keep her in honorable confinement. Mary at once began to plot her escape and always declared that this was all she wanted, but the three major conspiracies in which her complicity was proved—the Ridolfi Plot of 1571, the French duc de Guise's enterprise of 1582, and the Babington Plot of 1586—all included the murder of Elizabeth in their objectives, and all three envisaged an invasion of England by the dreaded Spanish Army that was engaged in repressing the resistance of the Netherlanders. This army's reputation for remorseless cruelty helps to explain the fear and hatred aroused by the Queen of Scots and the fierce pressure put upon Elizabeth by Parliament, after the discovery of each of the major plots, to have her executed.

The announcement of Pope Gregory XIII in 1580 that it would be no sin to send Elizabeth out of the world, and the murder of 1584 of the other great heretic leader, William I the Silent, who had organized Dutch resistance against the Spanish, drove the English into making a Bond of Association in 1584, signed by men in every shire, by which they swore that if Elizabeth were murdered they would prosecute to the death the persons by whom and for whom the act had been done. In the Parliament of 1585 this resolve was incorporated into the Act for the Preservation of the Queen's Safety, in the drawing up of which Elizabeth herself made a statesmanlike alteration to the effect that the interested person should be condemned to death and his or her heirs disqualified from the succession, only if, upon trial, he or she was convicted of guilty foreknowledge. This amendment was not only just but made it possible for Mary's son, James VI of Scotland, to remain in the position of Elizabeth's heir presumptive, provided he did not engage in treason against her; and the tacit understanding thus created shielded England from possible future attempts by France or Spain to gain the vacant throne. The next year, on the discovery of the Babington Plot, Mary was tried under this Act and found guilty, in October 1586, of conniving at Elizabeth's attempted murder. It took the government three months to persuade the Queen to sign the death warrant. Elizabeth undoubtedly meant the sentence to be carried out, but when the Privy Council acted on the warrant without telling her, she had a hysterical collapse, declaring she had never authorized or intended the final step. Nevertheless, popular rejoicing at the news of the Queen of Scots' execution on February 8, 1587, was unequivocal; all over London the church bells were rung for 24 hours.

The Bond of Association

Defeat of Spain and the Armada. Elizabeth had caused Spain as much injury as she could without provoking a full-scale retaliation. Sir Francis Drake, who had returned from his three-year voyage around the world in 1580, had a private arrangement with her that he should pillage Spanish treasure ships returning from America, and that, if he were caught, she was to disavow him. In 1585 she dispatched a small force under Leicester to aid the Netherlands in their resistance; this was withdrawn in 1586. The execution of the Queen of Scots made the prospect of invading England more attractive to King Philip because it meant that if he now won the English crown, it would be for himself. A devastating raid by Drake on Spanish shipping and stores at Cadiz and Lisbon in April 1587 finally determined Philip to attack England before completing the subjection of the Netherlands.

The Spanish Armada, launched in July 1588, was intended to pick up the Spanish Army on the Flemish coast and land it at the Thames estuary. The defeat of the Armada, caused partly by the superiority of English ships and seamanship, partly by the Netherlanders' ships that kept the Spaniards blockaded in the ports, and partly by the gales that worked havoc on the Spanish galleons as they were driven into the North Sea, proved to be one of the decisive naval battles of history. The Queen's obstinate belief until the last moment that war could be avoided caused the English fleet to be short of stores and ammunition; but the superiority and seaworthiness of the fleet itself was the direct result of the program of shipbuilding she had steadily pursued, which was considered by contemporary historians to be one of the achievements of the reign. The news of the Armada's destruction had not come in when the Queen reviewed the land forces drawn up at Tilbury to resist the expected invasion. The speech she made on this occasion showed her total identification with her people and her pride in being their queen. She said she had come to live or die among them and added:

Speech at Tilbury

> I have but the body of a weak and feeble woman, but I have the heart and stomach of a king, and a king of England, too, and think foul scorn that any Prince in Europe should dare to invade the borders of my realm.

She spoke with complete confidence in the loyalty that her 30 years' work had earned her. Sir Francis Walsingham, Elizabeth's principal secretary, was told by one of his agents that "nothing so much displeased the king of Spain as the loyalty of the Catholics to their queen during the late enterprise."

The years of fruition. The last 15 years of Elizabeth's reign saw a great burst of national vigour, nowhere so remarkable as in literature. This era encompassed the work of the playwright Christopher Marlowe, the poet Edmund Spenser, and a host of other lyric poets and writers of prose, as well as the first stages of the amazing careers of Francis Bacon and Shakespeare. But the queen's own great work was done. She enjoyed immense popularity and prestige, but she became an old woman who had outlived her dearest friends; Leicester died in 1588, Walsingham in 1590, Hatton in 1591, Burghley in 1598. She was left with a group of men more anxious to gain advantage for themselves than to serve her; but the

tragic affair of Robert Devereux, earl of Essex, in the last years of her reign, showed that her great powers as a monarch were not in eclipse. This young man had attractive qualities and some abilities, but he was selfish and lacking in judgment. He was the stepson of Leicester, who had brought him forward to oust Walter Raleigh. The great favour and kindness the old Queen showed him made him altogether miscalculate the degree of his influence over her. Essex had distinguished himself for bravery and chivalry in 1596 in an action against Cadiz undertaken by the English because the Spaniards had laid siege to Calais and were, it seemed, planning a second invasion; but at home his envious and domineering temper caused him to make enemies of Burghley's son Robert Cecil, and of Sir Walter Raleigh and the lord admiral, the Earl of Nottingham.

Discontent with English rule had made Ireland dangerous throughout Elizabeth's reign because it offered an invasion base to Spain. In 1598 the Irish earl of Tyrone inflicted a crushing defeat on the English deputy at Armagh, and rebellion broke out all over the country. A disciplinary expedition was necessary; Essex demanded the leadership and was given it, lavishly supplied with money and men; but he disobeyed the Queen's injunctions as to how he was to proceed, with the disastrous result that he failed while refusing to carry out her orders. Rushing back to England (also against the Queen's positive commands) to tell his own story, he was put under arrest by the Privy Council. An enquiry cleared him of treason, but he was not allowed to enter the Queen's presence again. This complete reversal of his previous position was too much for his overbearing nature; he made his house a focus for disaffected people of all kinds; and, hearing that his arrest was ordered, he set out with some 200 followers in an attempt to raise the City of London, meaning to return with crowds at his back to arrest his enemies on the council and overawe the Queen. For this act of high treason he was executed on February 25, 1601.

Rebellion of Essex

The story that Elizabeth had given Essex a ring, promising that if he were ever under her displeasure and sent it to her, it should gain his pardon, and, further, that he had, in fact, attempted to send it to her but that it was held up by Lady Nottingham, the wife of his enemy, would be too absurd to require notice, except for its undying persistence. The notion that Elizabeth, apparently foreseeing trouble with him, would have given to a subject a virtual license to commit treason with impunity hardly deserves comment. A vague reference to such a story is found in John Webster's play, *The Devil's Law Case*, written in 1623; but it is first told at length by an Italian, Gregorio Leti, in his *Historio o vero vita di Elisabetha, regina d'Inghilterra* (History of the True Life of Elizabeth, Queen of England), a romancing pseudo-biography published in 1692. The story is dismissed by such 19th-century historians as the German Leopold von Ranke and the Englishman John Lingard; more important, it is demolished by Elizabeth's own words to the French ambassador in June 1602. He could see that she still mourned Essex' death, but she said to him: "I had put up with but too much disrespect to my person, but I warned him that he should not touch my sceptre." She said further: "When the welfare of my state was concerned, I dared not indulge my own inclinations."

In the last few years of Elizabeth's reign she found her careful financial policy defeated by conditions she could not control. Her reserves had been exhausted, and prosperity brought inflation; she sold crown lands and much of her jewelry. She also salaried some of her beneficiaries by granting them the exclusive right to sell a commodity; since they then possessed the monopoly, they could force up the price. That this practice was causing wide discontent was brought to the queen's notice by her last Parliament, that of 1601. She at once declared that the most offensive items on the list should be removed. When a deputation waited on her the following day to thank her, she made them the greatest speech of her life, in which she said: "Though God hath raised me high, yet I count this the glory of my crown, that I have reigned with your loves," adding, in words that summed up her inspiration of 45 years, "I have cause to wish nothing more than to content the subject, which is a duty that I owe."

James VI of Scotland had long been informally recognized as her heir. The last service she performed for her country was on her deathbed, at the palace of Richmond in London, on March 23, 1603. Almost speechless, she was asked to make a sign if she acknowledged him as her successor. She made the gesture that lent her authority to a peaceful succession and died early the next morning. King James's bishops, in their preface to the King James Version of the Bible, published in 1611, spoke for their own time and for posterity when they described her death as "the setting of that bright, *Occidentall Starre* Queene Elizabeth of most happy memory."

Character and achievements. Elizabeth's famous reign of 45 years was a personal achievement to which every trait she possessed contributed. At her accession, England, if it were to survive as an independent power, was in urgent need of unity. The nation was divided between Catholic and Protestant. The Queen, while committed to the Protestant cause, was devoid of religious fanaticism and wanted a compromise acceptable to as many of her subjects as possible. A marriage alliance with France or Spain would have dragged England into war behind one of these great powers, but in any case the Queen did not want to marry. The resources of the country were at a low ebb; Elizabeth, alone of his descendants, had inherited the financial ability of her grandfather, Henry VII. The charges of government were defrayed by the sovereign's own revenues; where these proved inadequate, Parliament was summoned to vote money. Elizabeth, by her stringent economy and vigilant oversight, made a great contribution to the independence of the crown. Finally, the vigorous but small nation of 4,000,000 needed a figure to inspire them with national consciousness.

Elizabeth was exceptionally gifted for this role. Twenty-five years old at her accession, she was of striking appearance, very erect, her face pale and aquiline, with the red-gold hair of her forebears of the Yorkist dynasty. Highly strung but energetic, she was an accomplished horsewoman, a rapid walker, and fond of dancing. Her interest in dress and her passion for precious stones helped to invest her public appearances with a glamorous quality. She had been highly educated and spoke Latin, French, Spanish, Italian, and Flemish. The Spanish ambassador at the court of her elder sister, Mary Tudor, had said: "The Princess Elizabeth is very dangerous; she has a spirit full of incantation."

Personal appearance

From the day of her coronation procession until her death, the people knew that their queen loved them. With this emotional response, Elizabeth combined an acute capacity to assess talent and character. She was brilliantly served by the men she chose to help her, above all by William Cecil. At her accession she made him chief secretary of state, and in 1571, when he took the office of lord treasurer, relinquishing his previous post to the second greatest of her public servants, Francis Walsingham, Elizabeth created him Lord Burghley.

The Queen was not easy to work with; her high spirit was accompanied by a neurotic caution and a dread of committing her small kingdom at times when her advisers thought safety lay in decisive action. But though she sometimes abused Burghley passionately, their partnership was based on deep mutual trust, Elizabeth saying: "No Prince in Europe had such a counsellor as she had of him."

The nation's enthusiasm for Elizabeth was based on a strong economic foundation. Her fierce reluctance to make war secured for it a prosperity that steadily increased. The scenes that accompanied her opening of the Royal Exchange built by the financier Thomas Gresham (1571), as shown in William Harrison's *An Historicall Description of the Islande of Great Britayne* (1577), indicate the development of national wealth and domestic comfort; and the trade demonstration arranged for the Queen's visit to Norwich in 1578, where a rhyming ad-

dress contained the words: "We bought, before, the things that now we sell,/God's peace and thine we hold and prosper well," point to the nation's consciousness of success with which she was associated.

The external dangers threatening England were, throughout Elizabeth's reign, very great. Spain or France would take any opportunity of acquiring England, the possession of which would tip the scale in the struggle for European supremacy; and the papacy, seeing that the Queen was committed irrevocably to Protestantism, was anxious to see this bastion of heresy destroyed. Elizabeth did what she could to protect herself by giving cautious aid to those who were resisting her enemies—the Huguenots (Protestants) in France and the various provinces of the Netherlands struggling to preserve their liberties against the ferocious tyranny of Philip of Spain. Even the defeat of the Armada did not finally avert the menace of Spain, which was continued through its attack on Henry IV of France in the 1590s and caused Elizabeth to authorize raids upon Spanish bases in 1596 and 1597. But after 1588 her European fame developed a legendary, almost supernatural quality; and the judgment made on her by the French ambassador in 1597, that "she is a very great Princess, who knows everything" reflected a general verdict.

BIBLIOGRAPHY. Some of Queen Elizabeth's formal letters, and her personal additions to letters written on her behalf, may be found in the *Calendar of British State Papers: Domestic Series*, the 2 vol. for the years 1547–80 and 1581–90, ed. by R. LEMON (1856–65) and the 3 vol. for the years 1591–94, 1595–97, and 1598–1601, ed. by M.A.E. GREEN (1867–69). The *Acts of the Privy Council of England*, in 25 vol. covering the years 1558–1601 and ed. by J.R. DASENT (1893–1905), show the Queen at work with her council. Selections from Elizabeth's official addresses are published in *The Public Speaking of Queen Elizabeth*, ed. by G.P. RICE (1951). Her *Letters* were edited by G.B. HARRISON (1968), her *Poems* by L. BRADNER (1964), and *A Book of Devotions: Composed by Her Majesty Queen Elizabeth* with translations by A. FOX was published in 1970.

Contemporary writings about Elizabeth include FRANCIS BACON, *Felicity of Queen Elizabeth*, written in 1608, to be found in full collections of his *Works* or in *Essays Civil and Moral*, ed. by J.S. BLACKIE (1886). JOHN FOXE, "Imprisonment of the Princess Elizabeth" is included in his *Actes and Monuments. The Diary of Dr. John Dee*, the Queen's astrologer, was edited by J.O. HALLIWELL (1842).

The best biography is E. JENKINS, *Elizabeth the Great* (1958). SIR JOHN NEALE, *Queen Elizabeth I* (1959), is authoritative; as is J.B. BLACK, *The Reign of Elizabeth, 1558–1603*, 2nd ed. (1959). Earlier biographies are those of AGNES STRICKLAND (1870) and MANDELL CREIGHTON (1896). Readable biographies include N.J. WILLIAMS, *Elizabeth, Queen of England* (1967); and J.M. LEVINE, *Elizabeth I* (1969). R.C. STRONG, *Portraits of Queen Elizabeth I* (1963), is a valuable study.

Some excellent and authoritative works deal with special aspects of Elizabeth's life. These include the three by CONYERS READ, *Mr. Secretary Walsingham and the Policy of Queen Elizabeth*, 3 vol. (1925); *Mr. Secretary Cecil and Queen Elizabeth* (1955); and *Lord Burghley and Queen Elizabeth* (1960). E. JENKINS explored the Queen's relation with Liecester (1961), LYTTON STRACHEY that with Essex (1928), and W.F. OAKESHOTT that with Raleigh in *The Queen and the Poet* (1960). Elizabeth's earliest years are dealt with by W.H. TAPP, *Anne Boleyn and Elizabeth at the Royal Manor of Hanworth* (1953); and her adolescence with insight and style by EDITH SITWELL, *Fanfare for Elizabeth* (1946). DAME EDITH'S *The Queens and the Hive* (1962), dealt with Elizabeth's relations with her sister Mary Tudor, with Mary Queen of Scots, and with Catherine de Medici. M.A.S. HUME gave an account of *The Courtships of Queen Elizabeth* (1896, reprinted 1964). SIR JOHN NEALE, *Elizabeth I and Her Parliaments*, 2 vol. (1953–57), is a standard work. The Queen's part in the Elizabethan church settlement is dealt with in C.S. MEYER, *Elizabeth I and the Religious Settlement of 1559* (1960); and in W.P. HAUGAARD, *Elizabeth and the English Reformation* (1968). C.H. WILSON in *Queen Elizabeth and the Revolt of the Netherlands* (1970), shows the Queen's attitudes in foreign policy. Court details can be gleaned from F.S. BOAS, *Queen Elizabeth, the Revels Office and Edmund Tilney* (1938); and I. DUNLOP, *Palaces and Progresses of Elizabeth I* (1962). The impact of the Elizabethan legend in France is shown in J.H. GREW, *Elisabeth d'Angleterre: la reine vierge dans la littérature française* (1932).

An antidote is P. COMPTON, *Bad Queen Bess* (1933). Curiosities engendered by the Elizabethan legend include A. DODD, *The Marriage of Elizabeth Tudor* (1940), to Leicester, with Bacon and Essex as issue; and G.E. SWEET, *Shake-Speare: The Mystery* (1963), which maintains that Elizabeth was the author of Shakespeare's works.

(E.Je./Ed.)

Elopiformes

The order Elopiformes contains about 12 species of marine and brackish water fishes, the best known of which are the bonefish, tarpons, and ladyfishes. A few are prized as gamefishes, but only the Pacific tarpon (or oxeye) is of economic importance as a food fish, supporting a major fishery in Southeast Asia. The terms "ladyfish" and "bonefish" have both been used for *Elops saurus* and for *Albula vulpes*. In this article the name ladyfish is applied only to *Elops* and bonefish only to *Albula*. The tarpons and ladyfishes are fast-swimming predators with adult lengths of up to 2.5 metres in tarpons and about 1 metre in ladyfishes. The bonefish and the Japanese gisu are specialized bottom feeders. All except the gisu are coastal fishes of warm oceans, most common in latitudes from 20° N to 20° S.

Primitive position among teleosts

Elopiforms are of interest to the ichthyologist as the most primitive living teleost fishes, standing in much the same relation to the higher bony fishes as do the egg-laying mammals (monotremes) to other mammals. As is usual with primitive groups, the elopiforms have an extensive fossil record, with many more fossil than recent species.

General features. Despite their archaic structure, elopiforms are related by their life history to more specialized groups such as the eels. Like eels, elopiforms have a ribbonlike, translucent, pelagic larva (leptocephalus) that undergoes a striking metamorphosis involving shrinkage to about half the maximum larval size. The bonefish, tarpons, and ladyfish spawn close to shore. The eggs are shed and fertilized in shoal water, sinking to the bottom. Elopiforms are prolific breeders; a large Atlantic tarpon (*Tarpon atlanticus*) was estimated to contain more than 12,000,000 eggs, about seven times as many as in the proverbially fecund cod. The newly hatched leptocephali may be carried out to sea by offshore currents, but metamorphosis only occurs close inshore, and it is probable that larvae carried far out to sea die.

During or immediately after their metamorphosis, the postlarvae migrate inland and accumulate in brackish pools or creeks, often connected with open water only at extreme high tide. Such environments are stagnant and low in oxygen, and air breathing (see below) is an important aid to survival. The juvenile fish feed on small crustaceans, insect larvae, and other small animals, moving back to the sea as young adults. The gisu (*Pterothrissus gissu*), which differs from other elopiforms in inhabiting deeper and colder waters, is apparently an exception to this pattern of development, larval life and metamorphosis taking place in deep water, without the postmetamorphic inland stage that is experienced by other elopiforms.

With the exception of the gisu, elopiforms are coastal fishes, able as adults to enter brackish or fresh water. Adult ladyfishes and tarpons are typical predators of coastal waters, feeding mainly on other fishes. The Atlantic tarpon is renowned for leaping out of the water; the Pacific tarpon (*Megalops cypinoides*) and ladyfishes (several species of *Elops*) behave similarly, "rolling" at the surface. The purpose of this behaviour seems to be the intake of air. Like all of the other primitive teleosts, the elopiforms possess an open duct to the swim bladder, and air that is taken in at the mouth can be passed into the swim bladder.

Behaviour related to air intake

In tarpons the swim bladder is lunglike, partially compartmented and highly vascularized. Tarpon are obligate air breathers, dying from asphyxiation if prevented from reaching the surface, an unusual condition for a species in which adults normally inhabit well oxygenated waters.

Such an adaptation, however, is certainly advantageous in the stagnant pools where postlarval life is spent. The tarpons exhibit a further modification of the swim bladder, a pair of forward outgrowths that contact the auditory region of the braincase and are partially enclosed in bony bullae, a modification that presumably improves the sense of hearing.

The bonefish is a bottom feeder in shallow water, coming in with the tide and grubbing with its snout for worms and shellfish, which it crushes with its rounded palatal teeth. The deep-water gisu feeds mainly on worms, probably by similar methods.

Paleontology and classification. *Fossil history.* The family Elopidae is the only extant teleostean family whose fossil record extends back into the Jurassic (about 140,000,000 years ago). The Upper Jurassic genus *Anaethalion* is included in this family on the basis of some forms that were extremely similar to the modern *Elops*. The genera *Notelops*, from the Lower Cretaceous of Brazil, and *Osmeroides*, widely distributed in Upper Cretaceous seas, were probably true elopids, but the allocation of numerous little-known Cretaceous genera currently placed in the Elopidae, often on the basis of negative evidence, must be considered tentative. The earliest known member of the tarpon family appears to be the fossil *Sedenhorstia*, from the Upper Cretaceous of Europe and Lebanon. Fossils assigned to *Megalops* appear in Eocene deposits. The earliest member of the extinct suborder Pachyrhizodontoidei, *Rhacolepis* (from the Lower Cretaceous of Brazil), was small and resembled the ladyfishes, but later, Upper Cretaceous, members of this group become considerably more specialized. *Pachyrhizodus*, from the Cretaceous chalks of Europe and North America, exceeded three metres in length and superficially resembled a tuna. Pachyrhizodontoids may well have been large fast-swimming predators of the open sea, a niche which is now filled by the tunas.

The genus *Albula* has been recorded in the Upper Cretaceous of North America and in the Paleocene of Europe, but typically albulid tooth plates occur in Lower Cretaceous deposits. Several Cretaceous and Tertiary albulid genera have been described on the basis of otoliths ("ear-stones"). *Istieus*, from the Upper Cretaceous of Germany, had characteristics that were very similar to those of the modern *Pterothrissus*.

Annotated classification.

ORDER ELOPIFORMES

Elopomorph fishes having a long but not eellike body; scales containing bone-cells; teeth on the parasphenoid bone; a transverse sensory canal across the snout; broad infra-orbital bones extending back to the operculum, 55–90 vertebrae, a primitive tail skeleton with 2 free vertebrae supporting 6–7 hypural bones, tail fin large and forked.

Suborder Elopoidei

Mouth terminal and snout unmodified; 2 supramaxillaries; many branchiostegal rays (23–35); teeth small; large gular plate between the lower jaws; 7 hypural bones.

Family Elopidae (ladyfish or tenpounder)

Upper Jurassic to Recent. Very generalized fish, the living forms having 32–35 branchiostegal rays and the swim bladder unmodified. Length to 0.9 metre; weight to about 13 kilograms. One living genus, *Elops*, with 5 or 6 species, circumtropical. Numerous fossil genera.

Family Megalopidae (tarpons)

Swim bladder partially cellular, lunglike, and connected with the ear; scales large; 23–25 branchiostegal rays. Two living genera, *Megalops* and *Tarpon*, each with one species, recently shown to differ in the ear to swim-bladder connection. *Tarpon* confined to the Atlantic; *Megalops* to the Indian and Pacific oceans, ranging from Africa to the Philippines. *Tarpon* length to 2.5 m, weight to 150 kg; *Megalops* to 1.5 m.

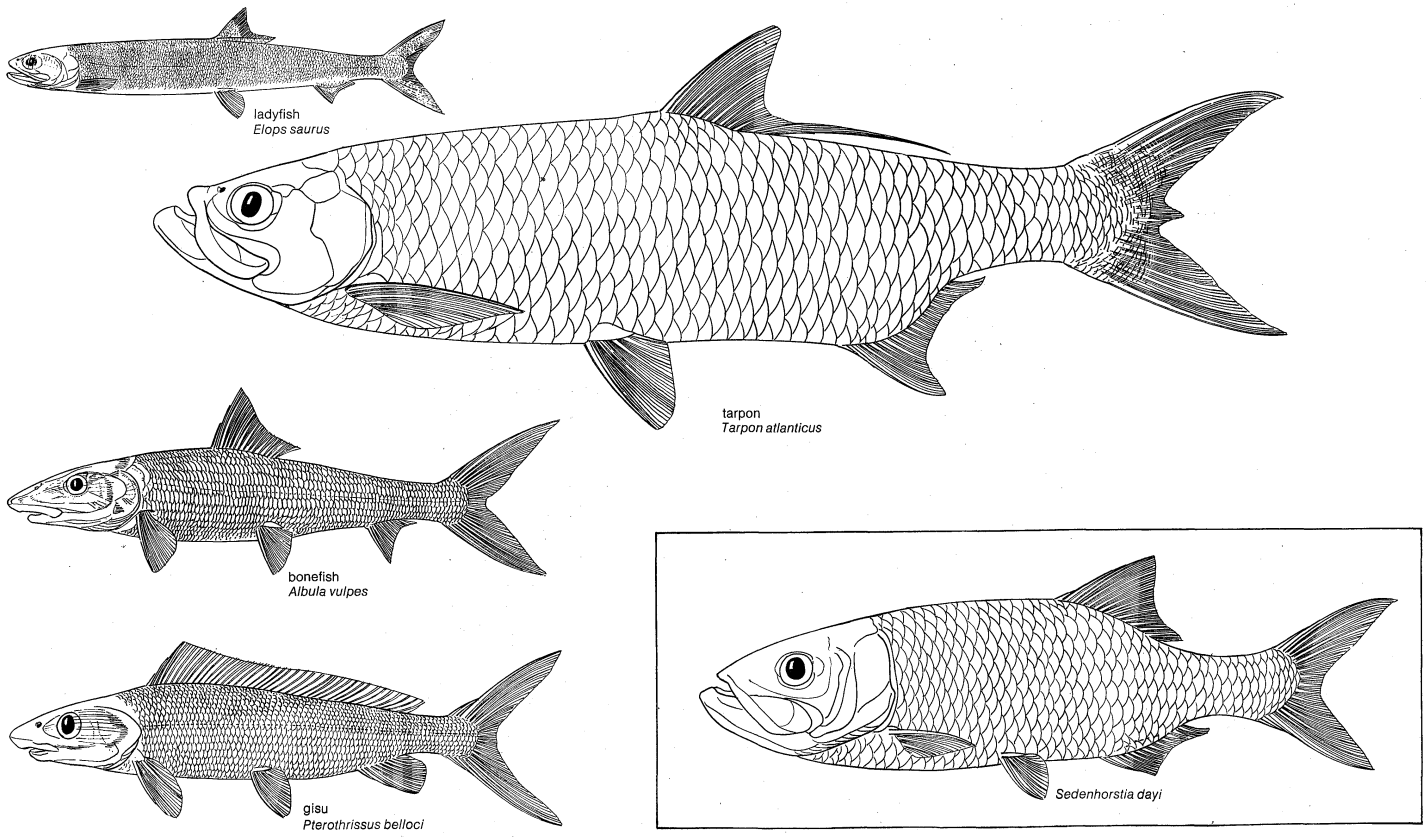
†Suborder Pachyrhizodontoidei

Fossil; Lower to Upper Cretaceous. Parietal bones separated by supra-occipital; gular plate absent; teeth large, pointed, and set in deep sockets. Length to 3 metres.

Suborder Albuloidei

Snout enlarged; mouth small and underslung; crushing teeth on palate; single supramaxillary bone; gular plate small or absent; 6 hypural bones.

From (*Elops saurus, Tarpon atlanticus, Albula vulpes*) *McClane's Standard Fishing Encyclopedia* edited by A.J. McClane, illustrated by Richard E. Younger, Copyright © 1965 by Holt, Rinehart and Winston, Inc., reproduced by permission of Holt, Rinehart and Winston, Inc.; (*Pterothrissus belloci*) from M. Poll, "Resultats scientifiques de l'Expedition oceanographique belge dans les eaux cotieres africaines de l'Atlantique Sud (1948-1949)," III, Teleosteens malacopterygiens, vol. IV, fasc. 2, 1953, p. 14, fig. 5



Body plans of several elopiform fishes.

Family Albulidae (bonefishes and gisu)

Upper Cretaceous to Recent. Characteristics those of suborder. Two living genera: *Albula* (2 species) and *Pterothrissus* (2 species). Length to 70 cm, weight to about 6.5 kg. *Albula* inhabits coastal waters, *Pterothrissus* deeper waters (below 200 m).

†*Family Phyllodontidae*

Fossil. Upper Cretaceous and Early Tertiary. Known only from thick tooth plates, with many superimposed series of replacement teeth. Five genera; from Africa, Europe, and North America.

Critical Appraisal. The chief problem in defining and classifying the elopiforms is the reliance on primitive features, which are merely negative evidence against relationships with other, more advanced groups. Similar problems arise among the most primitive members of all animal groups. The possession of a leptocephalus larva places the elopiforms as the basal order of the cohort Elopomorpha, which also contains the eels, halosaurs, and notacanths. But it is only on negative evidence (absence of specializations) that the elopoids and albuloids are associated, and the absence of any notable specializations in elopoids makes the placing of fossil forms in this group rather doubtful.

BIBLIOGRAPHY. S.F. HILDEBRAND, "Family Elopidae" and "Family Albulidae," in H.B. BIGELOW *et al.*, *Fishes of the Western North Atlantic*, vol. 3, pp. 111–147 (1963), full discussions of the structure, biology, and classification of ladyfishes, tarpons, and bonefishes; R.A. WADE, "The Biology of the Tarpon, *Megalops atlanticus*, and the Ox-eye, *Megalops cyprinoides*, with Emphasis on Larval Development," *Bull. Mar. Sci. Gulf Caribb.*, 12:545–622 (1962), a detailed account of the biology of the Pacific tarpon, or oxeye, and the Atlantic tarpon; P.H. GREENWOOD, "Skull and Swimbladder Connections in Fishes of the Family Megalopidae," *Bull. Br. Mus. (Nat. Hist.), Zool.*, 19:119–135 (1970), details of the specialized respiratory system of the tarpons; P.L. FOREY, "A Revision of the Elopiform Fishes, Fossil and Recent," *Bull. Br. Mus. (Nat. Hist.), Geol.*, suppl. 10 (1973), a full account of the structure and evolution of elopiforms.

(C.P.)

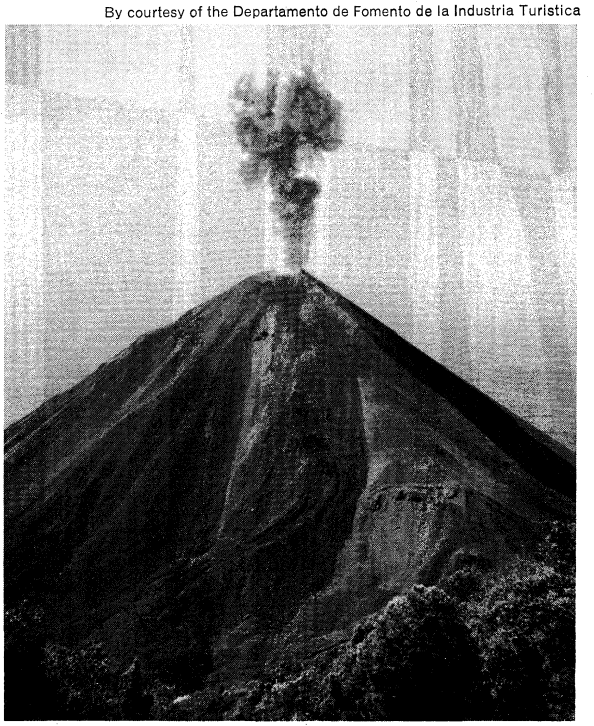
El Salvador

El Salvador, a republic, is the smallest and most densely populated country in Central America. With a population of about 3,541,000 people concentrated in an area of 8,098 square miles (20,975 square kilometres), El Salvador has a population density of about 460 persons per square mile. To the south it has a 199-mile-long coastline on the Pacific Ocean; it is the only Central American country without a Caribbean coastline. To the west it is bounded by the Republic of Guatemala, and to the north and east by the Republic of Honduras. Its capital is San Salvador (population 337,000). The boundary with Honduras, which caused a small war between the two states in 1969, remains in dispute.

The great majority of Salvadorans are mestizo (of mixed Spanish–Indian descent) and Catholic. El Salvador, a land of lakes and volcanoes, is the fourth largest producer of coffee in Latin America and—together with Costa Rica, Guatemala, Honduras, and Nicaragua—a member of the Central American Common Market (CACOM); for coverage of historical aspects, see CENTRAL AMERICAN STATES, HISTORY OF.

The landscape. *Relief features.* The territory of El Salvador includes mountain ranges, volcanoes, plains, lakes, and river valleys. Five regions may be distinguished. These are: (1) the narrow coastal plains, situated between the coastal ranges and the Pacific Ocean, comprising about 10 percent of El Salvador's area; (2) the southern mountain chain, which is volcanic, has an average height of about 4,000 feet and comprises about 30 percent of the territory; (3) the central plain, constituting about 25 percent of the territory, which it crosses from the west to the southeast at heights that vary from about 1,300 to 2,600 feet; (4) the valleys of the Lempa and Grande de San Miguel rivers, which are cut into the central plain and account for about 20 percent of the territory; (5) the northern mountain region, which covers about 15 percent of the country's surface and consists of the Metapán and the Chalatenango mountain chains; the altitudes in this region of El Salvador vary from 4,900 to 6,500 feet.

By courtesy of the Departamento de Fomento de la Industria Turistica



Volcán de Izalco, El Salvador, during an eruption.

Volcanoes

More than 20 volcanoes are situated in the southern mountains, being clustered into five groups which stretch across the country. The westernmost are Izalco (6,004 feet [1,830 metres] high and known as the "Lighthouse of the Pacific"), San Marcelino, Santa Ana (which is 7,812 feet [2,381 metres] high and is the highest mountain in the country), Los Naranjos, Aguila, San Juan de Dios, Apaneca, Tamajaso, and La Laguneta; to the east of these are located San Salvador (6,430 feet) and Las Pavas; next come San Vicente and Ciguatepeque; then Tecapa, Taburete, Buenapa, Usulután, Chinameca, and San Miguel (6,985 feet); finally, farthest to the east, is located Conchagua (4,078 feet), near which are some smaller volcanoes. Several of these volcanoes are still active.

Drainage and soils. El Salvador has more than 300 rivers, of which the most important is the Lempa, which is the only one that is in part navigable. The Lempa, which has many tributaries, rises in Guatemala, crosses a corner of Honduras, cuts across the northern range, flows eastward for more than 80 miles, and then flows south for 65 miles across the southern range to the Pacific. The Grande de San Miguel River System drains eastern El Salvador. The coastal plain is well drained not only by the Lempa and Grande de San Miguel rivers, but also by many other short ones. There are a number of lakes, of which the two largest are Lago de Ilopango (about 25 square miles), and Laguna de Coatepeque (about 15 square miles).

The most productive soils in the country are those of the southern mountain ranges; they are derived from lava and volcanic ash. The soils of the central plains and the valley region are generally shallow and stony. The soils of the coastal plain are alluvial.

Climate. The climate is tropical but is moderated by altitude in the interior; in general it is warm rather than hot, varying between 59° F and 73° F (15° and 23° C). Heavy rains, known as the "temporales," fall in the winter season from May to October. The dry *verano* (summer) season lasts from November to April. There is considerable climatic variation in the different regions. The Pacific lowlands and low areas in the middle Lempa Val-

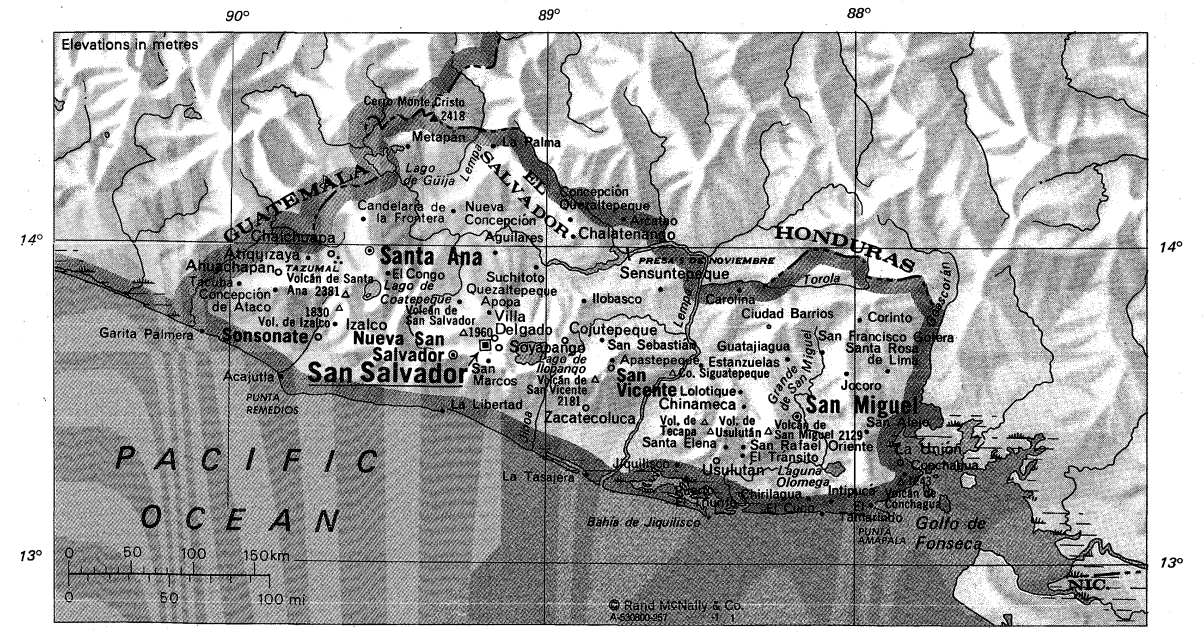
ley have mean monthly temperatures between 78° and 85° F (25° and 29° C). At San Salvador, the capital, which is 2,156 feet above sea level, the maximum monthly mean temperature is 95° F (35° C) in March and the lowest monthly mean is 58° F (14° C) in January. In the mountains, above 4,800 feet, mean monthly temperatures vary between 63° and 72° F (17° and 22° C). Annual precipitation on the Pacific lowlands averages about 68 inches; on the southern and northern mountain ranges, at heights between 2,000 and 3,500 feet, the average is between 70 and 97 inches; the higher mountains receive a little more. Annual precipitation recorded in the deeper valleys and surrounding plateau-like areas is between 45 and 60 inches.

Vegetation and animal life. The higher mountain regions have temperate grasslands and the remnants of deciduous oak and pine forests. On the central plains and in the valleys small deciduous trees, bushes, and subtropical grasslands are found. The coastal plains and the lower slopes of the southern mountains are covered with either savanna (parklike grassland) or deciduous forests. Among the many species of trees is the balsam, known for its beauty and soft perfume; its trunk yields excellent lumber as well as resin that is used in antiseptics and medicinal gums.

Trees and tree products

Valuable wood also is obtained from the cedar, mahogany, laurel, nispero (a fruit tree which also yields chicle), and madrecacao (a tree with pink flowers) trees, and is used for the manufacture of furniture. Particularly beautiful is the maquilishuat tree, whose flower is the national emblem of El Salvador; the tree has thick foliage and vividly coloured flowers when in bloom; it is found both in the countryside and in the cities. Several species of palm and coconut trees grow in the coastal zone. In addition, there are many varieties of tropical fruit, such as coconut, tamarind, melon, watermelon, and mango.

Because of the amount of land under cultivation, El Salvador is considerably less rich in wild animal life than most Central American countries. Rodents, reptiles, and insects of many kinds, however, are common. There is a wide variety of bird life, which includes wild duck, the white and the royal heron, the urraca (which has a blue breast and a grey head, and is known for its call which resembles a scoffing laugh), the blue jay, and many more, some of which have fine plumage. Although a great



EL SALVADOR

MAP INDEX

Cities and towns

Acajutla 13·36n 89·50w
Aguilares 13·58n 89·12w
Ahuachapán 13·55n 89·51w
Apastepeque 13·40n 88·47w
Apopa 13·48n 89·11w
Arcatao 14·05n 88·45w
Atiquizaya 13·58n 89·46w
Candelaria de la Frontera 14·07n 89·38w
Carolina 13·51n 88·19w
Chalatenango 14·03n 88·56w
Chalchuapa 13·59n 89·41w
Chinameca 13·30n 88·21w
Chirilagua 13·13n 88·08w
Ciudad Barrios 13·46n 88·16w
Cojutepeque 13·43n 88·56w
Concepción de Ataco 13·52n 89·51w
Concepción Quezaltepeque 14·06n 88·58w
Conchagua 13·19n 87·52w
Corinto 13·49n 87·58w
El Congo 13·54n 89·30w
El Cuco 13·10n 88·07w
El Tamarindo 13·11n 87·54w
El Tránsito 13·22n 88·21w
Estanzuelas 13·38n 88·30w
Garita Palmera 13·44n 90·05w
Guatajiagua 13·40n 88·13w
Ilobasco 13·51n 88·51w
Intipucá 13·12n 88·04w
Izalco 13·45n 89·40w
Jiquilisco 13·19n 88·35w
Jocoro 13·37n 88·01w
Jucuapa 13·31n 88·24w
La Libertad 13·29n 89·19w
La Palma 14·19n 89·11w
La Tasajera 13·16n 88·52w
La Unión 13·20n 87·51w
Lolotique 13·33n 88·21w
Metapán 14·20n 89·27w
Nueva Concepción 14·08n 89·18w
Nueva San Salvador 13·41n 89·17w
Puerto el Triunfo 13·17n 88·33w
Quezaltepeque 13·50n 89·17w
San Alejo 13·26n 87·58w
San Francisco Gotera 13·42n 88·06w
San Marcos 13·39n 89·11w
San Miguel 13·29n 88·11w
San Rafael Oriente 13·23n 88·21w
San Salvador 13·42n 89·12w
San Sebastián 13·44n 88·50w
Santa Ana 13·59n 89·34w
Santa Elena 13·22n 88·25w
Santa Rosa de Lima 13·37n 87·53w
San Vicente 13·38n 88·48w
Sensuntepeque 13·52n 88·38w
Sonsonate 13·43n 89·44w
Soyapango 13·42n 89·09w
Suchitoto 13·56n 89·02w
Tacuba 13·54n 89·56w
Usulután 13·21n 88·27w
Villa Delgado 13·43n 89·10w
Zacatecoluca 13·30n 88·52w

Physical features and points of interest

Amapala, Punta, *point* 13·10n 87·55w
Coatepeque, Lago de, *lake* 13·52n 89·33w
Conchagua, Volcán de, *volcano* 13·16n 87·51w
5 de Noviembre, Presa, *dam* 13·59n 88·44w
Fonseca, Golfo de, *gulf* 13·08n 87·40w
Goascorán, *river* 13·25n 87·48w
Grande de San Miguel, *river* 13·14n 88·22w
Güija, Lago de, *lake* 14·16n 89·31w
Ilopango, Lago de, *lake* 13·41n 89·03w
Izalco, Volcán de, *volcano* 13·49n 89·38w
Jiboa, *river* 13·22n 89·04w
Jiquilisco, Bahía de, *bay* 13·14n 88·34w
Lempa, *river* 13·15n 88·50w
Monte Cristo, Cerro, *mountain* 14·25n 89·21w
Olomega, Laguna, *lake* 13·18n 88·03w
Pacific Ocean 13·10n 89·45w
Paz, *river* 13·44n 90·10w
Remedios, Punta, *point* 13·31n 89·50w
San Miguel, Volcán de, *volcano* 13·28n 88·16w
San Salvador, Volcán de, *volcano* 13·44n 89·17w
Santa Ana, Volcán de, *volcano* 13·50n 89·39w
San Vicente, Volcán de, *volcano* 13·36n 88·50w
Siguatepeque, Cerro, *mountain* 13·37n 88·35w
Sucio, *river* 14·02n 89·17w
Tazumal, *ruins* 13·59n 89·41w
Tecapa, Volcán de, *volcano* 13·29n 88·31w
Torola, *river* 13·53n 88·30w
Usulután, Volcán de, *volcano* 13·25n 88·29w

variety of fish, as well as turtles and crocodiles, inhabit streams, lakes, and rivers, most of the fish caught commercially or for sport come from offshore waters and coastal lagoons; they consist chiefly of mullet, grunts (food fish found in tropical and subtropical waters which emit grunting sounds), snappers, jacks, groupers (a species of sea basses found in tropical and subtropical waters), sharks, and anchovies.

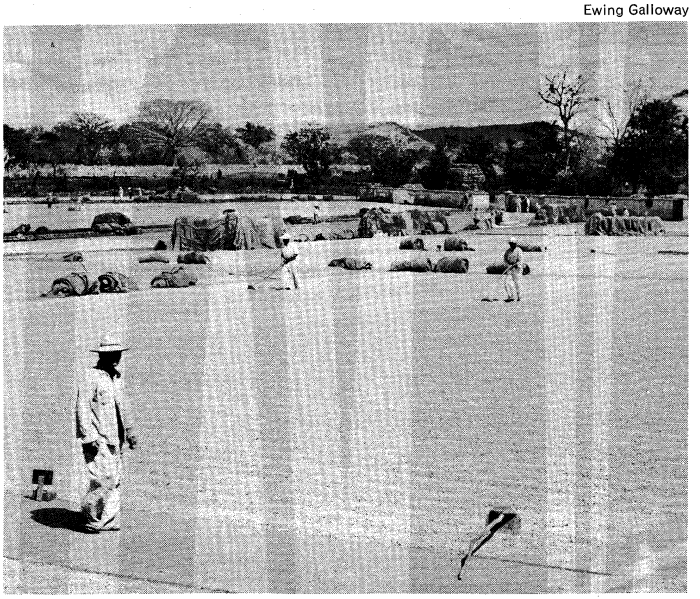
Human settlement. Sixty percent of the population live on the hot coastal plains and hills. Rice and cotton are cultivated in this region, but much more land is used either for pasture or forest than for crops. A significant proportion of the nation's beef, forest products, sesame, sugar, corn, and fish is produced here. The southern volcanic ranges, which consist of cone-shaped volcanoes and areas of gently sloping land, have more than 60 percent of the country's population. San Salvador, the capital, is situated in this region, as also are most of the cities with more than 10,000 inhabitants. Some 90 percent of El Salvador's coffee, tobacco, and manufactured goods are produced in this area, as also are two-thirds of its corn, beans, fruits, vegetables, and sugar, as well as half of its cotton, rice, and animal and forest products. The central plains and the valley regions are inhabited by 25 percent of the population, who live in rural areas and small towns. It is the driest region in the country and has more pastureland than cropland. A cattle-raising area, it also produces corn, beans, vegetables, henequen (a fibre used for making ropes and coarse fabrics), and simple manufactures.

The northern mountain region, inhabited by only about 6 percent of the population, has no towns. The people are dispersed in deep valleys and live chiefly by subsistence farming and cattle raising, although they also grow a little coffee and engage in some lumbering.

San Salvador, the capital, founded in 1525, is situated in the Valle (valley) de Las Hamacas and contains about a quarter of the country's urban population. A road and rail centre, it boasts several beautiful parks and gardens and some sturdy administrative buildings. Its residential suburbs are continuing to grow and it has a number of manufacturing industries. It is six miles to the west of Lago de Ilopango which, lying in the crater of an ancient volcano, is the focus of a popular resort area. In the east of the country, San Miguel is located on the slope of the volcano San Miguel; it is a thriving city where the old Spanish colonial and modern life styles merge. The city of Santa Ana is the commercial centre of western El Salvador.

People and population. Before the Spanish arrived in Salvadoran territory in the 16th century, it was occupied by five Indian tribes—the Pokomans, Chortís, Pipils (or Yakis), Lencas, and Úluas. The Pokomans, Chortís, and Lencas were the more ancient, but the Pipils, whose civilization resembled that of the Aztecs in Mexico, were predominant. The Úluas were represented by scattered groups. The Pipil name for their territory as well as for their capital was Cuscatlán, meaning "Land of the Jewel"; the name is still sometimes applied to El Salvador today. Archaeological ruins dating from Indian times are to be seen at Tazumal, Pampe, El Trapito, and San Andrés. Of several large towns founded by the Indians, Sonsonate and Ahuachapán are still in existence.

Ewing Galloway



Plantation workers spreading coffee beans for drying at Ateos, El Salvador.

El Salvador, Area and Population

	area		population	
	sq mi	sq km	1961 census	1971 census‡
Departments				
Ahuachapán	686	1,777	131,000	178,000
Cabañas	420	1,089	95,000	129,000
Chalatenango	622	1,611	130,000	172,000
Cuscatlán	299	774	113,000	156,000
La Libertad	646	1;672	203,000	284,000
La Paz	478	1,239	131,000	182,000
La Unión	841	2,177	148,000	224,000
Morazán	519	1,344	119,000	156,000
San Miguel	773	2,002	232,000	316,000
San Salvador	341	884	463,000	731,000
Santa Ana	702	1,819	259,000	336,000
San Vicente	477	1,236	113,000	157,000
Sonsonate	465	1,204	167,000	228,000
Usulután	829	2,147	207,000	292,000
Total El Salvador	8,098*	20,975*	2,511,000†	3,541,000†

*According to the FAO the total water area in 1961 was 417 sq mi (1,080 sq km). Because no breakdown of the water area by department was available, only total areas are given in this table. †Figures do not add to total because of rounding. ‡Preliminary. Source: Official government figures.

Few Spanish settlers arrived to colonize the area, but the intermarriage of these few Spaniards with the Pipil Indians resulted in a racially homogenous people. Almost 90 percent of the contemporary population is mestizo; about 10 percent remains Indian (notably the Izalco Indians, and the Panchos from the village of Panchimalco near the capital), and one percent is white. During the pre-Spanish epoch various Indian dialects were spoken, the most important of these being Nahuatl, spoken in the central region of the country, and Poton, spoken in the east. After the initial conquest, Spanish became the official language and the Indian dialects slowly fell into disuse. While Nahuatl still survives, its use is dwindling; the government, in conjunction with the University of El Salvador, has founded a school devoted to its preservation. About 75 percent of the people are of Roman Catholic religion.

Mestizos

Demographically, El Salvador is confronted with the problem of overpopulation. As mentioned, the population density is about 460 persons per square mile, and the rate of growth exceeds four percent a year. About 300,000 immigrants from El Salvador live in neighbouring Honduras. It is estimated that by 1985 the population will approach 6,000,000 people.

The national economy. The economy is predominantly agricultural, although there has been rapid industrial expansion in recent years. The most important agricultural products are coffee, cotton, and sugarcane. Coffee alone accounts for one-third of the value of the total agricultural production and one-half of the value of exports; it is sold to the United States, Europe, and Japan. Cattle raising is also an important activity. While some woods are used for furniture making, as mentioned earlier, the main forest products are dye woods and balsam, a medicinal gum of which El Salvador is the world's main source. Commercial fishing, regulated by the government, increased considerably in the 1960s and has added to the country's export earnings; shrimps as well as fish are caught. There is no mineral exploitation of significance. The main power source, meeting most of the country's needs, is the hydroelectric project on the Río Lempa 35

miles northeast of San Salvador, which is administered by a government agency.

Industrial development

In the 1960s there was a steadily increasing investment in industry, stimulated by the Central American Common Market. Industrial plants rose throughout the country and existing facilities were expanded, helped by government incentives, an advanced banking system, and development credits from abroad. The manufacturing sector of the economy, which has been steadily growing, now represents about 20 percent of the gross national product. Manufactures include beverages, canned foods, organic fertilizers, cement, plastics, cigarettes, shoes, cotton textiles, leather goods, petroleum products, and machinery.

Tourism is an increasing source of income and is encouraged by both the governmental and private sectors. Since 1965 an international fair has been held every two years to encourage the development of commerce and industry. The Salvadoran banking system is well organized; apart from the Central Reserve Bank of El Salvador there are nine commercial banks. Governmental revenue is mostly raised by sales tax (21 percent), nontax revenue (19 percent), customs (19 percent), and export duties on coffee (over 12 percent).

Transportation. El Salvador has adequate transportation facilities except in some of the more remote areas, such as the mountains of the north. Two main routes of the Pan-American Highway system cross El Salvador from Guatemala to Honduras, forming the framework of a road system that reaches almost all parts of the country; one of these routes runs across the central plains, the other across the coastal plains. Traffic has been interrupted, however, by the closing of the frontier by Honduras after a war between the two states in 1969. Out of some 5,500 miles of roads, about 750 are paved, about 1,900 can be used in all weathers, and about 2,700 are for dry-weather use. There are about 430 miles of narrow-gage railroad, operated by two companies, the International Railways of Central America (IRCA) and the FES (Ferrocarril de E Salvador, or El Salvador Railway Company). For sea commerce, El Salvador relies on three ports—Acajutla, La Libertad, and La Unión. El Salvador's outlet to the Atlantic is through the Guatemalan port of Puerto Barrios, with which San Salvador is linked by road and rail, via Guatemala City. Ilopango Field, a modern airport five miles from San Salvador, is linked to the international airways network.

Administration. The constitution, adopted in 1962, provides for a representative form of government with three branches—legislative, executive, and judicial. All Salvadorans, male and female, who have attained the age of 18, are permitted to vote. Executive power is exercised by the president, his cabinet ministers, and undersecretaries of state. The president is elected by direct, popular vote for a term of five years and is ineligible to succeed himself. Legislative power is vested in a unicameral National Legislative Assembly (Asamblea Nacional Legislativa) that meets in regular session on the first of June and the first of December each year. The number of deputies is determined by apportionment. They are elected by popular vote for two-year terms and can be re-elected. The judicial branch is composed of a Supreme Court of Justice, whose members are selected by the National Assembly, and of other tribunals as established by statute. The Supreme Court consists of ten magistrates, one of whom serves as president.

Second- and third-order administrative divisions

El Salvador's territory is divided into 14 departments, each department is divided into *distritos* (districts) and *municipios* (municipalities); each department has a governor and a substitute governor, appointed by executive power; and each municipality has a popularly elected municipal council composed of a mayor, a secretary, and aldermen, the number of whom are in proportion to the population. Independent in their local functions, the municipalities are limited to economic and administrative activities.

Defense. The Army has a strength of 4,500 men. There is a national guard, a national police, and a treasury police. There is also a small air force, and a navy consisting of two patrol boats and some small craft. Military service for at least one year is compulsory for all Salvadoran males from 18 to 30 years of age. The strength of the army is fixed annually by the national assembly; in no case can it be less than 3,000 men.

Social conditions. The standard of living of artisans in El Salvador is high, but that of agricultural labourers remains low. The government is attempting to ensure a more equitable distribution of income by emphasizing worker's education, sponsoring projects for literacy, and training workers for union activity and community life. The Salvadoran Social Security Institute, established in 1949, provides insurance for all workers affiliated to it. The Institute of Urban Housing is attempting to help overcome a serious housing shortage by providing low-cost housing.

Education. All public and private institutions of learning are under the jurisdiction of the Ministry of Education. Since 1968 the school system has been composed of preschool, primary, and secondary educational categories, followed by university level education. Basic education is free and compulsory, but the illiteracy rate is high. Diversified education takes three years and offers high school degrees in subjects such as science and literature, industrial training, hotel management, agriculture, and cattle raising. El Salvador has pioneered in the educational use of television. There are four institutions of higher education: the University of El Salvador, the Higher Normal School, the School of Social Service, and the José Simeón Cañas University (Universidad Centroamerica José Simeón Cañas). There is also an Institute of Fine Arts and a National Agricultural Institute, as well as schools of public administration, social service, and nursing, and various other vocational schools.

Health. By 1969 it was estimated that there was one doctor for every 4,000 people; most of the doctors service urban areas. There are a number of hospitals and health centres.

Prospects for the Future. El Salvador has, especially since the early 1960s, made considerable economic progress, with industry becoming of increasing significance to the economy. Membership in the Central American Common Market has done much to stimulate the pace of development, although the 1969 war with Honduras, which diverted efforts and resources and hampered regional cooperation, proved a setback. Overpopulation poses a basic challenge that has still to be met. The government is also confronted with the problem of helping to raise living standards for the rural population.

BIBLIOGRAPHY.

Topography and general: DAVID BROWNING, *El Salvador: Landscape and Society* (1971), deals with the population as well as the country, while H.F. GODFREY, *Your El Salvador Guide* (1968), is a general account. See also the DIRECCION GENERAL DE ESTADISTICA Y CENSOS, *Statistical Yearbook 1968*, vol. 2, *Demography and Health* (1968); and the handbooks and bulletins issued by the Republic of Salvador. L. DE J. OSBORNE, *Four Keys to El Salvador* (1956), is a good travel book, supplemented by R.N. ADAMS, *Cultural Surveys of Panama-Nicaragua-Guatemala-El Salvador-Honduras* (1957), which places El Salvador in the Central American context.

Economics and history: D.R. RAYNOLDS, *Rapid Development in Small Economies: The Example of El Salvador* (1967), scrutinizes the economy, to whose understanding the BANCO CENTRAL DE RESERVA, *Foreign Exchange Control Law and Regulations for the Application of the Foreign Exchange Control Law* (1962); and the PAN AMERICAN UNION, *Foreign Trade of El Salvador, 1945–56* (1958), contribute. On early-20th-century history, W.H. KOEBEL, *Central America: Salvador* (1914); JOAQUIN LEIVA, *The Republic of El Salvador in Central America* (1913); and P.F. MARTIN, *Salvador of the Twentieth Century* (1911), are all illuminating.

(R.S.V.)

Elton, Charles

Biologist and one of the founders of the modern science of animal ecology, Charles Elton has had a profound influence on the development of this subject throughout the world.

Born on March 29, 1900, at Liverpool, the son of Oliver Elton, young Elton was educated first at Liverpool Col-

lege and then at New College, Oxford, from which he graduated with first class honours in zoology in 1922. Like many others, Elton rebelled against the strong emphasis on comparative anatomy in zoology at that time. Although most of his contemporaries turned to the physical and chemical analysis of animal mechanisms in the laboratory, Elton, at heart a naturalist, went in the opposite direction—to use the scientific method to study the lives of animals in their natural habitats and interrelationships with their surroundings. He set out to turn natural history into science—the science of ecology. The naturalists are the pioneer observers preparing the ground for the ecologists, who follow with their more quantitative and experimental studies. When Elton began his work, he described it as "the sociology and economics of animals."

Influences on Elton's thought

The great naturalists of the past had a marked influence on Elton's outlook. He wrote of Alexander von Humboldt as "perhaps the first ecologist" in that he "created a stirring picture of the plant and animal world as a whole, with its majestic settings and its complex interplay of forces." Moreover, Elton recalled that "Humboldt's writings in turn inspired Charles Darwin, who went further than most of his generation in grasping the tremendous intricacy and importance of plant and animal interrelationships." Elton was particularly impressed by the methodology of U.S. ecologist Victor Ernest Shelford in his book *Animal Communities in Temperate America as Illustrated in the Chicago Region* (1913). He had an opportunity to apply Shelford's ideas in 1921, when, still an undergraduate, he acted as assistant to Julian Huxley on the Oxford University expedition to Spitsbergen; at this time he was given a free hand in making an ecological survey of local animal life. He continued this project on three subsequent expeditions to the Arctic, in 1923, 1924, and 1930. Just as Darwin had been much influenced by the *Essay on the Principle of Population*, of Thomas Malthus, so was Elton influenced by *The Population Problem* (1922) of Alexander Carr-Saunders, who was also a member of the Oxford University expedition to Spitsbergen and who provided Elton with a first-hand introduction to his ideas.

Basic principles of ecology

Elton's first book, *Animal Ecology*, published in 1927, was a landmark not only for his brilliant treatment of animal communities but also for the fact that the main features of his discussion have remained as leading principles of the subject ever since: food chains and the food cycle, the size of food, niches, and the "pyramid of numbers." He also developed more comprehensive ideas about the factors that govern animal numbers. As a result of his Arctic experience, Elton had become a biological consultant to the Hudson's Bay Company, a position that allowed him to make his important studies of the fluctuations in the populations of various furbearing mammals revealed in the trappers' records, which date from 1736. This study in turn led to his research on the fluctuations in Britain's mouse and vole populations as they are affected by their changing environmental conditions.

In 1930 appeared his provocative book, *Animal Ecology and Evolution*, in which he says that "The balance of nature does not exist and perhaps never has existed." Moreover, "in periods of stress it is a common thing for animals to change their habitats and usually this change involves migration." And again, "we are face to face with a process which may be called the *selection of the environment by the animal*, as opposed to the *natural selection of the animal* by the environment." Unfortunately, his increasing concern with population numbers and their fluctuations did not allow him to pursue these exciting evolutionary ideas.

In 1932, Elton established his Bureau of Animal Population at Oxford. It became both a world centre for the collection of data on variations in animal numbers and a research institute in terrestrial ecology. It attracted workers from many countries, providing training for younger men who carried the Elton tradition to distant places, such as California and British Columbia in one hemisphere and Australia and New Zealand in the other. In the same year he became editor of the new *Journal of Animal Ecology*, which was launched by the British Ecological Society largely under his influence. In 1936 Oxford appointed Elton reader in animal ecology, and Corpus Christi College elected him a senior research fellow.

Fluctuating population

Elton's extensive work on mice and voles enabled him to assign his bureau at the outbreak of World War II the task of finding practical methods of controlling rodent pests, a study that saved his country much loss of food during those critical years. The methods he and his colleagues developed and the results they achieved are described in *The Control of Rats and Mice* (1954), which has become the model for such work all over the world. In 1942 he published his study on *Voles, Mice and Lemmings*, and in 1958 he discussed in *The Ecology of Invasions of Animals and Plants* the effects produced by the spread of organisms introduced into an area by both natural and human agencies. Since World War II, Elton has been engaged primarily in habitat studies; these studies formed the basis of his important volume *The Pattern of Animal Communities* (1966). This work gives more general principles, particularly the inverse pyramid of habitats, which corresponds to his pyramid of numbers.

Elton was elected a fellow of the Royal Society of London in 1953, foreign member of the American Academy of Arts and Sciences, and was awarded the Gold Medal of the Linnean Society in 1967 and the Royal Society's Darwin Medal in 1970.

Elton perceived and stated more clearly than anyone else had yet done the outstanding principles governing the complex interplay between different animals and plants in any community, as well as their relations with their habitat. Moreover, he promoted these principles by his inspiration of many students at his Bureau of Animal Population at Oxford. He is widely acclaimed as the ecologist par excellence.

BIBLIOGRAPHY. SIR ALISTER HARDY, "Charles Elton's Influence in Ecology." *J. Anim. Ecol.*, 37:1–8 (1968), is a short general review of Elton's contributions to animal ecology.

(A.C.Ha.)

Embalming, Burial, and Cremation

From prehistoric times the institutionalization and ritualization of various methods of disposing of the dead have been characteristics of man's cultures and civilizations. Embalming, the art of preserving human remains, has been practiced in many societies, ancient and modern. Whether or not embalming is used, some method or methods of disposal of the body also are employed. The principal methods of disposal are burial—including inhumation (interment in the earth or a crypt, tomb, or similar shelter), water burial, exposure, and cannibalism —and cremation, the reduction of the human remains to ash.

This article covers the practices of embalming, burial, and cremation and the role of burial societies in ancient and modern civilizations (see also DEATH RITES AND CUSTOMS).

Attitudes toward death. Socrates' admonition to his grieving friends—"Be of good cheer and say that you are burying my body only"—represents an attitude toward death that has not been widely accepted, although it is basic to most of the world's important philosophies and religions. Death, consequently, has been celebrated with ceremony and often with splendour to a greater degree than any other rite of passage. In some ancient cultures, as widely distant as Egypt and Peru, the preoccupation with mortuary matters was so intense that a major portion of the social resources—capital and labour, as well as artistic, scientific, and technological skill—was devoted to building a civilization of tombs. In many cultures the treatment of the dead body before, during, and after burial mirrors the way of life and thought to such a degree that graves have yielded to archaeologists and anthropologists answers to the mystery of long-vanished peoples.

Rationale for burial practices

Sociologists vary in their explanations of funeral practices. Some hold them to be mechanisms for releasing tensions and re-establishing group morale; others look upon them as mechanisms for creating tensions and affirming group traditions. Generally, however, most investigators of human behaviour agree that burial practices have their origin in ambivalent emotional responses toward death and the dead: hope and terror, grief and guilt, love and loathing. The dominance of one or the other determines how the corpse shall be treated. Thus, it may be handled lovingly and kept close to the living, as was done among early Christians and continues to be among the Dayak of the Malay Archipelago; or it may be regarded as vile and relegated to a forbidden area, an attitude shared by the ancient Persians and modern Navajo Indians.

Attitudes toward the dead are also derived from concepts of the relationship between life and death. Where the two are closely identified, the hereafter is held to be a spirit world that is an extension of the material here and now, a belief common among primitive people from Homeric Greeks to present-day Ashanti tribesmen, and burial rites are most elaborate. Thus, it is thought that if the dead are to enjoy a happy ever after, they must be provided with all that made life enjoyable: food, drink, clothing, jewels, furniture, and (sometimes) wives and attendants.

Burial practices are simplest among people who make clear distinctions between spirit and matter. To Jews, Christians, Buddhists, and members of other religious groups who believe that death is a "translation" to a spiritual realm, the body is of minor importance after the soul has departed, and material trappings, therefore, are considered irrelevant and irreverent. Originally, the burials of followers of those sects were simple and stark. But ancient customs persist, as the lavishness of modern-day funerals the world over indicates.

EMBALMING

History. For practical as well as theological reasons a well-preserved body has long been a chief mortuary concern. The ancient Greeks, who demanded endurance of their heroes in death as in life, expected the bodies of their dead to last without artificial aid during the days of mourning that preceded the final rites. Other societies, less demanding of their greats, developed a wide variety of preservatives and methods to stave off decay or minimize its effects. Corpses have been pickled in vinegar, wine, and stronger spirits: the body of the British admiral Lord Nelson was returned from Trafalgar to England in a cask of brandy. Even the Greeks sometimes made concessions: the body of Alexander the Great, for example, was returned from Babylon to Macedonia in a container of honey. The application of spices and perfumed unguents to minimize putrefaction was so common a practice that the English word embalming had as its original meaning "to put on balm." Generally, however, the word is used to describe a less superficial procedure, the introduction of agents into the body to ensure preservation.

The beginnings of the art and techniques of embalming are associated principally with ancient Egypt, where, as in parts of Asia and South America, a dry soil and climate encouraged its development. The early practice of wrapping the dead in cloth and burying them in charcoal and sand beyond the reach of the Nile waters preserved the corpses, which retained form and features for a long period. Those naturally preserved mummies are believed to have influenced the religious doctrines because they seemed to prove that the individual existed after his death. A belief in immortality and physical resurrection was central to Egyptian religion, both to the sun worship of early periods and to the later cult of Osiris. Central to the latter was the belief that when all of the elements that were present in life—soul, name, shadow, heart, and body—were reunited, the person would be resurrected, as the god Osiris was after his brother killed and dismembered him.

Since the body had to be attractive enough to lure back the soul and other elements, the highly skilled and trained embalmers took exquisite care to preserve it. Although it is held that embalming skill reached a peak during the New Kingdom period between 1738 and 1102 BC, the most detailed description of the various methods was given by the 5th-century-BC Greek historian Herodotus. The most elaborate method, at first reserved for the royal dead, involved surgical procedures. The brain, intestines, and other vital organs were removed, washed in palm wine, and placed in vases, known as canopic jars, filled with herbs. The body cavities were filled with powder of myrrh and other aromatic resins and perfumes. The incisions were stitched, and the body was placed in nitre (potassium nitrate, or saltpetre) for 70 days, after which it was washed, wrapped in cotton bandages, dipped in a gummy substance, and finally coffined and entombed. In a less expensive procedure, oil of cedar was injected into the body, which was then placed in nitre for 70 days. When the body was removed, the oil was withdrawn along with fleshy parts of the body, so that only skin and bones remained. A third method, employed on the bodies of the poor, consisted of purging the intestines and covering the body with nitre for the prescribed period.

Methods of embalming

A number of other early peoples also practiced embalming of a sophisticated nature. Archaeologists have found evidence of a high degree of embalming skill in the burial chambers of the prehistoric Paraca Indians of Peru. The Guanches, aborigines of the Canary Islands, used methods much like those of the Egyptians, removing the viscera and filling the cavity with salt and vegetable powders. The Jívaro tribes of Ecuador and Peru took the additional precaution of ensuring the immortality of their chiefs by roasting their embalmed bodies over very low fires. In Tibet some bodies are still embalmed according to an ancient formula: the corpse is put in a large box and packed in salt for about three months, after which it is in mummified condition.

The ancient Babylonians, Sumerians, and Greeks seldom practiced any but the most superficial kind of embalming, anointing the body with unguents, perfumes, and spices. Nor did the Jews employ embalming procedures, with the notable exception of Joseph, who ordered embalming for himself and his father, and further departed from Jewish custom by having his body placed in a coffin. Among the ancients who profoundly influenced Western culture only the Romans employed cavity embalming, not for religious reasons but for the temporary preservation of bodies exhibited for some time before burial. Although there is evidence that some early Christians were embalmed, generally they rejected embalming as well as cremation, considering them pagan customs that mutilated the corpse. Such scruples were sometimes overcome by the desire to have an outstanding person linger on, a desire that was reinforced by the belief that the bodies of some of the devout were kept intact after death as a mark of divine favour. Consequently, some Christians were embalmed, a notable example being Charlemagne, whose embalmed and richly dressed corpse was placed in a sitting position in his tomb at Aachen after his death in 814. The body of the 11th-century Spanish epic hero el Cid, which remained seated on an ivory chair in the monastery of San Pedro de Cardeña for a decade before burial, is also presumed to have been embalmed.

During the Middle Ages and the Renaissance a modest amount of embalming was done by specialists who employed the elaborate Egyptian method. For its return from France to England, the body of the 12th-century English king Henry I was embalmed and eviscerated and the cavities stuffed with medicinal herbs. But during the Middle Ages, embalming was such an expensive procedure that even most royal personages could not afford it; involving, as it did, the use of costly spices, unguents, wax, and wrappings in addition to the prices charged by skilled embalmers. Moreover, religious opposition was so strong and skill so limited that few would consider it. Instead, cerecloths, strips of fabric impregnated with wax and wrapped snugly around the body to exclude air, were used. This method of preservation was so prevalent that cerement became a synonym for grave clothes. The great

Practices in the Middle Ages and Renaissance

interest in anatomy and surgery during the Renaissance stimulated experiments with other embalming methods. Leonardo da Vinci, who dissected at least 50 cadavers for study, developed a method of venous injection for preserving them that anticipated modern embalming procedures. One 17th-century Florentine physician is reported to have turned a corpse to stone by injecting the tissues with a solution of silicate of potash and then immersing the body in a mild acid solution.

Development of modern embalming. Embalming by arterial injection as a mortuary practice is considered to have begun in England in the 18th century. The technique had actually been developed in the first half of the 17th century by the noted English physiologist William Harvey in experiments leading to his discovery of the circulation of blood, during which he injected coloured solutions into the arteries of cadavers. Later the Dutch and German scientists Frederick R. Ruysch and Gabriel Clauderus are believed to have used similar arterial-injection techniques to prevent cadavers from decomposing. The Scottish anatomist William Hunter (1718–83), however, is credited with being the first to report fully on arterial and cavity embalming as a way to preserve bodies for burial. His discovery attracted wide attention after his younger brother, John Hunter, in 1775 embalmed the body of a Mrs. Martin Van Butchell, whose will specified that her husband had control of her fortune only as long as her body remained above ground. To meet that condition, Van Butchell had her embalmed, placed her fashionably dressed body in a glass-lidded case in a sitting room, and held regular visiting hours.

The demand for embalming grew in England and particularly in the United States, where it was promoted by a newly emerging group of undertaker-businessmen as superior to the customary but awkward and often unsatisfactory method of preserving bodies for transportation or for viewing by packing them in ice or laying them on "cooling boards," with a concave, ice-filled box fitted over the torso and head. Some of the more enterprising entrepreneurs exhibited well-preserved "cases" in the windows of shops, or took them on tour so that persons in rural areas and small towns could see the latest development.

Acceptance of embalming

The U.S. Civil War was the turning point in breaking down public resistance to "mutilating" the body and in establishing arterial embalming as a common practice in the United States. Although the government had established national cemeteries for the war dead, it freely awarded contracts to undertakers and embalmers to prepare the bodies of soldiers for shipment home. The widespread use of this service by soldiers' families and the embalming of such notable dead as Pres. Abraham Lincoln's son Willie and later of Lincoln himself brought about increased acceptance of the practice and even caused it to become associated with patriotic activity. Early practitioners included a number of vigorous salesmen, including Joseph H. Clarke, a road salesman for a coffin company. Impressed by embalming's possibilities and profits, he persuaded a staff member of a medical college in Cincinnati to institute a brief course in embalming in 1882, thus establishing the basis of mortuary education in the United States. Embalming remains the only specific skill required in the undertaking business.

Modern procedures. In the modern procedure of embalming, the blood is drained from the veins and replaced by a fluid, usually based on Formalin (a solution of formaldehyde in water), injected into one of the main arteries. Cavity fluid is removed with a long hollow needle called a trocar and replaced with preservative. Cavity fluid is also based on Formalin mixed with alcohols, emulsifiers, and other substances (like embalming fluid) to keep the body from shriveling and turning brown. Arterial embalming is not permanent; even such carefully prepared corpses as that of Lenin, on view in the Kremlin, must be given periodic renewal treatment. The chief purpose of embalming is rather to give the body a lifelike appearance during the days in which it is being viewed by mourners. To enhance this, cosmetics and masking pastes are often applied.

In the United States, embalming is a standard practice as a result of the government support it has received, and is mandatory when bodies are being transported by common carrier, and, in many states, usually when there is an interval of more than 48 hours between death and burial. In Europe, however, embalming is rarely practiced. In many countries permits are required; in most it is performed only by medical practitioners, and the costs are relatively high.

BURIAL

Geography, religion, and the social system all influence burial practices. Climate and topography determine whether the body is buried under the ground, placed in water, burned, or exposed to the air; religious and social attitudes determine how elaborate the burials should be. Rank may determine whether the body is placed in the shallow trench of a simple burial or in an underground chamber of impressive dimensions and construction; whether the body is thrown into a lake, river, or ocean or launched on a raft or in a ship; whether bonfires or vast pyres are lighted to roast the flesh from the bones and reduce them to ashes; whether the body is left on the ground to be consumed by scavengers or weathered by sun, rain, and wind or placed on scaffolds or high towers. In modern Tibet, for example, where four methods are used, water disposal is considered adequate for the members of the lowest classes; burial in the earth and exposure to the air are for the middle ranks; and cremation is reserved for the most exalted.

Rites for kings

The burials of kings and religious leaders have almost universally been characterized by lavish display. From ancient Sumer to modern England, the rites accorded temporal or spiritual rulers have been sumptuous and their tombs costly—whether provided by admiring followers or designed by the lofty themselves. Often the rites have involved great sacrifice of life in addition to wealth. Until recently, when an Ashanti king died—and members of that African tribe kept secret the fact that he was dead—an attendant would be killed at each stage in the preparations so that in death the ruler would have someone with him to perform the services of bathing, dressing, and feeding. To ensure him of company, a number of handsomely garbed and jewelled wives were also dispatched; and a year after death, when the coffin was finally placed in the royal mausoleum, an entourage of wives, cooks, and other servants was entombed with the king.

Modern heads of state are not guaranteed such postmortem attention, but even in the most democratic countries their burial is usually marked by elaborate and costly ceremonies. A rare exception was the modest burial of Pres. Charles de Gaulle, retired president of France, in 1970. At his request he was buried with utmost simplicity in the small cemetery of Colombey-les-Deux-Églises in a plain wooden coffin made by the village carpenter and with only his family, a few close friends, and his neighbours in attendance. His burial contrasted sharply with the pomp and circumstance he enjoyed in life. When the burden of caring for the dead becomes intolerable or excessive, reaction has often occurred, led by religious or by secular authorities and resulting in sumptuary laws limiting expenditures, enacted by such diverse peoples as the ancient Romans, Renaissance English, and the Puritans in the United States.

Inhumation. Burial in the ground by hollowing out a trench in the earth for the body or covering it with rocks or dirt dates back at least to Middle Paleolithic times. Inhumation is the most universal method of burial, as well as being one of the oldest; indeed, some animals employ it. Inhumation of various kinds may be simple or elaborate. Some Eskimo people cover the corpse with a pile of stones or, if stones are not available, with a small ice igloo; the old Norse people built barrows that sometimes reached enormous heights. Similarly, graves may be mere shallow pits, or they may be intricate and beautifully fashioned subterranean palaces sunk deep into the earth and spacious enough to accommodate vast numbers of persons. Excavations of the royal graves

of Ur (dating back to *c.* 3000 BC) revealed, in an inner chamber of one, the body of a ruler with a few intimate attendants and, in surrounding chambers, servants, ministers, dancing girls, charioteers with vehicles and animals, and other persons who had been slain to provide service in death. Skeletons of Bronze Age heroes who had been buried together were found in the royal graves at Mycenae. Recent discoveries in Peru revealed that the Paraca burial chambers, hewn out of solid rock 18 feet (about five metres) below the surface of the ground, were large enough to accommodate as many as 400 corpses with all of the belongings it was thought they would need in the afterworld. Customarily, however, graves have been planned for the burial of individuals.

Burial in caves

Caves, a natural refuge of man, have also been used for the dead. The ancient Hebrews used natural single-chamber caves and hewed oblong recesses lengthwise into the walls to accommodate the dead, a custom that encouraged the building of mausoleums. At first regarded as sacred places, they came to be considered unclean; by the time of Christ they were coated with lime so that they could be recognized and avoided—the literal origin of the metaphoric "whited sepulchres." Among many people, however, the sepulchral caves continued to be regarded as sacred and eventually became places of worship. Among them are thousands of rock temples in western India and in Sri Lanka (Ceylon), some of which received elaborate architectural and sculptural treatment. In Egypt, rock temples were sometimes fronted with an open-air exterior building; an example is Abu Simbel, which extended 180 feet (55 metres) into the rock formation. Both caves and earth graves encouraged the development of other burial practices: the use of coffins and rich graveclothes and burial goods.

Even the positioning of the body came to acquire significance, generally of a religious nature, though involving other considerations. Customarily the body is placed in an extended position, in or out of a coffin, as if in sleep. Bodies of Muslims are laid so that the right side faces Mecca; those of Buddhists are laid with the head to the north. The bodies of ancient Egyptians were placed to face toward the west, perhaps as an indication of the importance of the land of the dead. The Dagari (Dagaba) people of Africa base distinctions on sex. According to one scholar, in death, men are laid on their right side facing east so that the rising sun will tell them to prepare for hunting or farming, and women are laid facing west so that the setting sun will warn them the time has come to prepare the evening meal. Not all groups prefer the sleeping position. Early cultures buried their dead in a crouching or squatting position. In Babylon and Sumer, the sleeping position was reserved for the more exalted; servants killed and buried with their rulers were placed in a crouching position so that they would be ready to serve at royal command. Indians of the Americas often bury their dead in a fetal position, sometimes in a basket or clay urn, with knees under the chin and the body neatly tied into a death bundle. Upright burial has been favoured by other people, particularly for warriors; in Vienna in 1970, it was reported that a club for the "Vertically Buried Loved Ones" was enjoying popularity; burial was being effected in a plastic tube placed upright in a deep hole.

Water burial. The association between water and immortality is reflected in the myths of many cultures, myths that often centre around a god-hero who sails away from his people in death with the promise to return again. The bodies of chiefs and heroes, therefore, have often been set adrift on rivers and oceans in death ships. Among the Norse, even those who were interred were sometimes given such a bier—a custom that was widespread from Iceland to England during the 7th and 8th centuries AD. Perhaps the most famous of such ship burials that have been excavated was that at Sutton Hoo in Suffolk, England. In one mound, archaeologists found the remains of a wooden boat for 38 rowers, 85 feet (26 metres) long, that had been dragged a half mile (about one kilometre) from the river and lowered into the ground. Even women were so honoured, as revealed by excavations at Oseberg in Norway of an impressive Viking ship that had been used as a grave for two women.

Ship burials

Water burials have been common in other cultures. In the South Pacific it was customary to place the dead in a canoe and launch that on the water. Not all water burial involves a ship or raft, however. In the Solomon Islands, bodies are simply laid on a reef to be eaten by sharks; in other places they are wrapped and weighted with stones. In Tibet the bodies of the poor, as well as those of beggars, lepers, and babies, are sometimes thrown into rivers and streams. No special preparations are made in some places; in others, weights are attached before bodies are thrown in; in still others, they are dismembered before being put in the water. In Western cultures, water burial is still employed on occasion when people die during a sea voyage. Recently, there has been a growth in the practice of scattering the ashes on water after cremation—particularly along the Pacific coast of the United States. That custom obtains also in various parts of Asia. In Laos the ashes of the cremated are often strewn at sea; in India a bone-throwing ceremony concludes the Hindu death observances. Within a year after death the remains are taken to the Ganges and thrown into the sacred water; if it is not possible to do that, they are thrown into another river or stream with the hope that they will eventually make their way to the Ganges.

Exposure. Placing the body where it may be eaten by scavenging birds and animals or weathered to its essential elements has been held by many groups to be the most desirable form of disposal for spiritual as well as material reasons. The Zoroastrians have been perhaps the most widely known practitioners of this type of burial, which developed out of the belief that the corpse is so unclean that to inter or to cremate it would contaminate the "pure elements" of earth, fire, and water. Since the 6th century BC it has been their custom to leave bodies on mountains or hills at a distance from the community. In Bombay the Parsis (as the Indian descendants of the Persian refugees are called) maintain "towers of silence," high circular structures. The dead are carried to them, and funeral servants place them on stone beds surrounding a central pit. After the hovering vultures have stripped the flesh from the bones—usually within a few hours—the bones are gathered and dropped into the central pit. In Tibet, where cremation, water, and ground burial are also employed despite difficulties, exposure is common, partly because of the religious belief that bodies without souls have no significance. The Mongolian nomads, who consider it honourable as well as expedient to die "on the plain," practice several types of exposure. Whenever possible, the sick and the infirm are taken from the tents before the moment of death. Usually, the body is left to be eaten by scavengers. Bodies of infants and children are sometimes placed in a sack and abandoned at a crossroad; the first Mongol thereafter to pass the spot is obliged to open the sack so that the spirit may take flight into the afterlife. A number of people who expose the dead use trees and platforms (tree burial). Among them are the primitive Balinese, the Nāga tribes of India, the tribes of central Australia, and the Sioux and other North American Indian groups. Commonly, the Sioux robed the dead in their best clothing, sewed them into a deerskin or buffalo shroud, and carried them to a platform about eight feet (2.5 metres) high. Various possessions and gifts were placed on the scaffold, and the body was allowed to remain there for a year; at the end of that time it was taken down and given an earth burial.

Tree burial

Cannibalism. Ritualistic cannibalism, or anthropophagy, includes eating the flesh of slain enemies to obtain their strength and virtues; some believe that by ingesting the ghost or soul, which resides in the flesh, they can render it harmless, destroy it utterly, or make it subservient to the will of the eater. That is a reason for exocannibalism's (eating the flesh of persons outside the tribe or group) having been fairly common among Australian and Maori tribesmen, the Eskimos, and some Central African people. Not all corpses are eaten. In some cases,

only fallen warriors, medicine men, and women who die a violent death are consumed.

Endocannibalism, the eating of relatives or members of the social unit, has been widely practiced. In the funeral feasts of coast tribes of north central Australia, relatives eat the dead so that they may no longer be sad. Southern Australian tribes used to carry the corpse around, nibbling a piece of it when grief overcame the survivors. Slavonic people ate their parents; so did the Kallatiai people of India, and, according to Strabo's account, the people of Ireland also considered it desirable to eat dead relatives. The Mayorunas and Acumas along the Orinoco River felt it was better to be eaten by friends than by worms; and the Cocomas, also of South America, considered it a show of honour to the dead.

Second burial. Among many people, particularly in primitive cultures, a period of waiting occurs between the first and second burial that often coincides with the duration of decomposition. The origin of this practice is considered to be the different concept of death held by these peoples. In modern societies, death is regarded as instantaneous; it is not so in primitive societies, where it is held to involve a slow change, a passage from the visible society of the living to the invisible one of the dead. During the period of decomposition the corpse is sometimes treated as if it were alive, provided with food and drink, and surrounded by company. Often the dead were kept at home with the living; among the Dayak people of Borneo, chiefs and social leaders were kept in the longhouse with the entire village—a practice to which Dutch officials objected so strenuously that temporary houses were built for the dead. Some groups, the Indonesians, for example, attached mystical importance to the disintegration of the body, collecting and carefully disposing of the liquids produced by decomposition—sometimes mixing them with rice to ingest them. In some cultures the body is mummified or given earth burial or exposed to air during the interval before final burial, as mentioned above. In Bali, where the soul is believed to have left completely by the 42nd day after death, the body, which has been mummified or given earth burial, is cremated. Among Australian Bushmen the bones are allowed to bleach in the sun for several months, after which they are cleaned, painted, and placed in a hollow pole, which has been elaborately decorated with the totem symbols of the dead warrior. In Romania it was the custom to disinter the dead at fixed intervals—three years after burial for a child, five for a youth, and seven for an older person. Then the bones were washed with wine and water, put into a linen bag, and taken to the church for a second burial. The final burial, which was widely believed to mark the entrance of the soul into the land of the dead, was characterized by ceremony and often involved the whole community. A funeral feast was customarily served, and at the end of the affair, the mourners re-entered society.

Concepts of death

In some rare instances the second burial is a collective one. Every ten or 12 years the Hurons of North America held a "feast of souls," a major festival during which the remains of the dead were gathered up and deposited in the communal burial place. Another type of secondary burial also occurs, but it is one that has to do with cemetery space. In a number of South American and European countries where space is relatively scarce, people often lease a grave. At the end of three or five years the body is exhumed, and the bones are put into a niche in an ossuary or disposed of in a common grave.

Collective second burials

Cemeteries. Where and how the dead are finally sheltered have preoccupied many people, modern and ancient, highly civilized and primitive. As with methods of burial, the answers have been dictated by geography, religious beliefs, and social attitudes and also by aesthetic and sanitary considerations. Cemeteries, consequently, may be simple or elaborate—built with a grandeur that overshines the community of the living, as is the case in many rural areas in Central and South America. Also, the places may be regarded as "holy fields" or taboo areas. In countries such as Japan and Mexico, cemeteries are festival places on certain occasions set aside to honour the dead. In other countries and among other religious groups, they are simple and stark, shunned except at time of need.

In most cultures, providing a place for the dead was originally a family obligation because of the widespread belief that ties of kinship last beyond death. The land bought by the biblical Abraham from the sons of Heth had as a principal feature a cave in which his dead could be buried. Having a family mausoleum or graveyard is a custom that has endured in various parts of the world, East and West. And their locations have often been selected with great care: in China *feng shui* ("augury") experts picked sites calculated to provide "good wind and water"; Koreans traditionally hired geomancers to divine auspicious locations, out of the range of vision of "baleful spirits." The desire to be united with ancestors has been very strong. Dutiful Oriental sons return the bodies of their parents to Japan and China at sometimes enormous cost; in the Western world, there is much traffic, with bodies shipped by air, rail, and boat "back home." Even when the tribe or the community took over the obligation, burial in the communal graveyard was a jealously guarded privilege. Strangers could dwell in towns and cities but could not be buried in their cemeteries. Special cemeteries were set up by the ancient Jews, the Romans, and other people for criminals and foreigners and the poor. The infamous 30 pieces of silver returned by Judas were spent to provide a grave for strangers. In Europe from the medieval period until well into the 19th century, convicted witches and murderers, along with suicides, were excluded from the cemeteries.

Sanitary precautions have influenced the nature and location of cemeteries. Romans and Jews, for example, regarded cemeteries as hazardous and established their graveyards outside the walls of Rome and Jerusalem. The ancient Egyptians and the Chinese also shared the concern for sanitation. Christians, on the other hand, had no such concern: they used catacombs as combined mass graves and places of worship, and when they were allowed to practice their religion freely, they buried the dead in churches and churchyards. Overcrowding became very common after the 6th century, when many secular authorities decided to revert to the Roman custom of permitting burial only outside the walls of the city. Church land was not subject to secular sanitary laws, and during the Middle Ages and Renaissance the problem intensified. The secularization of cemeteries came about in the Western world on a mass scale during the early years of the 19th century, largely because of the growing interest in public health and sanitation. In France and other parts of western Europe, many cemeteries were set up outside the cities, managed as public services by city and state governments. In the United States and England, private individuals and organizations began to establish cemeteries. One of the largest 19th-century projects was England's Brookwood, organized by the London Necropolis Company. It had a private railway station in London and two in the cemetery, its own telegraphic address, and special areas for different religions, nationalities, social organizations, and professions. Perhaps the most famous of the type is California's Forest Lawn. In the United States there continue to be public cemeteries, cooperative cemeteries, church cemeteries, and large, mutually owned cemeteries. In addition to state, county, and municipal cemeteries, the federal government operates a complex of national cemeteries in the United States and abroad for military servicemen and members of their families.

Concern for sanitation

In recent years elaborate monuments and gravestones have in many places been replaced by markers flush with the ground, so that from a distance the cemetery seems to be a grassy field—an appearance that has stimulated the change of name from cemetery to memorial park. In other places the preference for above-ground burial has encouraged the building of high-rise mausoleums. Nationalistic preferences are evident in cemeteries in Paraguay: there, Germans, Jews, and Russians are usually buried in the earth, and those of Spanish and Slavic origin in niches and pantheons; Italians use both.

CREMATION

History. The practice of reducing the body to its essential elements by burning it on open fires was introduced to the Western world by the Greeks as early as 1000 BC. They seem to have adopted cremation from some northern people as an imperative of war, to ensure soldiers slain in alien territory a homeland funeral attended by family and fellow citizens. Corpses were incinerated on the battlefield; then the ashes were gathered up and sent to the homeland for ceremonial entombment. Although ground burial continued (even a symbolic sprinkling of earth over the body fulfilled requirements, as *Antigone* reveals), cremation became so closely associated with valour and manly virtue, patriotism, and military glory that it was regarded as the only fitting conclusion for an epic life.

Cremation of military heroes

The *Iliad* makes plain how elaborate and important cremations were. In that, Zeus himself forced Achilles to surrender Hector's body to his father so that he, King Priam of Troy, could have it cremated royally. The greater the hero, the greater was the conflagration. Achilles set the pattern in providing a pyre 100 feet (30 metres) square for his friend Patroclus. Achilles himself was incinerated even more gloriously after his death—in "raiment of the gods" after 17 days of mourning. After the flames were quenched with wine, his bones were bathed in oil and wine and placed in a golden urn with those of Patroclus. Lavish funeral feasting and funeral games followed, and a great tomb was erected for him on a headland above the Hellespont.

The Romans followed Greek and Trojan fashion in cremating their military heroes. Virgil's *Aeneid* scornfully contrasts the etiquette of the "unhappy" Latins with that of the Roman's Trojan ancestors. During the 12-day truce, declared so that both armies could cremate dead warriors, the Latins burned many without ritual or count and later heaped the bones together, covering them with a mound of earth. The Romans, on the other hand, observed all the proprieties. They covered the pyre with leaves and fronted it with cypresses; after it was set ablaze, troops shouting war cries circled it and cast trophies taken from the slain Latins into the fire. They poured the blood of animals on the flames, and, when the fires were quenched, washed the bones in wine and placed them in urns. Cremation became such a status symbol in Rome that constructing and renting space in columbariums (vaults or similar structures with niches in the walls to receive the ashes of the dead) became a profitable business. By about AD 100, however, cremations were stopped. Among the various explanations cited are the spread of Christianity and the influence of Oriental mystery cults, which abhorred fire. Although cremation was not explicitly taboo among Christians, it was not encouraged by them because of pagan associations and because of the concern that it might interfere with the promised resurrection of the body and its reunion with the soul. The most practical reason is that cremations were dangerously threatening wood shortages, since so much timber was being felled for pyres.

Explanations for cessation of cremation

Scandinavians favoured cremation, believing that it helped free the spirit from the flesh and also that it kept the dead from harming the living. While some communities preferred ship burials for their heroes (see above), the *Beowulf* epic indicates the esteem in which cremation was held and how closely the ritual conducted by the Geats to honour their leader paralleled the Greek and Roman epic cremations. After the Icelandic conversion in AD 1000, cremation was rare in western Europe until the 19th century, except in emergencies. During an outbreak of the Black Death in 1656, bodies of 60,000 victims were burned in Naples during a single week. The unfavourable associations, intensified during World War II, when hundreds of thousands of persons were incinerated in Nazi concentration camps, did much to discourage the use of cremation in the West. In India and some other countries where the custom is ancient, cremation is considered very desirable. It is the wish of all devout Hindus to be incinerated in Benares. The waterfront of that holy city is lined with concrete and marble slabs on which pyres are erected. The remains are then deposited in the Ganges. In some Asiatic countries it is available to only a favoured few: in Tibet it is usually reserved for the high lamas; in Laos it is for those who die "fortunately"; *i.e.*, of natural causes at the end of a peaceful and prosperous life. Cremation ceremonies in Bali are colourful and gay. On a "lucky" day, bodies of a number of worthies, which had been temporarily buried or embalmed, are carried to a high and decorative tower made of wood and bamboo and cremated. Forty-two days later a second tower, with effigies instead of bodies, is burned to assist the soul on its journey toward the highest heaven. The ashes of the towers, like those of the bodies, are scattered on the water.

Modern cremations. Cremation in the modern manner is very different. Open fires are not used; instead, the body is placed in a chamber where intense heat transforms it in an hour or two to a few pounds of white, powdery ash. That is disposed of in accordance with law and sentiment: scattered in a garden or some other preferred spot, placed in an urn and kept at home, or taken to a cemetery for burial in a small plot or in a columbarium.

The revival of interest in cremation in Europe and the United States began in 1874, when Queen Victoria's surgeon, Sir Henry Thompson, published his influential book *Cremation: The Treatment of the Body After Death*. He also organized the Cremation Society of England in association with Anthony Trollope, Sir John Tenniel, the dukes of Bedford and Westminster, and other articulate critics of burial practices. Although it was not until 1884 that a British court first ruled cremation a legal procedure, it won immediate support on both sides of the Atlantic. For more than a century social and religious reformers had protested such practices as the long and sometimes disorderly wakes, or watches, held over the bodies of dead persons prior to burial. Physicians and sanitary engineers were alarmed over cemeteries, holding the saturated burial grounds to be poisonous to men and animals. The British movement prompted action in the United States, where the first crematorium was built in Washington, Pennsylvania, in 1876. Five years later the New York Cremation Society was organized; in 1913 the Cremation Association of America was organized. Growth of the practice has been slow in the United States, however, where by the 1970s only about 8 percent of the dead were cremated. It has been more readily accepted in some European and Asiatic countries: the figure in England, West Germany, and Denmark, for example, is more than 50 percent; it has been climbing steadily. In Japan, where cremation was illegal in 1875, the practice has become almost universal.

Revival of interest in cremation

As the shortage of cemetery space in urban areas becomes more acute and as objections are answered, cremation may become the principal form of burial. Many Protestant churches have actively supported it; the Roman Catholic Church has announced that it is not prohibited. The Orthodox Jewish religion, however, continues to declare it forbidden. Legal objections—that it would allow crimes to go undetected—are being countered by improvements in techniques and standards in coroners' offices. Cemetery owners and undertakers have also minimized their opposition, since cremation has proved to be no less profitable than more traditional methods of burial.

BURIAL SOCIETIES

In small, cohesive groups, where each member has social significance, it is the community's obligation, understood and accepted, to assure him a burial according to established custom. As communities grow in size, complexity, and wealth, however, so do the fears of its members that they will be denied a "proper" funeral. In the urban environment the ordinary individual—socially insignificant—worries that his departure will be unmourned and that what he understands to be a decent funeral may be beyond his means.

History. Religious organizations have guaranteed that some final attention will be paid; and membership in some sects also eliminates financial concerns, since the faithful

Religious and secular societies

are provided with a funeral and burial without charge. Early Christians conducted funerals as acts of charity: members of the community, under clerical supervision, prepared the body, assisted at the religious ceremony, and carried it in joyful procession to a resting place held as common property. Those early practices have persisted into the 20th century in many places, even in the United States, where funerals are highly commercialized. Funerals and burials of members of "Old Order" Amish communities are carried out in that manner. Burial committees connected with Friends Meetings in Ohio and some other states handle the routines without pay, including making the coffin and arranging a memorial service.

In many urban societies, whatever the dominant religion, people have organized mutual-aid burial societies. In China, as in other parts of the Orient, mutual-aid groups used to be very common, although the present government guarantees burial. In Western cultures, secular burial societies date back to ancient Rome, where they were established by workingmen to prevent their bodies' being consigned to the common burial pit with the corpses of slaves and aliens. The societies used members' contributions to build columbariums for the cremated remains and to pay for other funeral services.

Secular burial societies were revived during the Middle Ages, after simple Christian rites had given way to more elaborate and costly ceremonies. For small initial sums and nominal additional contributions when a member died, the societies paid for prayers and masses for the souls of the departed and provided the necessary paraphernalia for the ceremony—hearse, pall cloth, bier, and candles. The craft guilds, which played such an important role in medieval life, levied special assessments to pay funeral expenses of members and to provide them with postburial prayers. In time, funeral activities became so extensive that the guilds and burial clubs had to hire men to perform some of the services: bidders, who summoned others to a wake; bell ringers, who took to the streets on the day of the funeral and its first anniversary, urging hearers to pray for the dead soul. The Reformation did little to minimize funerary extravagance; although the amount spent for prayers and masses diminished, that spent for goods and services increased—particularly after coffins came into vogue. During the 16th and 17th centuries, mutual-aid societies for burial continued to flourish in large cities. The need for them increased as industrialization and urbanization accelerated, shattering the communal pattern of funerals and encouraging lavish expenditure. By the late 19th century it had reached incredible heights all across Europe, even among the poorest, as the novels of Dickens and Dostoyevsky indicate. Burial societies offered survivors the only hope against ruinous expenses. European and Oriental groups carried the idea to the United States, where the need had been intensified by the secularization of cemeteries and the rise of the undertaking business.

Increasing expenditures for funerals

Some societies offered limited and occasional benefits, like the Basque Society in Florida, which bought coffins and held them for use by members at the time of death. Membership in some fraternal organizations offered a place in the member-owned cemetery and a funeral with all the group in attendance. Some of the mutual-aid societies transplanted Old World burial practices to teeming cities in the New World. Some had a permanent impact on art and culture, including Negro societies in Southern cities: the membership fee included a marching band to the cemetery. Organizations with wider interests found themselves in the business of funeral relief. Some trade unions bought cemeteries so that members could be buried at low cost; consumer cooperatives in the Middle West, like those in England and Scandinavia, offered the full range of goods and services at a modest fee. Responsible insurance companies tried to minimize display. A study of funeral costs in 1928, financed by the Metropolitan Life Insurance Company, focussed national attention on the ways in which policyholders' survivors were being penalized by funeral extravagance.

Modern societies. Religious and consumer groups began to explore methods of obtaining simple and dignified funerals at reasonable cost. The modern funeral and memorial societies movement in the United States was launched in 1939 at two ends of the country. Acting independently, members of the Community Church in Brooklyn and the People's Church in Seattle worked out agreements with local undertakers enabling members to select in advance the kinds of funerals they preferred. (No payment was made until the service was rendered.) Since then more than 100 nonprofit societies have been organized in the United States and Canada, affiliated with the Continental Association of Funeral and Memorial Societies. Although originally many of them were sponsored by churches and ministerial associations, all are nondenominational and controlled by members, in accordance with the requirements of the Cooperative League of the USA, to which the Continental Association belongs. By 1971 so many societies had been established north of the border that the Memorial Society Association of Canada was organized. In addition to offering economic benefits to members—prices available for some services have been a small fraction of the national average—all of the societies conduct educational activities to inform people in their communities about the religious and financial values of simple funerals. Their rejection of extravagant display has won the movement the endorsement of religious leaders of all denominations and of persons involved in consumer education. The societies have encouraged members to donate bodies for medical research—a growing tendency in the United States.

BIBLIOGRAPHY. Comprehensive accounts for the general reader include: EFFIE BENDANN, *Death Customs: An Analytical Study of Burial Rites* (1930, reprinted 1969); BERTRAM S. PUCKLE, *Funeral Customs: Their Origin and Development* (1926, reprinted 1968); ROBERT W. HABENSTEIN and WILLIAM M. LAMERS, *Funeral Customs the World Over*, rev. ed. (1963); and BARBARA JONES, *Design for Death* (1967). More specialized but significant are ROBERT HERTZ, *La Représentation collective de la mort* and *La Prééminence da la main droite* (1928; Eng. trans., *Death and the Right Hand*, 1960); and JACK GOODY, *Death, Property, and the Ancestors* (1962). Hertz's study, which centers on the double burial practices of the Dayak of Borneo, and Goody's analysis of the mortuary institutions of the Dagari (Dagaba or LoDagaa) of West Africa are comparative sociological works of wide implication. Insights into the religious and cultural differences in burial and funeral practices are given in L.A. WADDELL, "Death and Disposal of the Dead," in the *Encyclopaedia of Religion and Ethics*, vol. 4, pp. 411–511 (1912); H.R. DAVIDSON, *The Road to Hel: A Study of the Conception of the Dead in Old Norse Literature* (1943, reprinted 1968); ALFRED C. RUSH, *Death and Burial in Christian Antiquity* (1941); and MAURICE LAMM, *The Jewish Way in Death and Mourning* (1969). Modern embalming techniques are described in CLARENCE G. STRUB and L.G. FREDERICK, *The Principles and Practices of Embalming*, 2nd ed. rev. (1959). American funeral and burial practices, a matter of continuing controversy, have been explored in a number of books. A defense of elaborate funerals is provided by ROBERT HABENSTEIN and WILLIAM M. LAMERS, *The History of American Funeral Directing*, rev. ed. (1962). Three books that take a different approach are LE ROY E. BOWMAN, *The American Funeral: A Study in Guilt, Extravagance, and Sublimity* (1959); JESSICA MITFORD, *The American Way of Death* (1963); and RUTH MULVEY HARMER, *The High Cost of Dying* (1963). ERNEST MORGAN, *A Manual of Simple Burial*, 5th ed. (1971), provides information about funeral and memorial societies in Canada and the United States.

(R.M.Ha.)

Embryology, Human

The human body, like that of most animals, develops from a single cell produced by the union of a male sex cell and a female sex cell. Human development follows closely the basic vertebrate pattern, and it departs only in certain details from the type specifically characteristic of mammals. A prenatal period, in which most of the developmental advances occur, is followed by a long postnatal period. Only at about the age of 25 are the last progressive changes completed. Each new individual starts as a fertilized, activated egg. Further progress depends upon cell proliferation, the growth of daughter cells, the differentiation of organs and their tissues, and the integration of the body and its parts into a coordinated, func-

tioning mechanism. The present article is concerned with the development that occurs before birth, and it is in this sense that the term embryology is used. Postnatal development is discussed in the article DEVELOPMENT, HUMAN.

Normal development

GAMETOGENESIS

Preparatory to development in the usual sense is the production of mature sex cells, male and female. The main steps in their development are similar, even though the two end products differ in size, form, and function. Primordial sex cells become recognizable as large, pale elements in embryos about three weeks old and 2.5 millimetres (0.1 inch) long. From a location near the hind end of the embryo, they migrate forward and colonize the region of the future sex gland. Here they multiply and become the source of all definitive sex cells. The general process of advance by which the primordial cells and their progeny become mature, functional elements, or gametes, is called gametogenesis. In the female the production of large, mature eggs is specifically named oogenesis. The comparable stages in producing motile spermatozoa in the male constitute spermatogenesis.

Origin of sex cells

In both sexes there is a period of proliferation that increases cell numbers. More important still is the fact that all sex cells that progress toward maturity undergo two successive cell divisions of a special nature. These reduce the chromosome number by half. Each human cell has two sets of chromosomes, each set numbering 23. One group was provided by the male parent of an individual and the other group by the female parent. The process by which the double set of chromosomes in a sex cell is reduced to a single set is called maturation.

Reduction is accomplished by means of a sequence of two modified cell divisions in which the chromosomes divide but once. This process, named meiosis, accomplishes several important results: (1) One benefit is to prevent a further increase in chromosome number (beyond the normal 46) when a male and a female gamete unite and produce a new individual. (2) Another result is to prepare for the transmission of hereditary characteristics received from both parents, since any random combination of chromosomes from the original parental sets may regroup in the daughter cells. (3) Still another objective, attained in spermatogenesis, is to segregate the male and the female type of sex-determining chromosome and pass only one of these into a single male gamete, thus preparing for later sex determination. (For additional details on meiosis, see CELL AND CELL DIVISION.)

Prenatal formation of ovarian follicles; cyclic activities after puberty

Oogenesis. In the later months of female fetal life, a capsule consisting of a single layer of ordinary cells surrounds each primordial sex cell, called an oogonium. The encased oogonium and its cellular capsule are designated as a primary ovarian follicle. At birth such follicles number several hundred thousand, and proliferation ends. There are no further advances until cyclic activities begin at puberty. Henceforth, early in each menstrual cycle, a number of primary follicles enter into a period of growth. The encapsulating follicle cells become multilayered, and the enlarging oogonium, enclosed in a gelatinous secretion called the zona pellucida, gains the name of primary oocyte. Meanwhile, the solid follicle acquires a fluid-filled cavity and the oocyte is crowded toward the periphery. By the middle of the menstrual month, all but one of the growing follicles have discontinued their advances and have begun to degenerate. At this time the single surviving follicle is a prominent sac ten millimetres (about 0.4 inch) or more in diameter. Shortly before it ruptures, the contained primary oocyte completes the first of its meiotic divisions.

An outstanding feature of oogenesis, or the production of an ovum, is that while the nuclei of the daughter cells are essentially identical, the allotments of cytoplasm, the cell substance outside the nucleus, are very unequal. In the first division one daughter cell retains almost all of the cytoplasm and is then called a secondary oocyte. The other daughter cell is a tiny, nonfunctional element known as the first polar body. Each daughter cell, however, contains the reduced number of 23 chromosomes, since in this division each mated chromosome pair separates and only one of its components passes into a daughter cell. The final maturation division takes place subsequent to ovulation, and then only in the event of fertilization. During penetration of the spermatozoon into the cytoplasm of the secondary oocyte, the second polar body is pinched off in a meiotic division that is much like ordinary cell division, called mitosis. The first polar body may or may not subdivide, but theoretically the maturative process ends with the production of a ripe ovum (technically called an ootid) and three tiny, abortive polar bodies. If fertilization fails to occur, the secondary oocyte dies as such. Compared to ordinary cells of the body, the ripe egg is truly large (0.12 millimetre, or about 0.005 inch), being just visible to the naked eye. The production of a single egg instead of millions, as by some lower vertebrates, is correlated with greater surety of fertilization and with protected development within the mother's body.

Formation of sperm cell

Spermatogenesis. The primary sex cells, or spermatogonia, of the fetal testis remain dormant within the testis tubules until the time of puberty. They then enter into proliferative activity that continues indefinitely into adult life. Some daughter spermatogonia continue as mitotic stem cells (unspecialized cells from which more specialized cells develop). Others grow into primary spermatocytes and proceed through the two meiotic divisions that first produce secondary spermatocytes and then spermatids, the immediate precursors of spermatozoa. All of these final cells possess the reduced set of 23 chromosomes. The chromosomal behaviour parallels that occurring in the oocytes.

Spermatogenesis differs from oogenesis in several important respects. The spermatids are all equal in size and are potential spermatozoa but must first undergo a radical alteration in shape and organization. This process of transformation into atypical, motile cells is called spermiogenesis. Moreover, spermatids are of two kinds, half carrying the X-chromosome and half the Y-chromosome as the result of the reductional segregation of the unmatched X–Y pair of sex chromosomes. (By contrast, all ootids carry the X-chromosome.)

Each spermatid remodels drastically into a tadpole-shaped element with a relatively large head, a slender neck, and an extremely long, vibratile tail. The transforming spermatids bury their heads in the cytoplasm of certain tall, radially arranged supporting cells and possibly receive nourishment from them. When complete, the young spermatozoa become detached and are transported passively through the smaller ducts of the testis to the convoluted, or tortuous, epididymis. Here they are stored until finally discharged from the body. A spermatozoon is 0.06 millimetre (0.002 inch) long, which is one-half the diameter of the ovum. Yet in volume the ovum is 85,000 times as great. Both gametes differ only in matters of detail from those of other vertebrates, and most invertebrates.

OVULATION AND SEMINATION

Release of egg and discharge of sperm

Ovulation. Soon after the first polar body is shed, the follicle ruptures. Its growth is rapid in the final hours, but bursting involves a poorly understood weakening of the superficial follicular wall. The collapse tears the secondary oocyte away from the already loosened bed of follicle cells in which it has lain and also expels the oocyte on a wave of escaping follicular fluid. Both oocyte and its associated first polar body remain encased within the gelatinous zona pellucida, to which follicle cells will still adhere for a time. These several events make up the process of ovulation. The liberated oocyte is promptly taken up into the open end of the uterine tube and then is transported downward. The fertilizable life of the free oocyte is probably less than 24 hours.

Semination. Discharged seminal fluid contains some 250,000,000 spermatozoa, now motile by lashings of the tail. Forward progress is about two millimetres per minute (a little less than five inches per hour). When deposited in the vagina the sperms ascend through the uterus

and into the uterine tubes, mainly as the result of muscular movements in those organs rather than through swimming ability of the sperms. Apparently in about one hour some sperms reach the upper ends of the tubes. The total number attaining this level never increases greatly and remains small in comparison to the enormous number initially discharged.

FERTILIZATION

Penetration of the ovum by the sperm and their union

The development and liberation of the male and female gametes are steps preparatory to their union through the process of fertilization. Random movements first bring some spermatozoa into contact with follicle cells adhering to the secondary oocyte, which still lies high in the uterine tube. The sperms then propel themselves past the follicle cells and attach to the surface of the gelatinous zona pellucida. Some sperm heads successfully penetrate this capsule by means of an enzyme, hyaluronidase, which they secrete, but only one makes contact with the cell membrane and cytoplasm of the oocyte and proceeds farther. This is because the invading sperm head releases a substance that initiates surface changes in the oocyte cytoplasm that other competitors cannot master.

The sole successful sperm is engulfed by a conical protrusion of the oocyte cytoplasm and is drawn inward, tail and all. Once within the periphery of the oocyte, the sperm advances toward the centre of the cytoplasm. During this journey the head swells and converts into a typical nucleus, now called the male pronucleus, and the tail detaches. It is during the progress of these events that the oocyte initiates its final maturation division. Following the separation of the second polar body, the oocyte nucleus reconstitutes typically and is then called the female pronucleus of the ripe egg. It is now ready to unite with its male counterpart and thereby consummate the total events of fertilization.

The two pronuclei next approach, meet midway in the egg cytoplasm, and lose their nuclear membranes. Each resolves its diffuse chromatin material into a complete, single set of 23 chromosomes. Centrioles (complex particles involved in cell division), apparently supplied by the sperm, appear, and a mitotic spindle organizes with the two sets of chromosomes arranged midway on it—ready to proceed with a typical mitosis. This climax in the events of fertilization creates a joint product named the zygote. It contains all the essential factors for the development of a new individual.

The fundamental results of fertilization are the following: (1) reassociation of a male and female set of chromosomes (thus restoring the full number and providing the basis for biparental inheritance and for variation); (2) establishment of the mechanism of sex determination for the new individual (this depending on whether the male set of chromosomes included the X- or the Y-chromosome); (3) activation of the zygote, initiating a beginning toward the production of a new individual.

CLEAVAGE AND BLASTULATION

Cleavage. The onset of mitosis (ordinary cell proliferation by division) in the activated zygote is a first step toward development in the ordinary sense of that term, and the cells so produced are the first external sign of future body building. To this end, the relatively enormous zygote straightway subdivides into many smaller cells of conventional size, suitable as early building units for the future organism. This process is called cleavage and the resulting cells are blastomeres (Figures 1A–D).

Formation of blastomeres

The tendency for the progressive increase in cell numbers to follow a doubling sequence is soon disturbed and then lost. Each blastomere receives the full complement of paternal and maternal chromosomes.

Subdivision of the zygote into blastomeres begins while it is still high in the uterine tube. The cohering blastomeres are transported downward chiefly, at least, by muscular contractions of the tubal wall. Such transport is relatively rapid until the lower end of the tube is reached, and here cleavage continues for about two days before the multicellular cluster is expelled into the uterus. The full reason for this delay is not clear, but it serves to retain the cleaving blastomeres until the uterine lining is suitably prepared to receive its prospective guest.

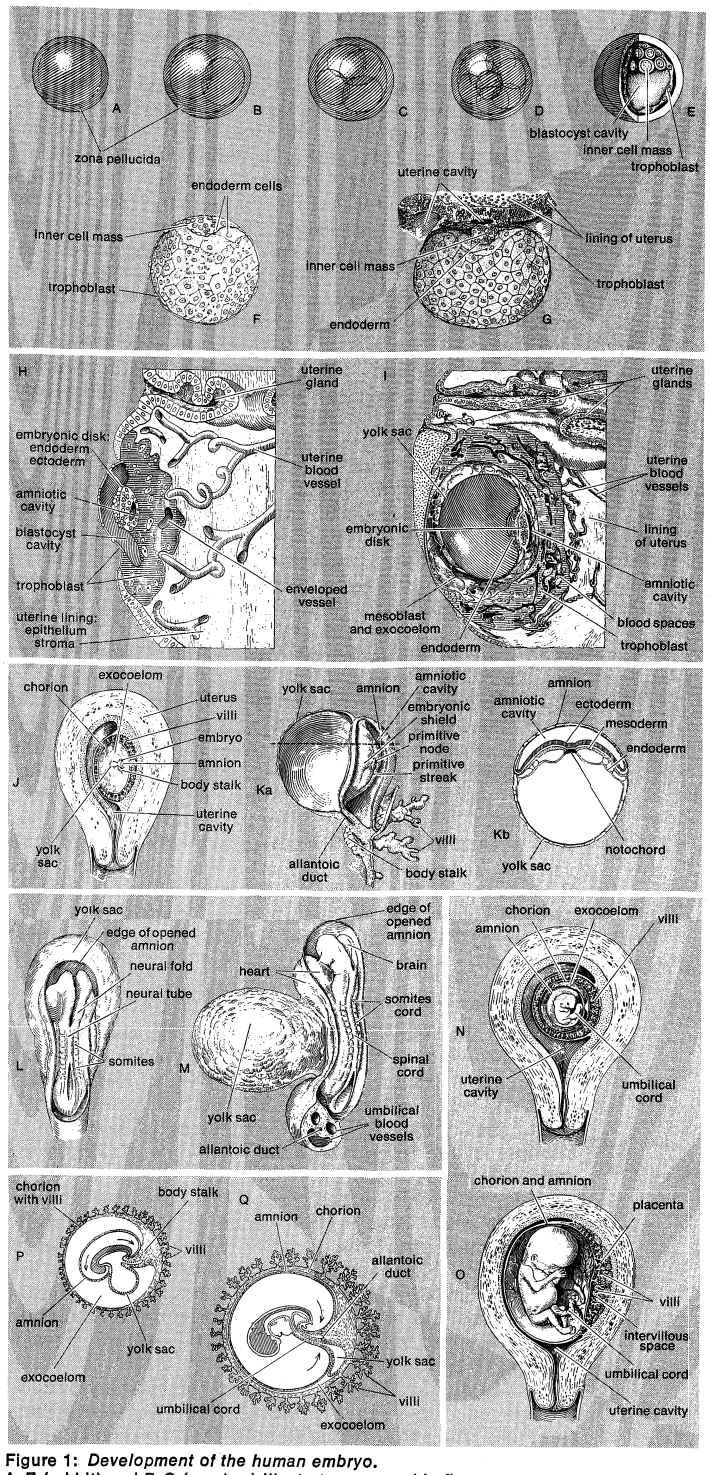


Figure 1: *Development of the human embryo.*
A–E (rabbit) and F–G (monkey) illustrate comparable first stages of human development. (A–D) Cleavage of ovum (magnified 80 ×). (E) Early free blastocyst, partially opened (magnified 80 ×). (F) Older free blastocyst, halved (magnified 60 ×). (G) Late attached blastocyst, halved (magnified 70 ×). (H–J) Half sections of embryos implanted in uterine lining: (H) at 7½ days (magnified 80 ×); (I) at 13 days (magnified 37 ×); (J) at 23 days (magnified ¼ ×). (Ka) Embryo of 18 days at disk (or shield) stage, three-quarter view (magnified 9 ×). (Kb) Cross section through same embryo at level of broken line in Ka. (L) Embryo of 21 days with amnion opened, back view (magnified 8 ×). (M) Embryo of 23 days with yolk sac and opened amnion (magnified 9 ×). (N,O) Embryo of six weeks and fetus of three months, respectively, within halved amnion, chorion, and uterus (magnified ¼ ×). (P,Q) Diagrams showing especially amnion growth and formation of umbilical cord; stages correspond to J and N.
Drawing by James F. Didusch

Since the human egg contains little inert yolk material, and since this is distributed rather evenly throughout the cytoplasm, the daughter cells of each mitosis are practically equal in size and composition. This type of cleavage is known as total, equal cleavage. The sticky blastomeres adhere and the cluster is still retained for a time within the gelatinous capsule—the zona pellucida—that had enclosed the growing and ovulated oocyte. There is no growth in the rapidly dividing blastomeres, so that the total mass of living substance does not increase during the cleavage period. Cleavage is not a mechanism designed primarily to distribute particular developmental qualities, either nuclear or cytoplasmic, to specific blastomeres, which then carry out assignments that give rise to particular parts of the embryo. In a sense, therefore, cleavage is essentially a nonspecific fractionating preliminary rather than a process of truly constructive development.

Passage of blastomeres into uterus; formation of morula and blastocyst

Morula and blastocyst. By the fourth day after ovulation, a cluster of about 12 blastomeres passes from the uterine tube into the uterus. When the cluster numbers 12 to 16 blastomeres it is called a morula (Figure 1D). By the time some 30 blastomeres have been produced, pools of clear fluid accumulate between some of the internal cells, and these spaces soon coalesce into a common subcentral cavity. The resulting hollow cellular ball is a blastula of a particular type that occurs in mammals and is called a blastocyst; its cavity is the blastocoel (Figure 1E, F).

An internal cellular cluster, eccentric in position and now named the inner cell mass, will develop into the embryo. The external capsule of smaller cells, enveloping the segregated internal cluster, constitutes the trophoblast. It will contribute importantly to the formation of a placenta. During its stay within the uterine cavity, the blastocyst loses its gelatinous capsule, imbibes fluid, and expands to a diameter of 0.2 millimetre (0.008 inch); this is nearly twice the diameter of the zygote at the start of cleavage. Probably several hundred blastomeres have formed before the blastocyst attaches to the uterine lining.

IMPLANTATION AND PLACENTATION

Late in the sixth day after ovulation the naked, sticky blastocyst comes into contact with the uterine lining and adheres to it (Figure 1G). The site of attachment is variable and not predetermined. The uterine lining has already been preparing, under the influence of ovarian hormones, for the reception of the blastocyst. Among these preparations has been the elaboration and expulsion, by the uterine glands, of a secretion that serves as nourishment for the blastocyst, both when it is free and during its implantation. Directly after blastocyst attachment come its establishment within the thickened uterine lining and the participation of its trophoblastic capsule in the differentiation of a placenta, a structure that enables the developing embryo to enter into a prompt, physiological dependence on the mother.

Implantation. The trophoblast of the blastocyst exerts an enzymic, destructive influence on the swollen uterine lining, leading to erosion of both the superficial epithelium of the uterine lining and also its deeper, cellular connective tissue (Figure 1H). This early stage of invasion ends in a few days; the blastocyst is then completely buried within a more superficial and compact layer of the total uterine lining (Figure 1H, I). While the blastocyst is completing this phase of implantation, its original shell of cellular trophoblast steadily proliferates a multitude of cells that lose their outermost membranes and merge. The result is a thick peripheral layer consisting of a common mass of cytoplasm in which many nuclei are embedded. This external investment is called syncytial trophoblast.

The implanted blastocyst next proceeds to establish itself as a parasite within the uterus. The syncytial trophoblast becomes a spongy shell containing irregular cavities (Figure 1I). This expanding mass destroys cellular connective tissue and glands encountered in its path. Both the cellular and derivative syncytial trophoblast have the capacity of ingesting such tissue. The erosive process also taps capillaries and small blood vessels, and blood liberated in this way is likewise taken up into the trophoblast. Such engulfed blood cells and fragments of tissue are presumably utilized for nourishment in the early period of establishment of the expanding blastocystic sac. Yet these erosive activities decline in intensity by the end of the third week of development, and at this time the sac is completing the first phase of its specialization.

Syncytial trophoblast; its formation and effects

Occasionally a fertilized egg fails to reach the uterus, implanting and beginning to develop elsewhere. This outcome is called an ectopic or extra-uterine pregnancy. The commonest ectopic site is the uterine tube, but the peritoneum lining the abdominal cavity and even the interior of the ovary are also involved, though rarely. The unsuitability of all these sites for continued development usually leads to early death and resorption of the embryo.

Placentation. The irregular strands of invasive syncytial trophoblast constitute a first stage in the formation of true villi (which form part of the placenta and are briefly described below). Primitive connective tissue soon lines the interior of the blastocyst wall, and this complex (trophoblast and connective tissue) is then named the chorion. Connective tissue promptly grows into the trophoblastic strands, and blood vessels develop in the tissue. The result is the production of many chorionic villi, each resembling a tiny, branching bush (Figure 1J, Ka).

In the fourth week of development, the essential arrangements have been established that make possible those physiological exchanges between mother and fetus that characterize the remainder of pregnancy. The deepest embedded portion of the chorionic wall becomes the so-called chorionic plate of the developing placenta. From the plate extend the main stems of chorionic villi, which give off progressively subdividing branches (Figure 1N, O). In general, the side branches are free, whereas apical ones tend to attach to the maternal tissue and serve as anchors. The villous trees occupy a labyrinthine space between the villi that was created by erosion of the uterine lining. Trophoblast not only covers the chorionic plate and its villi but also spreads like a carpet over the eroded surface of the maternal tissue.

Tapped uterine arteries open into the trophoblast-lined intervillous space, their blood bathing the branches and twigs of the villous trees as it circulates among them. This blood drains from the intervillous space through similarly tapped veins. Arterial blood of the embryo, and later fetus, passes through vessels of the umbilical cord to the chorionic plate. Thence it is distributed to the villous stems, branches, and twigs through vessels in their connective-tissue cores. Return of this blood to the fetus is by a reverse route.

The circulation of maternal blood through the intervillous space is wholly separate from fetal blood coursing through the chorion and its villi. Communication between the two is solely by diffusive interchange. The barrier between the two circulations consists of the trophoblastic covering of villi, the connective tissue of the villous cores, and the thin lining of the capillaries that are contained in the villous cores. The placenta serves the fetus in several ways, most of which involve interchanges of materials carried in the bloodstreams of the mother and fetus. These functions are of the following kinds: (1) nutrition, (2) respiration, (3) excretion, (4) barrier action (*e.g.*, prevention of intrusions by bacteria), and (5) synthesis of hormones, certain foodstuffs, and enzymes. For some functional details and for a description of the mature, expelled placenta, see PREGNANCY.

Functions of the placenta

EXTRA-EMBRYONIC MEMBRANES

The way in which the encapsulating membrane of the blastocyst becomes the chorion, and the most deeply embedded part of it becomes the fetal placenta, has already been described. There are still other important extra-embryonic membranes that develop from those portions of the inner cell mass of the blastocyst that are not directly involved in becoming an embryo.

Yolk sac. Cells split off from the inner cell mass of the blastocyst and fashion themselves into a primitive yolk

sac (Figure 1J, Ka). The roof of the sac then folds into a tubular gut, whereas the remainder becomes a vascularized bag that attains the size of a small pea (Figure 1M). In other vertebrates, such as amphibians and birds, the yolk sac is large and contains a store of nutritive yolk. But in man and other true mammals there is practically none. A slender neck, the yolk stalk, soon connects the rapidly elongating gut with the fast growing yolk sac proper. The stalk detaches from the intestine early in the second month, but the shrunken sac commonly persists and can be found in the expelled afterbirth.

Amnion. A cleft separates the outermost cells of the inner cell mass of the blastocyst from the remainder, which then becomes the embryonic disk (Figure 1H). The split-off, thin upper layer is the amnion, which remains attached to the periphery of the embryonic disk. As the disk folds into a cylindrical embryo, the amniotic margin follows the underfolding, and its line of union becomes limited to the ventral (frontward) body wall, where the umbilical cord attaches (Figure 1P, Q). The amnion becomes a tough, transparent, nonvascular membrane that gradually fills the chorionic sac and then fuses with it (Figure 1N, O). At the end of the third month of pregnancy, the nonplacental extent of this nearly exposed double membrane comes into contact with the lining of the uterus elsewhere. Fusion then obliterates the uterine cavity, which has been undergoing progressive reduction in size. For the remainder of pregnancy the only cavity within the uterus is that of the fluid-filled amniotic sac.

Clear, watery fluid fills the amniotic sac. The embryo is suspended in this fluid and thus can maintain its shape and mold its body form without hindrance. Throughout pregnancy the amniotic sac serves as a water cushion, absorbing jolts, equalizing pressures, and permitting the fetus to change posture. At childbirth it acts as a fluid wedge that helps dilate the neck of the uterus. When the sac ruptures, about a quart of fluid escapes as the "waters." If the sac does not rupture or if it covers the head at birth, it is known as a caul.

Allantois. The allantois, a tube of endoderm (the innermost germ layer), grows out of the early yolk sac in a region that soon becomes the hindgut. The tube extends into a bridge of mesoderm (the middle germ layer) that connects embryo with chorion and will become incorporated into the umbilical cord (Figure 1P, Q). The human allantoic tube is tiny and never becomes a large sac with important functions, as it does in reptiles, birds, and many other mammals. In the second month it ceases to grow, and it soon is obliterated. Blood vessels, however, develop early in its mesodermal sheath, and these spread into the chorion and vascularize it. Throughout pregnancy they will keep the embryo in close relationship with the mother's uterine circulation.

Umbilical cord. As the ventral body wall closes in, the yolk stalk and allantois are brought together, along with their mesodermal sheaths and blood vessels (Figure 1P, Q). Enclosing everything is a wrapping of amnion. In this manner a cylindrical structure, the umbilical cord, comes to connect the embryo with the placenta (Figure 1N, O). It will serve the embryo and fetus as a physiological lifeline throughout the period of pregnancy. The mature cord is about 1.3 centimetres (0.5 inch) in diameter, and it attains an average length of nearly 50 centimetres (two feet).

GASTRULATION

Formation of the three primary germ layers

The inner cell mass, attached to the deep pole of the implanted blastocyst, is sometimes called the embryoblast, since it supplies the materials used in the formation of an embryo. The cellular mass flattens and enters into the process of gastrulation, through which the three primary germ layers segregate and the gastrula stage, the next advance after the blastula, begins to take form. First, cells facing the cavity of the blastocyst arrange into a layer named the endoderm (Figure 1H, I). The thick residual layer, temporarily designated as epiblast, is the source of a definitive uppermost sheet, the ectoderm, and an intermediate layer, the mesoderm. In this second phase of gastrulation, some cells of the epiblast migrate to the midline position, then turn downward and emerge beneath as mesoderm. Such cells continue to spread laterally, right and left, between the endoderm and the residue of epiblast, which is now definitive ectoderm (Figure 1Kb).

The site where the migratory mesodermal cells leave the epiblast is an elongated, crowded seam known as the primitive streak (Figure 1Ka). Similar migrating cells produce a thick knob at one end of the primitive streak. Their continued forward movement from this so-called primitive knot produces a dense band that becomes the rodlike notochord.

The germ layers are not segregated sheets whose cells have predetermined, limited capacities and inflexibly fixed fates in carrying out organ-building activities. Rather, the layers represent advantageously located assembly grounds out of which the component parts of the embryo emerge normally, according to a master constructional plan that assigns different parts to definite spatial positions and local sites. Thus, although the germ layers have developmental potencies in excess of their normal developmental fates, their ordinary participation in organ forming does not deviate from a definite, standard program.

The derivatives of the primary germ layers can best be presented in tabular form. In naming the germ-layer origin of an organ, only the principal functional tissue is designated. In a few instances, such as the suprarenal (adrenal) glands and the teeth, a compound organ has important parts of different origin.

Table 1: Derivatives of Primary Germ Layers

ectoderm	mesoderm	endoderm
epidermis	epithelium of:	epithelium of:
cutaneous derivatives	circulatory system	pharynx
epithelium of:	spleen; lymph nodes	thyroid; thymus
mouth; oral glands	urogenital system	parathyroid
nasal passages	body cavities	digestive tube;
sense passages	connective tissues;	liver; pancreas
central nervous system	blood; bone marrow	larynx; trachea; lungs
peripheral nervous system	muscular tissues	urinary bladder; urethra
hypophysis;	skeletal tissues	vestibule; vagina
suprarenal medulla	suprarenal cortex	

GROWTH AND DIFFERENTIATION

Growth. Growth is an increase in size, or bulk. Cell multiplication is fundamental to an increase in bulk but does not, by itself, result in growth. It merely produces more units to participate in subsequent growing. Growth is accomplished in several ways: (1) Most important is synthesis, by which new living matter, protoplasm, is created from available foodstuffs. (2) Another method utilizes water uptake; a human embryo of the early weeks is nearly 98 percent water, while an adult is 70 percent fluid. (3) A third method of growth is by intercellular deposition; cells manufacture and extrude such nonliving substances as jelly, fibres, and the ground substance of cartilage and bone. Through these activities a newborn baby is several thousand million times heavier than the zygote from which it came.

Ways in which growth is accomplished; abnormal growth patterns

It is obvious that uniform growth throughout the substance of a developing organism would merely produce a spherical cellular mass. Local diversities in form and proportions result from differential rates of growth that operate in different regions and at different times. The particular program of starting times and growth rates, both externally and internally in the human embryo, constitutes its characteristic growth pattern. Abnormal growth occurs occasionally, and growth may be excessive or deficient. Also, such departures may be general or local, symmetrical or asymmetrical. General gigantism usually starts before birth, and the oversized baby continues to grow at an accelerated rate. On the other hand, an existing hereditary predisposition may not be aroused into action until some time during childhood. In a reverse manner, general dwarfism may exist before birth, the individual continuing to grow slowly and stopping at the usual time. Another type is normal in size at birth,

grows normally for a while, and then comes to a premature arrest. See also DEVELOPMENT, HUMAN; BIRTH DEFECTS AND CONGENITAL DISORDERS.

Differentiation. In a developing organism, differentiation implies increasing structural and chemical complexity. One kind of differentiation concerns changes in gross shape and organization. Such activities, related to molding the body and its integral parts into form and pattern, comprise the processes called morphogenesis. The processes of morphogenesis are relatively simple mechanical acts: (1) cell migration; (2) cell aggregation, forming masses, cords, and sheets; (3) localized growth or retardation, resulting in enlargements or constrictions; (4) fusion; (5) splitting, including separation of single sheets into separate layers, formation of cavities in cell masses, and forking of cords; (6) folding, including circumscribed folds that produce inpocketings and outpocketings; (7) bending, which, like folding, results from unequal growth.

Morphogenesis and histogenesis

A second kind of differentiation refers to progressive changes occurring in the substance and structure of cells, whereby different kinds of tissues are created. These changes, and the synthetic processes underlying them, constitute histogenesis. The zygote contains all the essential factors for development, but they exist solely as an encoded set of instructions localized in the genes of chromosomes and bearing no direct physical relationship to the future characteristics of the developing embryo. During histogenesis these instructional blueprints are decoded and transformed, through cytoplasmic syntheses, into the several types and subtypes of tissues that are the structural and functional units of organs. At first, the cells of each germ layer lack an identifiable shape and are similar in chemical composition, but an invisible type of chemical differentiation soon enters. After the elaboration of specific enzyme patterns and syntheses, certain groups of cells progressively assume distinctive characters that permit their fates to be recognized. Such early stages in definite lines of differentiation of cells are often designated by the suffix "-blast" (*e.g.*, myoblast; neuroblast).

The emerging cell types are discrete entities, without intermediates; for example, a transitional form between a muscle cell and a nerve cell is never seen. Neither can different, local parts of a cell carry out different types of tissue specialization, such as nerve at one end and muscle at the other end; nor can a cell, once fully committed to a particular type of specialization, abandon it and adopt a new course.

"Dedifferentiation"; the four types of tissue

Under certain conditions, differentiated cells may reverse their course and return to a simpler state. Thus, under a changed environment, cartilage may lose its matrix, and its cells may come to resemble the more primitive tissue from which it arose. Nevertheless, despite such reversal and apparent simplification ("dedifferentiation"), these cells retain their former histological specificity. Under suitable environmental conditions they can differentiate again but can only regain their previous definitive characteristics as cartilage cells.

The final result of histogenesis is the production of groups of cells similar in structure and function. Each specialized group constitutes a fundamental tissue. There are four main types of such tissues: (1) Each of the three germ layers gives rise to sheetlike epithelia, which cover surfaces, line cavities, and are frequently glandular. (2) Ectoderm also forms the nervous tissues. (3) Mesoderm also produces the muscular tissues. (4) It also differentiates into blood and the fibrous connective tissues (including two further specialized types, cartilage and bone).

THE ACQUISITION OF EXTERNAL FORM

At the end of the second week, the embryonic region is a nearly circular plate within its well-embedded, differentiating chorionic sac (Figure 1I). This embryonic disk consists of two layers (epiblast and endoderm), with a third layer (mesoderm) just starting to spread between them. A hollow, dome-shaped amnion sac attaches to the margin of the upper layer of the disk, and a hollow yolk sac is similarly continuous with the lower layer. A broad cellular bridge attaches the complex to the chorion.

Early in the third week the embryonic disk has enlarged and become pear-shaped in outline, and a well-formed primitive streak occupies the midline of its caudal (hind) half, which is narrower. Cells from the epiblast are passing through the streak and spreading laterally in both directions beneath the uppermost layer, now ectoderm (Figure 1Ka, Kb). In this way the embryonic disk becomes three layered, and the gastrula stage of development comes to an end. At the middle of the week a thickening, the head process, is extending forward from a knoblike primitive knot located at the head end of the primitive streak. These linear thickenings define the median plane of the future embryo and thus divide the embryonic disk into precise right and left halves.

Toward the end of the week the disk elongates and becomes slipper shaped in outline; a slight constriction demarcates it from the attached yolk sac. Growth has lengthened the region ahead of the now receding primitive streak. Here, in the midline, the ectoderm bears a definite gutter-like formation called the neural groove; it is the first indication of the future central nervous system. Beneath the groove the mesodermal head process presently rounds into an axial rod, the notochord, that serves as a temporary "backbone." By the end of the third week a head fold, paired lateral body folds, and a tail fold become prominent, demarcating a somewhat cylindrical embryo from the still broadly attached yolk sac (Figure 1L). The neural folds, flanking the neural groove, converge and begin to meet midway of their lengths, thereby producing a neural tube at that level. Mesoderm, alongside the notochord, begins to subdivide into paired blocks called somites, and the outlines of the somites show externally. From them, muscles and vertebrae will differentiate later. This stage, when the embryo is fashioning a neural tube, is often designated as a neurula.

In the fourth week the embryo goes beyond the external characteristics of vertebrates in general and becomes recognizable as a mammal. The week is marked by profound changes, during which the embryo acquires its general body plan. There is an increase in total length from about two to five millimetres, but size is quite variable among smaller specimens. Better correlated with the degree of development is the number of mesodermal somites, which attain their full number of about 42 during the fourth week. Some of the head of an early embryo arises from the embryonic disk in front of the primitive knot. But as the primitive streak shortens and its caudal retreat continues, such structures as the neural tube and notochord are added progressively in the wake of that retreat, and additional somite pairs also appear in steady succession.

Factors in formation of early embryo

The most important manoeuvre in the establishment of general body form is the transformation of the flat embryonic disk into a roughly cylindrical early embryo, which is attached to the yolk sac by a slender yolk stalk (Figure 1M). Three factors cooperate in producing this change: (1) There is more rapid expansion of both the embryonic area and the yolk sac than in the region joining the two. The enlarging embryonic area at first buckles upward and then overlaps the more slowly growing margin. Since growth is particularly rapid at the future head end and tail end, the embryo becomes elongate. (2) In conjunction with this overgrowth there is important underfolding, again most pronounced at the front and hind ends. Underfolding is produced by differential elongation in the regions of the brain and tail bud. Conspicuous is the change in the future cardiac (heart) and foregut region, which swings beneath the brain as on a hinge. (3) A certain amount of true constriction, through growth, purses all of these parts at the site of the future umbilicus.

Throughout the entire period when the body and its parts are being laid down, developmental advances tend to appear first at the head end and then progress tailward. For this reason, many structures that extend along the body for a distance show a gradation in development. The size advantage gained initially by the head end of the embryo is relinquished very slowly. Even in an infant the relatively large head and long arms are striking. A further tendency toward progressively graded development

occurs from the middorsal line in a lateral (sideward) and then ventral (frontward) direction. All such relations are the visible expressions of stages in growth and differentiation.

Early in the fourth week the cylindrical shape of the embryo is plain, even though the folding-off process is far from complete (Figure 1M). The neural tube is still open near both ends, and at the head end the broader neural folds indicate the future brain and even its three primary divisions. A pronounced bulge beneath the brain region denotes where the heart is forming precociously in order to institute a necessary, prompt circulation through the placenta.

During the middle and late days of the fourth week there are marked advances (Figure 2A). Accelerated growth along the dorsal region bends the total body length progressively until the embryo assumes a striking C-shape, with the tips of the head and tail not far apart. Continued growth and underfolding close in much of the ventral side of the embryo, so that a free head and upper trunk, and a lower trunk and a prominent tail, are easily recognizable. Forebrain, midbrain, and hindbrain can be identified, largely because of a sharp bend in the midbrain. Local outgrowths from each side of the forebrain produce stalked eye cups, and a pair of inpocketings of the ectoderm alongside the hindbrain sink beneath the surface as otic vesicles, forerunners of the inner ears. Bulges indicative of the heart and liver are prominent. Formations called branchial arches, reminiscent of the gill arches of fishes and aquatic amphibians, become conspicuous in the future jaw–neck region. Paired swellings ("buds") off the trunk foretell the locations of the upper and lower limbs.

Changes in the fifth to eighth weeks

A five-week embryo is about eight millimetres (about ⅓ inch) long (Figure 2B, C), whereas at six weeks the length is about 13 millimetres (Figure 2D). New external features are olfactory pits at the tips of the bent head. An umbilical cord becomes a definite entity, its proximal end occupying a low position on the abdominal wall. The sharply bent head joins the rest of the body at an acute angle. The first pair of branchial arches branch Y-fashion into maxillary and mandibular processes (primitive upper and lower jaws). The external ears are forming around the paired grooves located between each half of the

By courtesy of the Carnegie Institution of Washington

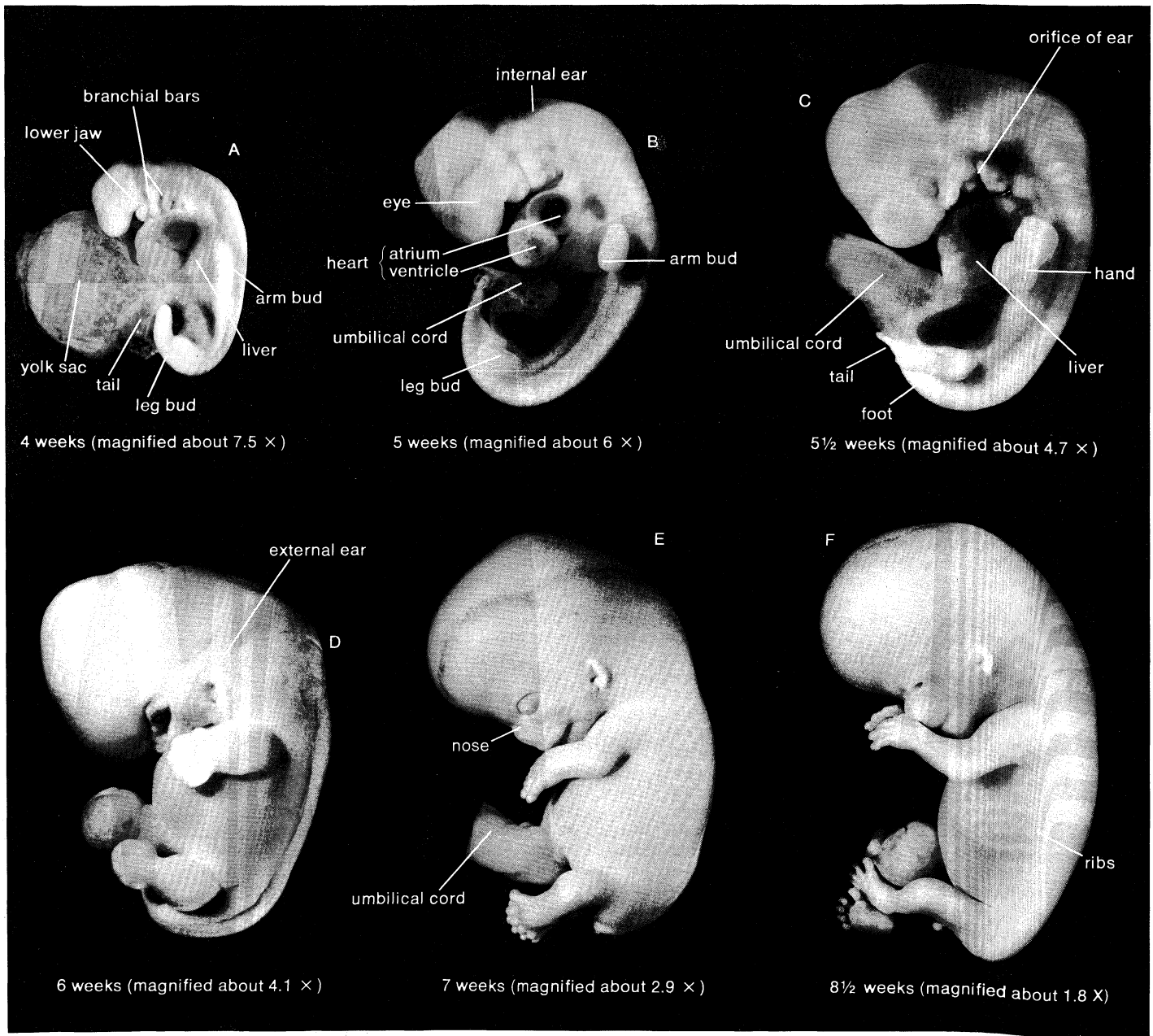


Figure 2: Human embryos from four to 8½ weeks.

mandible and each second branchial arch. The heart, which was previously the chief ventral prominence, now shares this distinction with the rapidly growing liver. Limb buds have elongated markedly and become flattened at their outer ends. A constriction on each bud separates a paddle-like hand plate or foot plate from a cylindrical segment attached to the body wall. Predictably, the upper limbs are somewhat further advanced than the lower pair.

In the latter weeks of the second month, developmental changes advance from those that distinguish primates to a state that is recognizably human (Figure 2E). At the end of the second month (30 millimetres', or 1.2 inches', length) the stage of the embryo ends; henceforth, until birth, "fetus" is the preferred term.

In the seventh and eighth weeks the head becomes more erect and the previously curved trunk becomes straighter (Figure 2E, F). The heart and liver, which earlier dominated the shape of the ventral body, yield to a more evenly rounded chest–abdomen region. The tail, which at an earlier time was one fifth of the embryo's length, becomes inconspicuous both through actual regression and through concealment by the growing buttocks. The face rapidly acquires a fairly human appearance; eyes, ears, and jaws are prominent. The eyes, previously located on the sides of the head, become directed forward; the nose lacks a bridge and so is of the "pug" type, with the nostrils directed forward instead of downward. A mandibular branch of each Y-shaped branchial arch combines with its mate to form the lower jaw. The maxillary branch on each side joins an elevation on the medial (inner) side of the corresponding nostril to produce the more complicated upper jaw. Branchial arches, other than those forming the jaws and external ears, are effaced through incorporation into an emerging, recognizable neck. Limbs become jointed, and the earlier hand plates and foot plates differentiate terminal digits. Primitive external genitalia appear, but in a nondistinctive, sexless condition.

Almost all of the internal organs are well laid down at the end of eight weeks, when the embryo is little more than 25 millimetres (one inch) long. The characteristic external features are established, and subsequent growth merely modifies existing proportions without adding new structure. Similarly, the chief changes undergone by internal organs and parts are those of growth and tissue specialization. At eight weeks the neuromuscular mechanism attains a degree of perfection that permits some response to delicate stimulation.

Development from the third month to birth

During the third month the young fetus clearly resembles a human being, although the head is disproportionately large (Figure 1O). The previous protrusion of much of the intestine into the umbilical cord is reduced through the return of its loops into the abdomen. The ears rise to eye level and the eyelids fuse shut. Nails begin forming; ossification (bone-forming) centres appear in most of the future bones; and the sex of external genitalia becomes recognizable. (In this paragraph, and in the next two, the months are lunar months, of 28 days).

At four months individual differences between the faces of fetuses become distinguishable. The face is broad but the eyes are now less widely separated. The umbilical cord attaches higher on the abdominal wall; this location is above an expanding region between the cord and the pubis (front bones of the pelvis) that scarcely existed previously.

At five months downy hairs (lanugo) cover the body, and some head hairs appear. The skin is less transparent. Fetal movements ("quickening") are felt by the mother. At six months eyebrows and eyelashes are clearly present. The body is lean, but its proportions have improved. The skin is wrinkled. At seven months the fetus resembles a dried-up old person. Its reddish, wrinkled skin is smeared with a greasy substance (vernix caseosa). The eyelids reopen. At eight months fat is depositing beneath the skin. The testes begin to invade the scrotum. At nine months the dull redness of the skin fades, and wrinkles smooth out. The body and limbs become better rounded.

At full term (38 weeks) the body is plump and proportions are improved, although the head is large and the lower limbs are still slightly shorter than the upper limbs. The skin has lost its coat of lanugo hair, but it is still smeared with vernix caseosa. Nails project beyond the finger tips and to the tips of toes. The umbilical cord now attaches to the centre of the abdomen. The testes of males are usually in the scrotum; the greater lips of the female external genitalia, which previously gaped, are now in contact. Cranial bones meet except at some angular junctions, or "soft spots."

The average time of delivery is 280 days from the beginning of the last menstrual period, whereas the duration of pregnancy (age of the baby) is about 266 days (38 weeks). Pregnancy may extend to 300 days, or even more, in which case the baby tends to be heavier. Premature babies born under 27 weeks of age rarely survive, whereas those more than 30 weeks old usually do survive. For the events of delivery, see PARTURITION, HUMAN.

Table 2: Age, Size, and Weight

age from date of conception (calendar months)	sitting height		weight	
	mm	in.	g	lb
Two	28	1	2.25	.75 oz
Three	75	3	25	1 oz
Four	135	5.3	170	6 oz
Five	185	7.3	440	14 oz
Six	225	9	820	1.75 lb
Seven	270	10.6	1,380	3 lb
Eight	310	12.2	2,220	5 lb
Nine	360	14.2	3,150	7 lb

Abnormal development

MULTIPLE BIRTHS

It is both unusual and abnormal for the human species to produce more than one offspring at a time. "Twins" and "twinning" are used as general terms to designate any number of multiple births, since the same basic principles apply.

Fraternal and identical twins

Fraternal twins stem from multiple ovulations in the same cycle. Each oocyte develops singly in a separate follicle, is shed and fertilized individually, develops within its own chorionic sac, and forms an individual placenta. In some instances, two blastocysts implant close together and the expanding placentas meet and fuse. In such double placentas, however, the two blood circulations rarely communicate. The word dizygotic technically designates two-egg twins. Such pairs obviously are independent in sex determination and bear no more resemblance (other than that due to their identical age) than do other children of the same parents. Properly speaking, they are merely littermates. Nearly three-fourths of all American twins are dizygotic, whereas the Japanese ratio is only one-fourth. A tendency toward such multiple births exists in some family lines.

Wholly different are those true twins who are always of the same sex and are strikingly similar in physical, functional, and mental traits. Such close identity is enforced by their derivation from a single ovulated and fertilized egg, and hence by their acquisition of identical chromosomal constitutions. This twin type is appropriately designated as monozygotic. Three-fourths of such pairs develop within a common chorionic sac and accordingly share a common placenta; one-fourth have individual sacs and placentas. The latter condition results from a mishap before implantation, when the cleavage cells separate into two groups and then become individually implanting blastocysts. There is no discernible hereditary tendency toward the production of monozygotic twins.

Several atypical processes involving the inner cell mass or embryonic plate can produce separate embryos within a single sac: (1) The inner cells of a blastocyst may segregate into two inner cell masses. (2) Somewhat later in time, two embryonic axes may become established on a single embryonic disk. (3) A single axis may subdivide by fission or budding. (4) Duplication of any sort may combine with secondary subdivision; the famous Dionne

quintuplets are believed to have followed this sequence, which is also normal for the regular quadruplets of the Texas armadillo.

The abnormal nature of monozygotic twinning is emphasized when the two individuals are conjoined as a "double monster." The degree of union varies from slight to extensive, and the possession of a single or double set of internal organs depends on the intimacy of fusion at any particular level. Union occurs by the heads, upper trunks, or lower trunks; the joining may be by the dorsal, lateral, or ventral surfaces. Sometimes there is a marked disparity in the size of the two components, and in such instances the smaller twin is called a parasite. Conjoining results from divergent growth at the front or hind end of the emerging primitive axis of an embryo, or at both ends.

FETAL DEVIATIONS

Human embryos are subject to disease, abnormal development, and abnormal growth. Decline and death can occur at any stage of embryonic and fetal life, but since most deaths occur in the first two or three weeks of development they usually escape notice. Probably little more than half of all zygotes reach full-term birth. Most abnormalities resulting from faulty development originate within the first seven weeks of pregnancy, before the prospective mother is surely aware of her condition. Abnormalities that do occur in living infants tend toward the milder types, since the severe mishaps commonly terminate development before birth.

Deep-rooted folklore maintains that a pregnant woman may "mark" her baby through incurring physical injuries or becoming subjected to horrors or repulsive sights. Since there are no nervous connections between mother and fetus, such beliefs lack a foundation. Moreover, practically all of the alleged causative experiences occur long after the "related" abnormalities have been established in the embryo.

Defective health of the mother can, in some instances, become a cause of the physical impairment or death of a fetus. Certain infectious diseases, for example, may result in fetal injury; such related causative organisms can be a virus (German measles), a spirochetal micro-organism (syphilis), or a protozoon parasite (toxoplasmosis). Also, placental disorders, malformations of the mother's reproductive organs, and inadequate functioning of her endocrines may provide an unfavourable environment for normal development. Birth itself imposes the risk of oxygen deficiency or other injury; either may result in some malfunctioning of the brain.

TERATOLOGY

Incidence, causation, and types of abnormality

Teratology is concerned with all features of abnormal generation and development of the embryo (embryogenesis) and their end products. The incidence of defective development is high. One infant in 14 that survive the neonatal period bears an abnormality of some kind and degree, and half of these babies have more than one malformation. Internal, concealed defects are more numerous than external ones, and some defects do not become apparent until childhood. One baby in 40 is born with a structural defect that needs treatment. Some types of abnormality are commoner in males (*e.g.*, pyloric stenosis, the narrowing of the opening between the stomach and the intestine), while other types predominate in females (*e.g.*, dislocated hip). Besides the obvious derangements of a gross nature, there are other aberrations at a submicroscopic, molecular level that are called inborn errors of metabolism. In these an enzyme deficiency blocks the normal course of intermediary metabolism and results in abnormal chemical functioning. Such errors involve proteins, carbohydrates, lipids, and pigments. The abnormal products may be stored or excreted. For some other information on teratology, see BIRTH DEFECTS AND CONGENITAL DISORDERS.

Important among causes of abnormalities are hereditary factors. Such include gene mutations, which may become Mendelian dominants (*e.g.*, fused fingers, which need be inherited from only one parent to appear in the offspring), recessives (*e.g.*, albinism, which does not become evident unless its gene is inherited from both parents), or sex-linked factors (*e.g.*, hemophilia, which becomes manifest in males). Besides the heritable defects, whose possibilities of recurrence can be estimated, there are many genetic results that are due to chance, are not passed on, and do not occur in other offspring. An unequal distribution of chromosomes during meiosis, leading to abnormal assortments, occurs in somatic (nonsex) chromosomes (*e.g.*, mongolism) and in sex chromosomes (*e.g.*, Klinefelter's syndrome).

Environmental factors, both external and internal, are also important. Among physical agents, mechanical pressures or blows are no longer considered to be significant because of the protection supplied by the uterus and the fluid-filled amniotic sac. On the other hand, irradiation is a wholly effective physical agency, as experiments (and atomic bombing) have amply proved. Various chemical agents, used experimentally on pregnant animals, are highly teratogenetic (productive of physical defects within the uterus), and at least one drug (thalidomide) has demonstrated its potency on human embryos. Deficiencies of some fetal hormones are associated causally with gross bodily defects (*e.g.*, male hormone and false hermaphroditism, a condition in which the gonads are of one sex but some appearances suggest the other). Similarly, hormonal excess can be productive (*e.g.*, growth-promoting hormone and gigantism). Vitamin deficiencies are effective experimentally but are not factors in ordinary substandard maternal diets. The role of infectious agents has been mentioned previously. Immunity incompatibilities may exist between mother and fetus (*e.g.*, Rh blood factor and fetal hemolysis, the release of hemoglobin from red blood cells).

There appear to be several ways in which teratogenetic agents can affect susceptible embryonic cells. But whatever the manner of interference may be, the final result is probably either cell impairment or death or a changed rate of growth. Either of these measures puts local development out of step with adjoining parts and upsets the coordinated schedule of development.

Development of organs

ECTODERMAL DERIVATIVES

Origin of the skin

Integumentary system. The skin has a double origin. Its superficial layer, or epidermis, develops from ectoderm. The initial single-layered sheet of epithelial cells becomes multilayered by proliferation, and cells nearer the surface differentiate a horny substance. Pigment granules appear in the basal layer, at least, of all races. The epidermis of the palm and sole becomes thicker and more specialized than elsewhere. Cast-off superficial cells and downy hairs mingle with a greasy glandular secretion and smear the skin in the late fetal months; the pasty mass is called vernix caseosa. The deep layer of the skin, or dermis, is a fibrous anchoring bed differentiated from mesoderm. In the later fetal months the plane of union between epidermis and dermis becomes wavy. The permanently ridged patterns are notable at the surface of the palm and sole.

Nails, hairs, and skin glands

Nails develop in pocket-like folds of the skin near the tips of digits. During the fifth month specialized horny material differentiates in proliferating ectodermal cells. The resulting nail plate is pushed forward as new plate substance is added in the fold. Fingernails reach the finger tips one month before birth. Hairs, produced only by mammals, begin forming in the third month as cylindrical pegs that grow downward from the epidermis into the dermis. Cells at the base of the peg proliferate and produce a horny, pigmented thread that moves progressively upward in the axis of the original cylinder. This first crop of hairs is a downy coat named lanugo. It is prominent by the fifth month but is mostly cast off before birth. Unlike nails, hairs are shed and replaced periodically throughout life.

Sebaceous glands develop into tiny bags, each growing out from the epithelial sheath that surrounds a hair. Their cells proliferate, disintegrate, and release an oily secretion. Sweat glands at first resemble hair pegs, but

the deep end of each soon coils. In the seventh month an axial cavity appears and later is continued through the epidermis. The mammary glands, peculiar to mammals, are specialized sweat glands. In the sixth week a thickened band of ectoderm extends between the bases of the upper and lower limb buds. In the pectoral (chest) region only, gland buds grow rootlike into the primitive connective tissue beneath. During the fifth month 15 to 20 solid cords foretell the future ducts of each gland. These later canalize and open into a pit that after birth elevates into a nipple. Until late childhood the mammary glands are identical in both sexes.

Mouth and anus. *Mouth.* The mouth is a derivative of the stomodeum, an external pit bounded by the overjutting primitive nasal region and the early upper and lower jaw projections. Its floor is a thin membrane, where ectoderm and endoderm fuse. Midway in the fourth week this membrane ruptures, making continuous the primitive ectodermal mouth and endodermal pharynx (throat). Lips and cheeks arise when ectodermal bands grow into the mesoderm and then split into two sheets. Teeth have a compound and relatively complicated origin, in which the cap of enamel develops from ectoderm, whereas the main mass of a tooth, the dentin, and the encrusting cementum about the root differentiate from mesoderm. For details, see TEETH AND GUMS, HUMAN. The salivary glands arise as ectodermal buds that branch, bushlike, into the deeper mesoderm. Berrylike endings become the secretory acini (small sacs), while the rest of the canalized system serves as ducts. The palate is described in relation to the nasal passages. A tiny pocket detaches from the ectodermal roof of the stomodeum and becomes the anterior, or frontward, lobe of the hypophysis, also called the pituitary. The anterior lobe fuses with the neural lobe of the gland (see below).

Anus. A double-layered, oval membrane separates the endodermal hindgut from an ectodermal pit, called the proctodeum, the site of the future anal canal and its orifice, the anus. Rupture at eight weeks creates a communication between the definitive anus and the rectum.

Origin, histogenesis, and morphogenesis of central nervous system

Central nervous system. Both the brain and the spinal cord arise from an elongate thickening of the ectoderm that occupies the midline region of the embryonic disk. The sides of this neural plate elevate as neural folds, which then bound a gutter-like neural groove (Figure 1L). Further growth causes the folds to meet and fuse, thereby creating a neural tube. The many-layered wall of this tube differentiates into three concentric zones, first indicated in embryos of five weeks. The innermost zone, bordering the central canal, becomes a layer composed of long cells called ependymal cells, which are supportive in function. The middle zone becomes the gray substance, a layer characterized by nerve cells. The outermost zone becomes the white substance, a layer packed with nerve fibres. The neural tube also is demarcated internally by a pair of longitudinal grooves into dorsal and ventral halves. The dorsal half is a region associated with sensory functioning and the ventral half with motor functioning.

The gray substance contains primitive stem cells, many of which differentiate into neuroblasts. Each neuroblast becomes a neuron, or a mature nerve cell, with numerous short branching processes, the dendrons, and with a single long process, the axon. The white substance lacks neuroblasts but contains closely packed axons, many with fatty sheaths that produce the whitish appearance. The primitive stem cells of the neural tube also give rise to nonnervous cells called neuroglia cells, with barrier and other functions.

Brain. The head end of the neural plate becomes expansive even as it closes into a tube. This brain region continues to surpass in size the remainder, or spinal cord region. Three enlargements are prominent, the forebrain, midbrain, and hindbrain. The forebrain promptly gives rise to two secondary expansions, called the telencephalon and the diencephalon. The midbrain, which remains single, is called the mesencephalon. The hindbrain produces two secondary expansions called the metencephalon and the myelencephalon.

The telencephalon outpouches, right and left, into paired cerebral hemispheres, which overgrow and conceal much of the remainder of the brain before birth. Late in fetal life the surface of the cerebrum becomes covered with folds separated by deep grooves. The superficial gray cortex is acquired by the migration of immature nerve cells, or neuroblasts, from their primary intermediate position in the neural wall. The diencephalon is preponderantly gray substance, but its roof buds off the pineal body, which is not nervous tissue, and its floor sprouts the stalk and neural (posterior) lobe of the pituitary. The mesencephalon largely retains its early tubular shape. The metencephalon develops dorsally into the imposing cerebellum, with hemispheres that secondarily gain convolutions clothed with a gray cortex. The myelencephalon is transitional into the simpler spinal cord. Roof regions of the telencephalon, diencephalon, and myelencephalon differentiate vascular choroid plexuses (including portions of the pia matter, or innermost brain covering, that project into the ventricles, or cavities, of the brain), which secrete cerebrospinal fluid.

Spinal cord. For a time the spinal cord portion of the neural tube tapers gradually to an ending at the tip of the spine. In the fourth month it thickens at levels where nerve plexuses, or networks, supply the upper and lower limbs; these are called the cervical and lumbosacral enlargements. At this time the spine begins to elongate faster than the spinal cord. As a result, the caudal (hind) end of the anchored cord becomes progressively stretched into a slender, nonnervous strand known as the terminal filament. Midway in the seventh month the functional spinal cord ends at a level corresponding to the midpoint of the kidneys. Both the brain and the spinal cord are covered with a fibrous covering, the dura mater, and a vascular membrane, the pia-arachnoid. These coverings differentiate from local, neighbouring mesoderm.

Histogenesis of peripheral nervous system

Peripheral nervous system. In general, each craniospinal nerve has a dorsal (posterior) root that bears a ganglion (mass of nerve tissue) containing sensory nerve cells and their fibres, and a ventral (anterior) root that contains motor nerve fibres but no nerve cells. Ganglion cells differentiate from cells of the neural crest, which is at first a cellular band pinched off from the region where each neural fold continues into ordinary ectoderm. Each of these paired bands breaks up into a series of lumps, spaced in agreement with the segmentally arranged mesodermal somites. Neuroblasts within these primordial ganglia develop a single stem and hence are called unipolar. From this common stem one nerve process, or projection, grows back into the adjacent sensory half of the neural tube; another projection grows in the opposite direction, helping to complete the dorsal root of a nerve. Neuroblasts of motor neurons arise in the ventral half of the gray substance of the neural tube. They sprout numerous short, freely branching projections, the dendrons, and one long, little-branching projection, the axon; such a neuron is called multipolar. These motor fibres grow out of the neural tube and constitute a ventral root. As early as the fifth week they are joined by sensory fibres of the dorsal root and continue as a nerve trunk.

Cells of the neural crest are versatile and differentiate into things other than sensory neurons. Among these variants are cells that encapsulate ganglion cells and others that become neurolemma cells, which follow the peripherally growing nerve fibres and ensheath them. The neurolemma cells cover some nerve fibres with a sheath of fatty substance called myelin.

Spinal nerves. Spinal nerves are sensorimotor nerves, with dorsal and ventral roots, as already described. A network called a brachial plexus arises in relation to each upper limb and a lumbosacral plexus in relation to each lower limb. The spine, elongating faster than the spinal cord, drags nerve roots downward, since each nerve must continue to emerge between the same two vertebrae. Because of their appearance, the obliquely coursing nerve roots are named the cauda equina, the Latin term for horse's tail.

Cranial nerves. Cranial nerves V, VII, IX, and X arise in relation to embryonic branchial arches but have origins similar to the spinal nerves. The olfactory nerves

(cranial nerve I) are unique in that their cell bodies lie in the olfactory epithelium (the surface membrane lining the upper parts of the nasal passages), each sending a nerve fibre back to the brain. The so-called optic nerves (II) are not true nerves but only tracts that connect the retina (a dislocated portion of the brain) with the brain proper. Nerves III, IV, VI, and XII are pure motor nerves that correspond to the ventral roots of spinal nerves. The acoustic nerves (VIII) are pure sensory nerves, each with a ganglion that subdivides for auditory functions and functions having to do with equilibrium and posture; they correspond to dorsal roots. Nerves X and XI are a composite of which XI is a motor component.

Autonomic nervous system. The autonomic nervous system is made up of two divisions, the sympathetic and the parasympathetic nervous systems; it controls such involuntary actions as constriction of blood vessels. Some

Origins of autonomic nervous system

cells of the neural crests migrate and form paired segmental masses alongside the aorta, a principal blood vessel. Part of the cells become efferent multipolar ganglion cells (cells whose fibres carry impulses outward from ganglions, or aggregates of nerve cells) and others merely encapsulate the ganglion cells. These autonomic ganglia link into longitudinal sympathetic trunks. Some of the neuroblasts migrate farther and assemble as collateral ganglia—ganglia not linked into longitudinal trunks. Still others migrate near, or within, the visceral organs that they will innervate and produce terminal ganglia. These ganglia are characteristic of the parasympathetic system.

Some cells of certain primitive collateral ganglia leave and invade the amassing mesodermal cortex of each adrenal gland. Consolidating in the centre, they become the endocrine cells of the medulla.

Sense organs. *Olfactory organ.* Paired thickenings of ectoderm near the tip of the head infold and produce olfactory pits. These expand into sacs in which only a relatively small area becomes olfactory in function (Figure 2E). Some epithelial cells in these regions remain as inert supporting elements. Others become spindle-shaped olfactory cells. One end of each olfactory cell projects receptive olfactory hairs beyond the free surface of the epithelium. From the other end a nerve fibre grows back and makes a connection within the brain.

Gustatory organ. Most taste buds arise on the tongue. Each bud, a barrel-shaped specialization within the epithelium that clothes certain lingual papillae (small projections on the tongue), is a cluster of tall cells, some of which have differentiated into taste cells whose free ends bear receptive gustatory hairs. Sensory nerve fibres end at the surface of such cells. Other tall cells are presumably inertly supportive in function.

Origins of eye and ear

Eye. The earliest indication of an eye is an optic vesicle (sac) bulging from each side of the forebrain (Figure 2B, C). It quickly becomes an indented optic cup, connected to the brain by a slender optic stalk. Most of the cup will become the retina, but its rim represents the epithelial part of the insensitive ciliary body and iris. The thicker inner layer of the cup becomes the neural layer of the retina, and by the sixth month three strata of neurons are recognizable in it: (1) visual cells, each bearing either a photoreceptive rod or a cone at one end; (2) bipolar cells, intermediate in position; (3) ganglion cells, which sprout axons that grow back through the optic stalk and make connections within the brain. The thin outer layer of the cup remains a simple epithelium whose cells gain pigment and make up the pigment epithelium of the retina.

The lens arises as a thickening of the ectoderm adjacent to the optic cup. It inpockets to form a lens vesicle and then detaches. The cells of its back wall become tall, transparent lens fibres. Mesoderm surrounding the optic cup specializes into two accessory coats. The outer coat, the tough, white sclera, is continuous with the transparent cornea. The inner coat, the vascular choroid, continues as the vascular and muscular ciliary body and the vascularized tissue of the iris. The eyelids are folds of adjacent skin; from the inside of each upper lid several lacrimal glands bud out (see EYE AND VISION, HUMAN).

Ear. The projecting part (auricle) of the external ear develops from hillocks on the first and second branchial arches (Figure 2C–E). The ectodermal groove between those arches deepens and becomes the external auditory canal. The auditory tube and tympanic cavity—the cavity at the inner side of the eardrum—are expansions of the endodermal pouch located between the first and second branchial arches. The area where ectodermal groove and endodermal pouch come in contact is the site of the future eardrum. The chain of three auditory ossicles (small bones) that stretches across the tympanic cavity is a derivative of the first and second arches.

The epithelium of the internal ear is at first a thickening of ectoderm at a level midway of the hindbrain (Figure 2B). This plate inpockets and pinches off as a closed sac, the otocyst. Its ventral part elongates and coils to resemble a snail's shell, thereby forming the cochlear duct, or seat of the organ of hearing. A middle region of the otocyst becomes chambers known as the utricle and saccule, related to the sense of balance. The dorsal part of the otocyst remodels drastically into three semicircular ducts, related to the sense of rotation. Fibres of the acoustic nerve grow among specialized receptive cells differentiated in certain regions of these three divisions.

MESODERMAL DERIVATIVES

Skeletal system. Except for part of the skull, all bones pass through three stages of development: membranous, cartilaginous, and osseous. The earliest ossification centres appear in the eighth week, but some do not arise until childhood years and even into adolescence.

Axial skeleton. The ventromedial walls (the walls toward the front and the midline) of the paired somites break down, and their cells migrate toward the axial notochord and surround it. Differentiation and growth of these segmental masses produce the jointed vertebrae. Ribs also grow out of each primitive vertebral mass, but they become long only in the thoracic region. Here their ventral ends join the sternum, which arises independently by the fusion of a pair of bars.

The skull has three components, different in origin. Its basal region is an ancient heritage whose bones pass through the three typical stages of development. By contrast, the sides and roof of the skull develop directly from membranous primordia, or rudiments. The jaws are derivatives of the first pair of cartilaginous branchial arches but develop as membrane bone. Ventral ends of the second to fifth arches contribute the cartilages of the larynx and the hyoid bone (a bone of horseshoe shape at the base of the tongue). Dorsal ends of the first and second arches become the three auditory ossicles (the small bones in the middle ear).

Development of the bones of the arms and legs; the joints

Appendicular skeleton. The limb bones develop in three stages from axial condensations in the local mesoderm. The shoulder and pelvic supports are comparable sets, as are the bones of the arms and legs.

Articulations. Some type of joint exists wherever bones meet. Joints that allow little or no movement consist of connective tissue, cartilage, or bone. Movable joints arise as fluid-filled clefts in mesoderm, which condenses peripherally into a fibrous capsule.

Muscular system. Much of each somite differentiates into myoblasts (primitive muscle cells) that become voluntary muscle fibres. Aggregations of such fibres become muscles of the neck and trunk. Muscles of the head, and some of the neck muscles, originate from mesoderm of branchial arches. Muscles of the limbs seemingly arise directly from local mesoderm. In general, muscle primordia may fuse into composites, split into subdivisions, or migrate away from their sites of origin. During these changes they retain their original nerve supply. Regardless of differences in source of origin, all voluntary muscle fibres are of the same striated type (marked by dark and light stripes). Spontaneous movements begin to occur in embryos about ten weeks old. In general, involuntary muscle differentiates from mesoderm surrounding hollow organs; only the cardiac muscle type is striated.

Vascular system. All hollow organs, including arteries, veins, and lymphatics, are lined with epithelium—

the principal functional tissue—and are ensheathed with muscular and fibrous coats.

Blood vessels. Primitive blood vessels arise in the mesoderm as tiny clefts bordered by flat endothelial cells. Growth and coalescence produce networks, out of which favoured channels persist as definite vessels, while others decline and disappear. A bilaterally symmetrical system of vessels is well represented in embryos four weeks old. This early plan is profoundly altered and made somewhat asymmetrical during the second month by fusions, atrophies, emergence of new vessels, and rerouting of older ones. The alterations reflect adjustments to changing form and pattern within the developing organ systems of the body.

Arteries cranial to the heart (headward of the heart) are mostly products of the paired aortic arches, which course axially within the branchial arches, thus interconnecting the ventral aorta with paired dorsal aortas. The third pair of aortic arches becomes the common carotids; the fourth pair, the aortic arch and brachiocephalic; the fifth pair, the pulmonary arteries and ductus arteriosus. The dorsal aortas fuse into the single descending aorta, which bears three sets of paired, segmental branches. The dorsal set becomes the subclavian, intercostal, and lumbar arteries. The lateral set becomes arteries to the diaphragm, the adrenal glands, the kidneys, and the sex glands. The ventral set becomes the celiac, mesenteric, and umbilical arteries. Axial arteries to both sets of limb buds emerge from an original plexus, but they undergo drastic alteration and extensive replacement.

The primitive veins are symmetrically bilateral. They consist of vitelline veins from the yolk sac, umbilical veins from the placenta, and precardinal and postcardinal veins from the cranial and caudal regions (the regions toward the head and toward the tail) of the body. Drastic transformations occur in all of these, and new pairs of veins (subcardinals and supracardinals) arise also, caudal to the heart. From the vitellines come chiefly the portal and hepatic veins. The left umbilical becomes the main return from the placenta by making a diagonal channel, the ductus venosus, through the liver to the heart. The precardinal veins change their names to the internal jugulars, but near the heart an interconnection permits both to drain into a common stem, then called the superior vena cava. Caudal to the heart, the postcardinals virtually disappear, and all blood return shifts to the right side as a new compound vessel, the inferior vena cava, becomes dominant. Pulmonary veins open into the left atrium. Veins from the limb buds organize from an early peripheral border vein.

Lymphatic vessels. The lymph vessels develop independently in close association with veins. Linkages produce the thoracic duct, which is the main drainage return for lymph. Masses of lymphocytes accumulate about lymphatic vessels and organize as lymph nodes. The spleen has somewhat similar tissue, but its channels are supplied with blood.

Origins of the heart

Heart. Fusion combines two endothelial tubes, and these are surrounded by a mantle of mesoderm that will become the muscular and fibrous coats of the heart. At three weeks the heart is a straight tube that is beginning to beat (Figure 1M). Starting at the head end, four regions can be recognized: bulbus, ventricle, atrium, and sinus venosus. Since the heart is anchored at both ends, rapid elongation forces it to bend. In doing this, the sinus–atrium and bulbus–ventricle reverse their original relations. Further development concerns the transformation of a single-chambered heart into one with four chambers.

The atrium becomes subdivided by the growth of two incomplete partitions, or septa, placed close together and each covering the defect in the other. The ventricle also subdivides, but by a single, complete partition. A canal, connecting atria and ventricles, becomes two canals. The bulbus is absorbed into the right ventricle, and its continuation (the truncus) subdivides lengthwise, forming the aorta and the pulmonary artery. The right horn of the sinus venosus is absorbed into the right atrium, together with the superior and inferior venae cavae, which originally drained into the sinus. The transverse portion of the sinus persists as the coronary sinus. The pulmonary veins retain their early drainage into the left atrium. Important valves develop and ensure flow within the heart from atria to ventricles, and outward from the ventricles into the aorta and the pulmonary artery.

Birth initiates breathing, and the abandonment of the placental circulation follows. These changes entail a drastic rerouting of blood through the heart. As a result, the two atrial septa fuse and no longer permit blood to pass from the right atrium to the left atrium. Blood in the pulmonary artery no longer virtually bypasses the lungs; previously it had passed to the aorta directly through a shunt offered by the ductus arteriosus. As a sequel to these changes, the abandoned umbilical arteries, umbilical vein, ductus venosus, and ductus arteriosus all collapse and become fibrous cords.

Urinary system. Vertebrates have made three experiments in kidney production: the pronephros, or earliest type; the mesonephros, or intermediate kidney; and the metanephros, or permanent kidney. All arise from the cellular plates called nephrotomes that connect somites with the mesodermal sheets that bound the body cavity. The vestigial pronephros is represented solely by several pairs of tubules; they join separately formed excretory ducts that grow downward and enter the cloaca, the common outlet for urine, genital products, and for intestinal wastes. Next tailward arise some 40 pairs of nephric (kidney) tubules that constitute the mesonephros; these tubules join the same excretory ducts, hereafter called the mesonephric ducts. The two sets of mesonephric tubules serve as functioning kidneys until the tenth week.

Origins of the kidney, urinary bladder, and urethra

Each permanent kidney, or metanephros, develops still farther tailward. A so-called ureteric primordium buds off each mesonephric duct, near its hind end. The ureteric stem elongates and expands terminally, thereby forming the renal pelvis and calices; continued bushlike branching produces collecting ducts. The early ureteric bud invades a mass of nephrotome tissue. The branching collecting ducts progressively break this tissue up into tiny lumps, each of which becomes a long secretory tubule, or nephron, and joins a nearby terminal twig of the duct system. Continued proliferation of ducts and nephric tissue produces over a million urine-producing tubules in each kidney.

The blind caudal end of the endodermal hindgut absorbs the stem of each mesonephric duct, whereupon the remainder of the duct and the ureter acquire separate openings into the hindgut. This expanded region of the gut, now a potential receptacle for feces, urine, and reproductive products, is known as a cloaca. It next subdivides into a rectum behind and a urogenital sinus in front. The sinus, in turn, will specialize into the urinary bladder and the urethra. The prostate gland develops as multiple buds from the urethra, close to the bladder.

Origins of sex glands and ducts and external genitalia

Genital system. The genital organs begin to develop in the second month, but for a time sex is not grossly distinguishable. Also, a double set of male and female ducts arise, and not until later does the unneeded set decline. Hence, this period is commonly called the indifferent stage.

Gonads. Sex glands develop in a pair of longitudinal ridges located alongside the mesentery, the anchoring fold of membrane to the gut. The primordial sex cells appear first in the cloacal wall, from which they migrate upward in the gut, pass through its mesentery, and finally invade the genital ridges, where they proliferate. The testes are the earliest type of gonad to organize. They begin by developing testis cords and a testis capsule. The cords radiate from one focal point at the periphery, and thin fibrous partitions segregate groups of the cords within wedge-shaped compartments. These cords do not gain channels and become semen-producing tubules until near the time of puberty. The ovaries organize somewhat tardily by differentiating an outer portion, the cortex, and a central portion, the medulla. The cortex contains the primordial sex cells; these become surrounded by a layer of ordinary cells, thereby forming primary ovarian follicles. The further advance of these follicles and that of the

testis tubules have been described earlier. Both the testes and the ovaries undergo relative shifts from their early sites to lower positions in the body. But only the testes make a bodily descent; this is into the scrotum.

Genital ducts. In the male, a few mesonephric tubules on each side do not degenerate but link up with the neighbouring testis tubules. The converted mesonephric tubules and the retained mesonephric ducts become the male sex ducts. Near their terminations they outpouch seminal vesicles and then open into the urethra. In the female, a pair of ducts develops from the epithelium clothing the mesonephric ridges. These ducts, known as the uterine tubes, mostly parallel the courses of the mesonephric ducts, but at their lower ends they unite into a common tube that becomes the uterus and vagina.

External genitalia. Both sexes develop a genital tubercle (*i.e.*, a knob) and a pair of urogenital folds flanked by a pair of genital swellings. At three months these rudiments begin to assume male or female characteristics. In the male, the tubercle and the united urogenital folds combine as the penis, thereby continuing the urethra to its end; the genital swellings shift toward the anus, fuse, and become the scrotum. In the female, the tubercle remains small, as the clitoris; it does not contain the urethra. The urogenital folds remain unclosed as the lesser vulvar lips and are flanked by the unshifted and unfused genital swellings, or greater lips.

Coelom. The lateral mesoderm, beyond the somites and nephrotomes, splits into two layers: the somatic layer and, underlying the somatic layer, the splanchnic layer. The intervening space is the coelom. As the embryo's body folds off, its coelom becomes a single closed cavity. In it can be recognized, regionally, a provisional pericardial cavity (cavity for the heart), two pleural canals (for the lungs), and a peritoneal cavity (for the abdominal contents). A thick plate of mesoderm, the transverse septum, constitutes a partial partition just ahead of the developing liver. Two pairs of membranes grow out from the septum. One set separates the pericardial cavity from the two pleural cavities; these membranes later expand into the pericardium and enclose the heart. The other pair of membranes separates the pleural cavities from the peritoneal cavity of the abdomen. The definitive diaphragm is a composite partition, much of which is furnished by the transverse septum; lesser contributions are from the lateral body walls and the paired membranes that separated the pleural and peritoneal cavities.

Origins of the body cavities

ENDODERMAL DERIVATIVES

Pharynx. The tongue is a product of four branchial arches, whose ventral ends merge in its midplane. Papillae elevate from the surface, and taste buds arise as specializations within the covering epithelium of some of them. Pharyngeal pouches are early lateral expansions of the local endoderm, alternating with the branchial arches. The first pair elongate as the auditory tubes and tympanic cavities. The second pair mark the site of the tonsils. The third pair give rise to the halves of the thymus, and the third and fourth pairs produce the two sets of parathyroid glands. The thyroid gland buds off the pharyngeal floor in the midplane and at the level of the second branchial arches.

Development of the esophagus, stomach, and intestines

Digestive tube. As the embryo folds off, the endoderm is rolled in as the foregut and hindgut. Continued growth progressively closes both the midbody and the midgut (Figure 1P, Q). The esophagus remains as a simple, straight tube. The stomach grows faster on its dorsal side, thereby forming the bulging greater curvature; the stomach also rotates 90° so that its original dorsal and ventral borders come to lie left and right. The intestine elongates faster than the trunk, so that its loops find temporary room by pushing into the umbilical cord. Later, the loops return, completing a rotation that gives the characteristic final placement of the small and large intestines.

When the gut folds into a tube it is suspended by a sheetlike dorsal mesentery, or membranous fold. In the region of the stomach it forms an expansive pouch, the omental bursa. Secondary fusions of the bursa and of some of the rest of the mesentery with the body wall produce lines of attachment from stomach to rectum inclusive, different from the original midplane course. Such fusions also anchor firmly some parts of the tract. A ventral mesentery, beneath the gut, exists only in the region of the stomach and liver.

Major glands. The liver arises as a ventral outgrowth of the foregut that invades the early transverse septum. Although rapid growth causes it to bulge prominently away from this septum, it remains attached to the septum and hence to the definitive diaphragm. The differentiating glandular tissue takes the form of plates bathed by blood channels. The stem of the original liver bud becomes the common bile duct, whereas a secondary outgrowth produces the cystic duct and the gallbladder.

The pancreas takes its origin from a larger dorsal bud and a smaller ventral bud, both off the foregut. The two merge and their ducts communicate, but in man it is the lesser, ventral duct that becomes the stem outlet. Secretory acini are berrylike endings of the branching ducts. Pancreatic islets arise as special sprouts from the ducts; these differentiate into endocrine tissue that secretes insulin.

Nose, larynx, trachea, and lungs

Respiratory system. *Nasal cavity.* The first part of the respiratory system is ectodermal in origin. The olfactory sacs (see above *Sense organs*) become continuous secondarily with a passage captured from the primitive mouth cavity. This addition is produced by a horizontal partition, the palate. It arises from a pair of shelflike folds that grow out from the halves of the primitive upper jaw and then unite. The final nasal passage extends from the nostrils to the back of the pharynx.

Larynx, trachea, and lungs. A hollow lung bud grows off the floor of the endodermal pharynx, just caudal (tailward) to the pharyngeal pouches and in the midline. It has the form of a tube with an expanded end. The entrance to this tube is the glottis, and the region about it becomes the larynx. The tube proper represents the trachea. Its terminal expansion divides into two branches, and these tubes elongate as the primary bronchi. Continued growth and budding produce two side branches from the right bronchus and one from the left. These branches and the blind ends of the two parent bronchi indicate the future plan of the lungs, with three right lobes and two left lobes. Continued branchings, through the sixth month, produce bronchioles of different orders. In the final months the smaller ducts and early respiratory alveoli (air sacs) appear, the lungs losing their previous glandular appearance and also becoming highly vascular. Until breathing distends the lungs, these organs remain relatively small.

BIBLIOGRAPHY. L.B. AREY, *Developmental Anatomy*, 7th ed. (1965), a comprehensive, modern text on human development; B.I. BALINSKY, *An Introduction to Embryology*, 3rd ed. (1970), a college-level treatment of development with emphasis on early stages and comparative conditions; E.B. BLECHSCHMIDT, *The Stages of Human Development Before Birth* (1961), an atlas of human embryos, sections, and reconstructed organ systems; G.W. CORNER, *Ourselves Unborn* (1944), a popular description of human development, its biological significance, and its aberrations; W.J. HAMILTON *et al.*, *Human Embryology*, 4th ed. (1972), a well-illustrated account of human development; J.E. MORISON, *Foetal and Neonatal Pathology*, 2nd ed. (1963), a discussion of the disturbances of prenatal life and of adaptations to extra-uterine existence; J. NEEDHAM, *A History of Embryology*, 2nd ed. rev. (1959), a historical review from antiquity through the 18th century; B.M. PATTEN, *Human Embryology*, 3rd ed. (1968), a largely structural account of human development with emphasis on the nervous and cardiovascular systems; A. SORSBY (ed.), *Clinical Genetics* (1953), a compilation of known relations of heredity to disease and abnormal development; J.G. WILSON, "Experimental Studies on Congenital Malformations," *J. Chron. Dis.*, 10:110–130 (1959), a summary of recent discoveries on causative factors related to abnormal development.

(L.B.A.)

Emerson, Ralph Waldo

Essayist, major poet, lecturer whose voice and rhetoric sustained the faith of thousands in the lecture circuits be-

tween 1834 and the U.S. Civil War, cultural middleman through whom the aesthetic and philosophical currents of Europe and Asia passed to America, Emerson led his countrymen during the burst of literary glory known as the American renaissance (1835–65), serving as catalyst for such writers as Henry David Thoreau, Walt Whitman, Nathaniel Hawthorne, and Emily Dickinson. As a principal spokesman for Transcendentalism, the American tributary of European Romanticism, he gave direction to a warm, intuitional, religious, aesthetic, philosophical, and ethical movement that stressed a theoretical and practical way of life and combined ancient classical humanism with varieties of Oriental supernaturalism to support belief in an oversoul and in the "infinitude" (spiritual potential) of every man. Among his bequests to the 20th century, in which, according to one critic of awakening Emerson scholarship, he continues "the least limited, the most permanently suggestive" of American literary artists, are a controlled optimism with an awareness of the tragic; a Western or Yankee mysticism; poetry with a uniquely American quality; and the ideal of a vital, skeptical mind heroically affirming its individuality.

Principal spokesman for Transcendentalism

Early life and works. Ralph Waldo Emerson was born in Boston, Massachusetts, on May 25, 1803, to the Rev. William Emerson, a Unitarian clergyman and minister of the First Church, friend to the arts in the lean years following the American Revolution, and a founder of the Anthology Club of Boston. The club was a literary society out of which came the Boston Athenaeum, the *Monthly Anthology and Boston Review* (1803–11), and, eventually, the *North American Review*. Through his father, Emerson inherited the profession of divinity, which had attracted all his ancestors in direct line from Puritan days. The family of his mother, Ruth Haskins, was strongly Anglican, her father, John Haskins, being a communicant of King's Chapel, Boston, and both father and daughter loved the Anglican *Book of Common Prayer* and memorized its principal offices so that phrases from it sprang up naturally in their daily conversation. When John Haskins died, Emerson, then 11, wrote a short elegy beginning, "See the calm exit of the aged saint. . . ." Among Emerson's principal teachers, in later years, were Anglican writers and thinkers like Ralph Cudworth, Robert Leighton, Jeremy Taylor, and Samuel Taylor Coleridge—a fact that possibly explains some ambivalences.

On May 12, 1811, William Emerson died, leaving Ralph (aged nine) largely to the intellectual care of Mary Moody Emerson, his spinster aunt, who took her duties seriously, declaring that her brother's children "were born to be educated." In 1812 Emerson entered the Boston Public Latin School, where his juvenile verses were encouraged and his literary gifts recognized. In 1817, at the age of 14, he entered Harvard College, where he began his journals, which, when eventually published without abridgment, may prove to be the most remarkable record of the "march of Mind" or intellectual and moral progress to appear in America. Perhaps a little too young to earn distinction under the college marking system, he nonetheless developed intellectually and emotionally, graduating in August 1821, to enter school teaching while preparing for part-time study in the Harvard Divinity School. Though Emerson was licensed to preach in the Unitarian community in 1826, fatigue from his strenuous reading program and incipient lung disease required a sojourn in Georgia and Florida. Ordained to the Unitarian ministry at the Second Church, Boston, on March 11, 1829, he began concentrating on weekly sermons in which are to be found the great themes later developed in his famous essays. At this time he was reading the works of the Swedish philosopher and religious writer Emanuel Swedenborg; Sampson Reed, a Boston Swedenborgian writer; the poets Coleridge and Wordsworth; the French prelate François de Salignac de la Mothe Fénelon, who effectively proclaimed for him the majesty of the moral law within. Passages in these sermons may be considered first drafts for a number of his later poems, the prose style being intensely lyrical. In the pulpit he gained experience in the rhetoric which was to be employed with distinction on the lecture platform throughout his life.

Ordination to the Unitarian ministry

On September 30, 1829, as surviving love letters and his journals indicate, he entered into the most vital emotional experience of his life by marrying Ellen Louisa Tucker of New Hampshire. When she died of tuberculosis on February 8, 1831, his grief drove him to question his beliefs and his profession, turning him hopefully to other religious sects for evidence of vital truth—especially to Methodists, Quakers, and Swedenborgians of the Boston area. A period of agonized meditation in the White Mountains produced "Gnothi Seauton" (written in July 1831), an impassioned creed hymn, or meditation, in free verse on the witness of the oversoul in the heart of man and on the compensations by which it governed man's morality. From this point onward Unitarianism had little appeal, and his sermon opposing the asserted anachronism of the Lord's Supper (September 9, 1832), drawn from Quaker sources, was more an excuse for abandoning the ministry than a sufficient cause. His resignation having been reluctantly accepted, he sailed for Europe on Christmas Day in quest of spiritual communion with the three Romantic writers whom he had been studying with excitement. (His account of the meetings with Wordsworth, Coleridge, and Thomas Carlyle is given in *English Traits*.) His startled recognition of occult relationships between himself and natural objects in the Muséum Nationale d'Histoire Naturelle in Paris gave particularity to his already developing theory (suggested by Swedenborg and Coleridge) of a spiritual correspondence between man and nature, and it encouraged his browsing in available popular science books. When he returned to Concord in October 1833, he drew upon these and current literary studies for his first public lectures. Sparkling with wit, anecdote, proverbs, and his early Romanticism, his lectures offered a liberal education to his first Boston hearers and introduced a diffused and easily absorbed Transcendentalism in a style midway between the pedagogy employed in the sermons and the highly intricate prose of his essays.

Meeting with Wordsworth, Coleridge, and Carlyle

By courtesy of The Library of Congress, Washington, D.C.

Emerson, lithograph by Leopold Grozelier, 1859.

Mature life and works. The death of his brother Edward Bliss, in the fall of 1834, was another blow that tested his optimism. In November, he took up permanent residence in Concord, which largely through his fame became the mecca of American Romanticism. By the end of the year, he had reached maturity as a poet, as evidenced in lyrics such as "Rhodora," "Xenophanes," "Each and All," and "The Snow-Storm." By June 1835, he had organized the wisdom from his principal sources, which he had made his own, into a cogent Transcendental credo or "Statement of the First Philosophy," which was a preparation for his first book. He had, meanwhile, married the slightly older Lydia Jackson ("Lidian") of Plymouth, Massachusetts, on September 14, 1835, and had lost another brother, Charles Chauncy, on May 9, 1836.

A complex classic—and his first major challenge to the

wasteland of his day—was *Nature* (1836), which may be classed with Whitman's "Song of Myself," Thoreau's *Walden*, Wordsworth's *Prelude*, and T.S. Eliot's *Four Quartets* as compressed and objectified personal experience, credal statement, confession, and prophecy combined. A prose lyric, it identifies the levels of nature's rich ministry to man in the conduct of his life and his quest for salvation, her role being analogous to Jesus' in the Christian dispensation. Influenced by ancient writers on the perennial philosophy (such as Plato and Plotinus) and moderns (such as Coleridge, Swedenborg and Goethe), it provides a compendium of Romantic doctrine and the highest point in Emerson's idealism, which was to be tempered but not compromised as, in middle life, he grew more aware of the limitations of man and society. The clarity of its structure reflects the instructional or programmatic character of the earlier lectures and sermons from which he drew many passages. All his later writings look back to this little book as their master plan and are variations of its themes. In this same year (1836), almost as if to illustrate the doctrines expressed at the end of *Nature*, his son Waldo was born and lovingly studied in terms of Romantic theories, infancy being defined as "the perpetual Messiah, which comes into the arms of fallen men, and pleads with them to return to paradise." (Emerson's other children were Ellen [1839], Edith [1841], and Edward Waldo [1844].

Theme of "The American Scholar"

His second challenge, "The American Scholar" (August 31, 1837), directed to the Harvard intelligentsia, under the auspices of Phi Beta Kappa, warned against pedantry, imitation of others, traditionalism, and scholarship unrelated to life, reminding the audience that since true wisdom flows from above by divine influx, the primary agency for acquiring it or (to use Plato's imagery) "unlocking" or "awakening" it in each man was the midwife nature. Books recording the great traditions of the past he considered of secondary importance, relegating them to the scholar's "idle times." To these channels of truth he added action, or full participation in human life. The address was a deliberate declaration of literary independence from England and an unconscious apology for his new vocation as artist and teacher on the lecture platform. His third major challenge (July 15, 1838) was the "Address at Divinity College," Harvard University, directed against a lifeless Christian tradition, especially Unitarianism as he had known it. Developing themes from "Gnothi Seauton" and emphasizing the "impersoneity" of the Transcendental and indwelling divinity, he dismissed religious institutions and the divinity of Jesus as dodges or failures to encounter deity directly through the moral principle or intuited sentiment of virtue. Since God so incarnates himself, men are differenced from Jesus and other spiritual leaders only in degree of reception and awareness. Alienated because of this impiety or infidelity, he now had few opportunities to preach in churches, and for many years Harvard had nothing to do with him; but young disciples joined the members of the informal Transcendental Club (founded in 1836) and encouraged him. In "Uriel" (written 1839), a parable of the difficulty in announcing or transmitting new truth and of society's inevitable resistance to its great philosophers and religious leaders, he apparently had in mind the career of Coleridge, who had suffered ignominy and rustication; but he used imagery suggesting the divinity school setting and unconsciously identified himself with the great English poet and critic.

Publication of *The Dial*

In July 1840, he helped launch *The Dial*, first edited by Margaret Fuller and later by himself, providing a publishing outlet for the new ideas which the Transcendentalists were trying to present to America. Though short-lived, it accomplished many of its purposes, providing the rallying point for the younger members of the school. Out of his continuing lecture series, meanwhile, Emerson gathered his *Essays* into two volumes (1841 and 1844), which became the basis of his international fame, linking him with other contributors to this genre such as Cicero, Plutarch, Michel de Montaigne, and Francis Bacon. In "History," the concept is interpreted not as a chronology of outward events but as an inner or existential awareness of reality, the record of the indwelling deity, or mind, as it manifests itself in the human condition. "Self-Reliance" (that is, God-reliance) celebrates the same divine intuitive principle in man, the voice of duty, which makes imitation of others or conformity to traditions a spiritual suicide. Nonconformity must be the rule of genius; intuition, its preservation. "Compensation" harks back to "Gnothi Seauton" in charting the instantaneous judgments from God or from the world that operate within man as he behaves, no element of chance being possible in the spiritual universe. One pays, rewards, and cheats himself. "Friendship" owes much to the Swedenborgian doctrines of influx and of the "association of spirits," the rule being that spiritual "like only can know or understand like." Human relationships like marriage gain little from arbitrary regulation or commitment. Without spiritual reciprocity no union can exist and no growth in character. "The Poet" predicted a bard who would choose America in the large for his theme and not the nightingales and skylarks of England. In his indigenous verse not metres "but a metre-making argument" would be apparent. Such a poet would be the complete man, receiving the greatest flow of power and truth and expressing in words his intuitions of eternal beauty. In "Politics," Emerson's doctrine of the God-reliant individual is related to government, which is best when it governs least and should expire altogether when it produces the wise man.

The death of little Waldo on January 27, 1842, clouded Emerson's middle years until his grief was partly relieved in "Threnody," one of the great elegies in the language. His earliest idealism, according to which all things seemed possible to man in union with the divine principle, began to accommodate itself to limitation, or fate. His later works show increasing acquiescence under the state of things, less reliance on the self, greater respect for society, and an awareness of the ambiguities and incompleteness of genius, which, at best, could only represent the divine completeness. His collected *Poems* (1846) were supplemented by others in *May-Day* (1867), the two volumes establishing his reputation as a major American poet. The first served as a visiting card in England when he returned there to lecture in 1847–48, helping him to establish friendships with younger poets, who were, perhaps, the first to appreciate such lyrics as "Mithradates," "Hamatreya," "The Humble-Bee," "Give All to Love," "Merlin," "The Apology," and "Bacchus"—to name only a few. While abroad this second time he was able to speak in behalf of Thoreau and Bronson Alcott, the Transcendentalist teacher and writer, and to invite like-minded Englishmen to the American shores.

Representative Men drew upon lectures and evaluations of his reading to praise the "uses" of men of genius without differencing them (except in degree) from ordinary men. Great men serve the human race directly by offering health, eternal youth, magical power, and prophecy; but they are not themselves original forces. Plato, the philosopher; Swedenborg, the mystic; Montaigne, the skeptic; William Shakespeare, the poet; Napoleon Bonaparte, the man of the world; and Goethe, the writer—these are the great servants of the race. *English Traits* (1856) is a character analysis of a people from which he himself stemmed and attests his admiration for the traditions of his forebears in spite of his strong declarations of independence.

Humanism of *The Conduct of Life*

The Conduct of Life (1860), Emerson's most mature work, reveals a developed humanism together with a full awareness of man's limitations, or partial development, in an evolutionary universe—man's freedom (resulting from the Spirit acting through his will) and "Fate" (one's boundaries established by time, space, environment and inheritance) both operating. Rich in the texture of human experience, it may be considered as partly confession. "Power" is achieved through the will, its value depending upon the kind of will exercised. "Wealth" is a good, especially if earned and spent by a "genius" ever enlarging its spiritual awareness. "Culture" (melioration) evaluates the popular procedures in getting an education, warns against what is called "success," and defines the title as

the process of correcting a man's imbalances or modulating the violence of his master tones. "Behavior" mounts from a consideration of manners that are deliberately acquired (thereby making society possible) to the unconscious flowering of character through self-reliance (God-reliance), which lies at the basis of love, friendship, and power. "Worship" (religion), the zenith of the cultural process, rests not upon history or dead traditions but upon the ever-living moral sense, discussed earlier in "Gnothi Seauton." "Considerations by the Way," which shows familiarity with Charles Darwin's ideas launched in the *Origin of Species* (1859), stresses goodness emerging from evil; the necessity of living with the witless, the unfit, the undeveloped; courage to be one's best through it all; and the importance of love, friendship, and service in the ascent of character. "Beauty" is ultimately a revelation of the moral sentiment whether it appears in outer form, human behaviour, or expression of thought. In "Illusions," he shifted his vision from the chaotic human scene already discussed with great psychological penetration and relegated it to what the Hindus call *māyā*, pointing to the super-realities (the "gods" on their thrones), sustaining the phenomenal universe, speaking in the moral sentiment, and beckoning mortals upward to their sphere. The majestic peroration contrasts with the sparkling wit, humour, metaphor, irony, and skepticism of the whole work.

His "Thoreau" (1862) brought together with some reservations his positive thoughts concerning the enigmatic disciple and in so many ways Emerson's teacher, who had died on May 6 and whose significance as a major American stylist possessing great charisma was to surpass all 19th-century expectations. Emerson had foreshadowed his brilliant protégé in "Woodnotes" (1840), much as he had predicted Whitman in "The Poet" (1844). He had praised Thoreau to the young versifiers in England and had seemed to commend his experimental return to nature at Walden Pond in the *Conduct of Life;* but his last word was scarcely prophetic of Thoreau's permanent place in the literature of the world.

Later life and works. The outbreak of the Civil War, which ushered in a period of Realism in American letters and threatened to end Transcendentalism as a literary force, found Emerson supporting the Union and abolition. He celebrated the Emancipation Proclamation in the "Boston Hymn" (1863) and followed that with "Voluntaries," lyric meditations on the black race, to which he promised recognition at last and for which the costly battle sacrifices were justified. These poems, in tone and movement, differ from earlier lyrics of the middle period, especially his Oriental pieces, such as "Hamatreya," "Saadi," and "Maia" (gathered into *Poems*), as well as "Brahma" and translations from the Persian (by way of a German text) in *May-Day*. "Terminus" acknowledged the limits and boundaries of his own life; "Days" stressed the illusoriness of human existence and the difficulty of knowing what to choose in it. *Society and Solitude* (1870), the last book he himself put together, had much great prose but revealed decreased powers.

Emerson's letters

Perhaps more significant for biography than either the published *Works* or *Journals* was the correspondence—nearly 5,000 letters written by and to him, much of it yet ungathered. At the centre are his communications with relatives, illustrating his affectionate concern and responsiveness; then with neighbours; with the Transcendentalists; with the larger literary fellowship, including Henry Wadsworth Longfellow, William Cullen Bryant, James Russell Lowell, and Whitman, at home, and Carlyle, Walter Savage Landor, John Stuart Mill, and Coventry Patmore abroad. Throughout all these letters his dignity, humour, frankness, courtesy, and wisdom are patent.

When, on July 24, 1872, the upper floors of Emerson's Concord home were destroyed by fire, neighbours saved his manuscripts and books, joining with his many friends throughout the country to raise money for a third trip abroad—to England, France, Italy, and Egypt, the farthest point being the isle of Philae, the supposed resting place of the god Osiris. Welcomed by the press wherever he went, he was met by friends all along the route. Meanwhile, in Concord, his house was restored and a major ovation planned for his return in May 1873.

He continued to make occasional public appearances—most of them near home—accompanied by his daughter Ellen. Between 1879 and 1880, he graced with his presence and an occasional lecture the new Concord School of Philosophy and Literature, founded by Bronson Alcott, his lifelong friend. He died on April 27, 1882, in his 80th year.

Relevance to the mid-20th-century world

Although the once-sharp outlines of Emerson's Transcendentalism were blurred as it became absorbed into the complex American mystical tradition, though the spiritual basis of his individualism has yielded largely to modern existentialisms, and though his recommended cultivation of the inner world of mind tends to be ignored in a collectivist and materialistic society, his humanistic ideals nevertheless are felt to remain fresh, the great problems that engaged him continue to challenge, and his attempted solutions command historic respect. As a writer of majestic prose he has few equals. His significance as a major poet is enlarging. His relevance to the mid-20th-century world is observable in ethical movements, the resurgence of youth, a revived Unitarianism, modern academic skepticism, religious tolerance, and concern with the world of the psyche. The proverbial and anecdotal quality of his writing has made him the most quoted American in the 20th-century press. Translated into dozens of languages, abroad he still represents the America of high goals, character values, and opportunities of which the immigrant dreams. Nineteenth- and 20th-century world figures such as the philosopher Nietzsche in Germany, the philosopher and writer Miguel de Unamuno in Spain, the poet, dramatist, and essayist Maurice Maeterlinck in Belgium, and the philosopher Henri Bergson in France have unhesitatingly confessed indebtedness to him, and the end is not yet.

MAJOR WORKS

POETRY: *Poems* (1846), including "The Sphinx," "Threnody," "The Problem," "The Rhodora," "Woodnotes," "Give All to Love," and "Bacchus"; *May-Day, and Other Pieces* (1867), including "Brahma" and "Terminus."

ESSAYS AND ADDRESSES: *Nature* (1836); "The American Scholar" (1837); "Divinity School Address (first published 1838, published under this title, 1903); *Essays: First Series* (1841); *Essays: Second Series* (1844); *Addresses and Lectures* (1849); *Representative Men: Seven Lectures* (1849, dated 1850); *English Traits* (1856); *The Conduct of Life* (1860); *Society and Solitude* (1870); *Letters and Social Aims* (1876); *Lectures and Biographical Sketches* (1884); *Miscellanies* (1884); *Natural History of Intellect* (1893); *Uncollected Writings* (1912); *Uncollected Lectures* (1932).

OTHER PROSE: *Journals*, 10 vol. (1909–14); *The Letters of Ralph Waldo Emerson*, 6 vol. (1939).

COLLECTIONS: *Journals and Miscellaneous Notebooks*, ed. by William H. Gilman *et al.*, 20 vol. (1960–); *Early Lectures*, ed. by Stephen E. Whicher and Robert E. Spiller, 3 vol. (1959–72); *The Collected Works* (1971–).

BIBLIOGRAPHY. RALPH L. RUSK, *The Life of Ralph Waldo Emerson* (1949), factually the latest and most complete biography, but it doesn't quite capture the many sides of the creative artist and man; OLIVER WENDELL HOLMES, *Ralph Waldo Emerson* (1885), a sympathetic study by a contemporary and a literary man; JAMES ELLIOT CABOT, *A Memoir of Ralph Waldo Emerson*, 2 vol. (1887), important because written by Emerson's literary executor; FREDERIC IVES CARPENTER, *Emerson Handbook* (1953), out of date but still useful; JONATHAN BISHOP, *Emerson on the Soul* (1964), perhaps the most radical interpretation of Emerson, employing close reading and new critical approaches; KENNETH WALTER CAMERON, *Emerson the Essayist*, new ed. (1971), basic for any discussion of intellectual backgrounds; *Index-Concordance to Emerson's Sermons, with Homiletical Papers*, 2 vol. (1963), the appendixes anticipate the kinds of evidence the sermons will eventually provide when they are completely edited; *Young Emerson's Transcendental Vision* (1971), indispensable for an understanding of "Gnothi Seauton," "The Statement of the First Philosophy," and *Nature* (1836); *Ralph Waldo Emerson's Reading: A Corrected Edition with Photographs of Literary Concord, Emerson, and His Family* (1962), especially valuable for the photographs; "Early Background for Emerson's 'The Problem,'" *Emerson Society Quarterly*, 27:37–46 (1962); ERIC W. CARLSON and J. LASLEY DAMERON

(eds.), *Emerson's Relevance Today* (1971); ARTHUR CHRISTY, *The Orient in American Transcendentalism* (1932); EDWARD WALDO EMERSON, *Emerson in Concord: A Memoir* (1889), a moving account of Emerson's home life and participation in Concord affairs, written by his son; MAURICE GONNAUD, *Individu et Société dans l'oeuvre de Ralph Waldo Emerson: Essai de biographie spirituelle* (1964), a study of Emerson by a perceptive Frenchman; LEYLA GOREN, *Elements of Brahmanism in the Transcendentalism of Emerson* (1959); DAVID GREENE HASKINS, *Ralph Waldo Emerson: His Maternal Ancestors, with Some Reminiscences of Him* (1887), sets forth the important dimension of Ruth Emerson's family; VIVIAN C. HOPKINS, *Spires of Form: A Study of Emerson's Aesthetic Theory* (1951), discusses Emerson's employment of Romantic organicism; LEONARD NEUFELDT, "The Vital Mind: Emerson's Epistemology," *Philological Quarterly*, 50:253–270 (1971); SHERMAN PAUL, *Emerson's Angle of Vision: Man and Nature in American Experience* (1952), a discussion of the Law of Correspondence, which lies at the core of Emerson's epistemology; HENRY POMMER, *Emerson's First Marriage* (1967); CARL F. STRAUCH, "The Year of Emerson's Poetic Maturity: 1834," *Philological Quarterly*, 34:353–377 (1955)—all Professor Strauch's articles are important in studying Emerson's poetry; HYATT H. WAGGONER, *American Poets, from the Puritans to the Present* (1968), Emerson is seen as a major poet, central in the development of American verse; STEPHEN E. WHICHER, *Freedom and Fate: An Inner Life of Ralph Waldo Emerson*, 2nd ed. (1971), portrays the modifications of Emerson's idealism.

(K.W.C.)

Emotion

In modern scientific usage, the term emotion has multidimensional referents that include: verbally expressible subjective experiences; concomitant internal physiologic changes; and observable motor behaviour (*e.g.*, facial expression, gesture, posture). These fundamental aspects are richly documented in human oral, written, and artistic traditions.

Emotional processes thus are characteristically complex, and in the early 1970s there was no theoretical frame that accounted for all aspects without contradiction. There are difficulties even in framing a simple definition.

THE NATURE OF EMOTION

While emotion may have subjective dimensions, not every inner experience is called an emotion. Emotions are characterized by visceral change; yet, not every alteration in blood pressure or perspiration is held to be emotional. Such manifestations may reflect adaptation to environmental temperature. People often weep when they say they are sad; yet, such responses can occur when there is local irritation of the eye, or even when the weeper reports happiness. A definition of emotion would have to qualify the relevant phenomena, to note conditions that bring them about, and perhaps to specify their consequences. Such a definition would rest on well-defined mechanisms of emotion that remain to be empirically established, providing grounds for theoretical controversy. Only by mid-20th century had approaches to emotion become relatively comprehensive. Earlier literature shows a variety of limited concepts that differ considerably in language, and in cultural and scientific traditions. Terms that continue to generate misunderstanding include: affect, sentiment, feeling, mood, and passion.

Terms used to define emotion

Affect is frequently used as a synonym for emotion; similar considerations apply to the concept of sentiment. Some use the term feeling to describe experiences that arise in the absence of immediate external stimuli. For example, one may feel sad for no apparent reason (presumably from unobserved changes within the body). Mood also has been defined as being dominantly determined internally, and as having relatively low intensity and brief duration.

The German term *Gefühl* indicates feeling and emotion in a positive, desirable sense. In French, *émotion* seems to refer to sudden, intense, rather negative states of excitement, the positive counterpart of which is *passion*; while *sentiment* is the most general concept, although mainly used in literature. In English, emotion is the most general term, feeling usually being limited to pleasant–unpleasant contexts. Other words connote explicit pathological manifestations (*e.g.*, depression, mania, apathy, anxiety, phobia, fury).

Terminological problems wane as objective (*e.g.*, physiologic) emotional indicators are discovered. With the appearance of each new book in psychology, the term emotion seems less likely to arise; authors seldom devote an independent chapter to the topic, tending to focus on the language of motivation. After the 1940s it seemed that the concept of emotion might dissolve in the complexities of motivational phenomena. There is, however, a common etymological origin for the words emotion and motivation, both being related to the Latin *movere*, "to move."

These semantic changes reflect major developments in research. Investigation into so-called emotional phenomena increasingly is being directed toward objective biological evidence. Nevertheless, the immediate, subjective experiences that people live through (*e.g.*, joy, hate) are undeniably essential to what are called emotional phenomena. An influential theoretician wrote in 1960: "Whether we like it or not, emotion is a 'mentalistic' phenomenon. If its character of experience is excluded, there is not much that remains to be discussed." While investigators may neglect this introspective aspect for temporary, strategic purposes (as in objective motivational studies), the subjective flavour is inescapable. Indeed, experimenters who turned away from introspective analysis of "felt," spontaneous emotion, enriched understanding of experiential qualities beyond that of traditional psychologists who focussed on the content of awareness.

INITIAL CONCEPTIONS OF EMOTION

Experimental analysis of introspective reports. Psychology as an experimental discipline developed during the second half of the 19th century. Its goal was to relate objectively definable, physically measurable environmental changes to the content of consciousness as it was available to self-observation (introspection). Wilhelm Wundt (1832–1920), called the father of experimental psychology, based his theory of emotion on an analysis of introspective reports of feelings (*Gefühlston*) generated by sensory stimuli. He concluded that emotions are describable in terms of three dimensions represented by pairs of contrasting words. Wundt's most fundamental dimension (*Lust–Unlust*) recurs in modern work as pleasant–unpleasant; another is translated as tension–release (*Spannung–Lösung*); and a third as excitement–relaxation (*Erregung–Beruhigung*). Only the first seems to have endured well; historically it is the oldest, being based on antecedents often identified as hedonic (see MOTIVATION). The other two, however, also touch on factors that reappear in empirical emotion research.

Wundt reported that near the sensory threshold, one perceives any stimulus in an emotionally neutral way; that, at higher intensities, pleasant and eventually unpleasant emotional qualities appear. Partial evidence for such a relationship later was reported, mainly in terms of taste (gustatory) stimuli (in studies of spontaneous food selection, through introspective reports, and by evidence of neural activity in gustatory experiments with humans and laboratory animals). Sensory and hedonic (pleasant–unpleasant) intensity in tasting salt solutions are not related in a straight-line fashion. When sensory intensity is smoothly varied, hedonic intensity abruptly changes at different levels of salt concentration.

Hedonic theories of emotion

The relation of stimulus intensity to affective quality appeared in more complex forms in later hedonic theories of emotion that added other cognitive factors. The effect of intensity in these theories is seen to depend on the perceiver's level of adaptation (or on his expectation of the stimulus effect). It is held that a relatively small discrepancy from anticipated stimulus intensity creates a pleasant effect, and that larger discrepancies are unpleasant; reduction of the discrepancy is said to lead to positive emotional experience, its persistence producing negative affect.

Evolutionary conceptions of emotion. Charles Darwin's *Expression of the Emotions in Man and Animals* (1872) had a profound influence in systematizing emo-

tion research. In his principle of "serviceable" association of habits, he suggested that emotional expressions are evolutionary remnants of previous adaptive behaviour, persisting (even uselessly) in a mild form (*e.g.*, snarling as a sign of aggression). In his principle of antithesis, concerning the multiplicity of emotional expression, he posited that "habitual actions" serviceable to "certain states of the mind" were replaced by contrary (antithetical) movement patterns that automatically tend to appear during the "opposite state of mind," even if they have no adaptive function. Darwin was suggesting that even "useless" emotional expression (and, consequently, its subjective correlates) derives from adaptive functions. Darwin's influence clearly can be seen in a later trend that dealt with emotions as a group of behaviour-controlling *processes* rather than as "mental" entities.

Patterns of emotional expression. In keeping with his initiative, some of Darwin's contemporaries began a continuing research effort toward understanding and classifying facial expressions and other movements that reflect emotional activity. A belief in universal patterns in human emotional expression is neither new nor entirely unfounded. Comic and tragic masks in European and Asian theatre provide ancient evidence.

In such research, statistical analysis is made of posed or active emotional mimicry by subjects; specific facial expressions are depicted in drawings, paintings, and photographs; and expressive behaviour is directly observed and evaluated. Some data provide cross-cultural comparisons, or concern emotional expression in the blind and among people of different ages.

Facial expression

Although partially validating some fundamental expressions (such as laughing or weeping) these investigations have not offered convincing evidence that mimicry in itself is a dependable indicator of subjective emotional states. Judgments of observers are clearly suspect.

Systematic analysis of criteria for evaluating emotional expression led U.S. psychologist Robert S. Woodworth (1869–1962) to devise a six-part scale: (1) love, happiness, mirth; (2) surprise; (3) fear, suffering; (4) anger, determination; (5) disgust; (6) contempt. Studies based on this scale led to a model that incorporated three emotional dimensions: pleasant–unpleasant, attention–rejection, and intensity. Such models are of uncertain value because of differences in cultural influence, in individual facial anatomy and muscular mobility (*e.g.*, deadpan), in tendency to dissimulate (*e.g.*, the false smile), and in emotional expression at different age levels.

Nevertheless, observation of the blind and studies of electrical stimulation to the brain among laboratory animals demonstrate distinct patterns of mimicry, gesture, posture, and vocalization in fundamental emotional phenomena (such as those interpreted as fear and anger). Among newborn infants (humans and anthropoid apes) indications of general excitement are about all that can be observed. Specific emotional expression (*e.g.*, smiling) first appears about the third month among human infants. Mimicry later becomes more differentiated, and it is characteristic of children aged three to five to recognize and readily produce mimicry related to joy; but they tend to have real difficulty with expressions of surprise and contempt. Emotionally expressive mimicry by blind adults is relatively poor, demonstrating that emotional expression is learned as well as genetically mediated.

Mimicry in infants and blind people

Efforts to interpret gestures, postural reactions, and vocalizations that accompany facial expressions still rest heavily on everyday observations with only limited scientific evidence available. It has been suggested that behind every interpretation one makes of emotional expression in others there lies a complicated, unreliable, intuitive act; just how unreliable such estimates can be is well demonstrated by prejudice and social misunderstanding.

Despite modest progress, efforts to classify emotional activity on the basis of expressive movements cannot be considered successful. Such activity is not as homogeneous as the simplified, abstract language of classification might suggest. Traditional terms based on manifest behaviour barely begin to reflect the transient, contradictory nature of emotional activity. The common confusion was in error; neurophysiological evidence suggests that such experiences may alternate quite rapidly.

THE PHYSIOLOGY OF EMOTION

Emotion as bodily sensation: James–Lange theory. William James (1842–1910) stated that his own theory of emotion would be as difficult to verify as to refute: "The best thing I can say for it is, that in writing it, I have almost persuaded *myself* it may be true!"

Evidence from studies of perceiving and of brain function (as well as his own introspections) led James to conclude that emotions are perceptions of changes within the body itself: ". . . the bodily changes follow directly the *perception* of the exciting fact . . . our feeling of the same changes as they occur *is* the emotion." According to James: "An object falls on a sense-organ and is apperceived by the appropriate cortical centre; or else the latter, excited in some other way, gives rise to an idea of the same object. Quick as a flash, the reflex currents pass down through their preordained channels, alter the condition of muscle, skin and viscus; and these alterations, apperceived like the original object, in as many specific portions of the cortex, combine with it in consciousness and transform it from an object-simply-apprehended into an object-emotionally-felt."

This abstract formulation may not convey the unexpected, shocking effect it had on everyday thinking. In its already classic popularization it meant that one does not run from an unexpected encounter with a bear out of fear; one is frightened *because* of running away.

James supported his postulate with excellent examples of major sensory impressions that evoke changes mediated by the autonomic nervous system (*e.g.*, blushing) evident even to naïve observers. A Danish physician, Carl Georg Lange (1834–1900), stressed changes in the circulatory system in his own version of the theory. It seemed to both James and Lange that emotional qualities were the result of perceived changes in bodily activity that followed sensory impressions. An enduring line of research has indeed emerged from this theory.

Visceral and chemical patterns. Beyond the James–Lange theory, elucidation of the structure and function of the autonomic nervous system marked a major contribution to the study of visceral correlates of subjective emotional states. Clear evidence was found of neural control over most viscera through the two autonomic divisions: sympathetic and parasympathetic. The sympathetic division functions through thoracic and lumbar spinal nerves; parasympathetic activity is mediated by brainstem structures (cranial nerves) and those of the sacral region of the spinal cord. In some organs the two have opposite effects (*e.g.*, heart rate is slowed by the parasympathetic and accelerated by the sympathetic). Its diffuse connections, however, make the sympathetic division more suited for general effects, while the parasympathetic tends to mediate selective, localized activity. Both systems can inhibit and excite; for example, sympathetic effects usually are to inhibit digestive function and selectively to dilate or to constrict blood vessels.

U.S. physiologist Walter B. Cannon (1871–1945) drew attention to the close relationship between emotional activity and sympathetic function. According to his emergency theory of emotion, the sympathetic division of the autonomic nervous system helps the organism face stress. A cat that faces an aggressive dog responds with increased heart rate and breathes more rapidly. In the cat's eyes, pupils dilate and blinking (nictitating) membranes retract; digestive function is inhibited; hair stands up; perspiration appears on the pads of the feet; the number of red cells in the circulating blood increases through contraction of the spleen; adrenal-gland activity is stimulated; glucose is mobilized into the bloodstream from glycogen storage in the liver. These symptoms all reflect sympathetic activation; Cannon's classical suggestion (1915) was that they increase resistance to stress and enhance defensive abilities (fight or flight).

Cannon's emergency theory

Nevertheless, a growing experimental literature indicates that Cannon's focus on the sympathetic division was too narrow. Among humans and other animals parasympathetic emotional effects also appear. Defecation and urination by frightened soldiers are under parasympathetic control. Among monkeys and human infants, a sudden auditory stimulus may produce parasympathetic penile erections, besides eliciting a distinctive, generalized bodily reaction (startle response).

Blood-sugar levels increase in response to sympathetic adrenal activation in rats after sudden, loud noise, and in cats frightened by a dog. Yet, after surgical removal of the adrenal glands, the same stimuli trigger pancreatic insulin secretion to decrease blood-sugar levels. This effect can be abolished by cutting the animal's (parasympathetic) vagus nerve (that initiates insulin production). Clearly, both autonomic sectors are active in marked emotional states. This and similar evidence casts doubt on Cannon's notions that sympathetic (or adrenaline) effects are the hallmark of emotional arousal and that they always are adaptive. Excessive elevation in heart rate, for example, will produce a maladaptive decrease in the volume of discharged blood. While adrenaline optimally mobilizes sugar stored in the liver as glycogen to decrease muscle fatigue, in larger doses it actually decreases blood flow in the functioning muscle.

Adrenal function

Emotional activity clearly shows characteristic combinations of sympathetic–parasympathetic effects. The earlier assertion that such visceral changes are essentially quantitative, and therefore unsuitable for distinguishing emotional qualities, has become less tenable. Bodily evidence of fear (or anxiety) gives promise of being discriminable through rigorously controlled psychophysiological experiments; electronic devices permit remarkably complex observation, and improved methods of hormone detection and quantification have become available.

During intense fear one's face usually is pale; it flushes during anger. Similarly, in chronic cases of gastric ulcer, when the lining of the stomach was observed directly, gastric secretion was inhibited during fear and the mucous membrane blanched; when the patient seemed angry or happy, the lining flushed. Similar findings have been reported for nasal and vaginal membranes.

The sympathetically controlled medulla (core) of the adrenal gland produces the hormones epinephrine (adrenaline) and norepinephrine (noradrenaline). Each hormone has different cardiovascular effects; from these it has been shown that people who exhibit fear react by secreting epinephrine; when they manifest anger or hostility, they produce norepinephrine. This is supported by findings that the adrenal glands of aggressive predators (such as dogs and cats) contain norepinephrine, while the fearful rabbit's contains epinephrine.

Monkeys under sudden, marked stress (apparently uncertainty or fear) showed increased blood-plasma levels of both hormones from the adrenal medulla, and of other substances (17-hydroxycorticosteroids) produced by the gland's cortex. When a monkey was prepared by a warning signal, however, epinephrine level remained normal, corticosteroids decreased and norepinephrine increased. Stimuli that elicted no signs of fear or anxiety tended to decrease blood pressure and pulse rate (parasympathetic effects). In other research, a quantitative relationship was reported between peripheral vasoconstriction (narrowing of blood vessels) and estimated degree of anxiety.

Serotonin, a substance found in many parts of the body, may have immediate emotional effects. An increase of serotonin that follows introduction of its biochemical precursor into the brain generates signs of ecstasy and euphoria. Drugs that imitate the effects of serotonin have analogous chemical structures and produce excitement, hallucination, and agitation; drugs that are serotonin antagonists tend to generate feelings of depression. Psilocybin (found in certain mushrooms), used in Yucatán for centuries to induce trance states, and LSD (lysergic acid diethylamide) are serotonin-like. Antagonists include the tranquillizers reserpine and chlorpromazine. Many hormones (*e.g.*, sex hormones and thyroxine) have significant influence on emotional function (see also HALLUCINOGEN).

Serotonin

Physiological detection of emotion. *The galvanic skin response (GSR).* A most sensitive indicator of sympathetic activity level is the galvanic skin response (GSR), also known as the psychogalvanic reflex (PGR), or the electrodermal reflex (EDR). The fluctuating electrical resistance (or conductivity) of any part of the body may be measured through electrodes on the skin (*e.g.*, on the palm and back of the hand). GSR fluctuations occur in response to such influences as frightening stimuli and are usually quick and transient; there also are more-or-less regular, slow changes during the course of a day. GSR latency (time lag) and voltage have not proved reliable indicators of specific emotional qualities, probably because they are too sensitive. Stimuli perceived as pleasant or as unpleasant can change GSR. In problem-solving tasks, subjects first tend to show elevated GSR activity; this decreases as tension dissipates with learning. Such changes usually are attributed to the subject's general level of physiological activity or to his degree of attention, as well as to his emotional intensity. Used with other indicators, the GSR has been helpful in studying emotional arousal or mood. Its doubtful use in lie detection presumably depends on sympathetic changes that accompany feelings of guilt (see below *Lie detection*).

Respiratory signs. Although respiration is partly under the control of autonomic brainstem centres, it also can be voluntarily influenced. Respiratory patterns once were expected to discriminate among what were thought to be primary emotional processes: enjoyment, pain, anger, astonishment (surprise), fear, and disgust. These hopefully were to be reflected in the changing inspiration/expiration ratio (I/E), based on the time the subject took to inhale and to exhale. It also was thought that specific I/E ratios could indicate lying as well, but well-controlled experiments dimmed most of these hopes. Weeping, however, can be differentiated by its sobbing interruption of inhalation, and in laughing, exhalation is spasmodic. Both patterns are observed as involuntary symptoms among those with injuries in the nervous system.

Muscle tone. Muscle tone, the degree to which a group of muscles is contracted, is another peripheral and general indicator of excitement, fear, and anxiety. Neck muscles exhibit particularly characteristic electrical changes (muscle potentials).

The startle reflex to unexpected, intense auditory stimulation usually is considered to be an emotional phenomenon. This complicated response involves skeletal muscles (particularly flexors) and viscera. First there is blinking (with a latency of about .04 of a second after stimulation); next the skeletal muscles react after a lag of about 0.1 of a second; after about one second, more complex signs appear (*e.g.*, GSR, blood pressure elevates, pulse rate quickens). In this hierarchy of latencies, some theorists see a distinctive sequence of the central nervous system.

Startle reflex

Lie detection. Word-association techniques have been used in criminal interrogation combined with physiological indicators of emotion. One is asked to respond to a number of words (some related to the given crime) with the first word he thinks of each time. Emotional tension elicited by critical words is assumed to be indicated by verbal hesitation and by such indicators as respiration, blood pressure, and GSR. Forms of a device (polygraph) are commercially available for such recordings. Most experts judge accuracy to depend heavily on the interrogator's experience and to be subject to serious error.

Soviet psychologist A.R. Luria suggested in 1932 that involuntary finger movements reflect emotional aspects of lying more sensitively than does word association. In the Luria method, one presses a key that registers pressure at each of his verbal associations. The pressure pattern is supposed to reflect increased muscle tension during emotional conflict.

Value of peripheral phenomena in classifying emotional states. Peripheral signs fail to give unequivocal support for the James–Lange hypothesis. Most visceral and chemical (humoral) indicators simply reflect a quantitative,

general level of emotional tension (although there are hormonal and vasomotor signs characteristic of fear–anxiety and joy–anger–resentment). The viscera and central nervous system also have routine regulatory functions that can cancel each other. Despite use of the polygraph to record complex peripheral responses simultaneously, chances are poor for developing emotional classifications based on visceral responses alone.

Further caution is indicated by work showing that the brain can mediate conditioning of visceral reflexes more precisely than once was assumed (particularly in autonomic change). Indeed, highly specific visceral and vascular reactions (as in high blood pressure) have been controlled experimentally by conditioning.

Cognitive effects on response to adrenaline

Apparently contradictory evidence regarding the James–Lange theory also comes from response to injections of such agents as adrenaline. Earlier experimenters reported genuine emotional effects in only a fraction of such cases. Carefully designed later experiments with humans suggested that the emotion-mobilizing effect of epinephrine varies considerably with the subject's cognitive state. (Some of the experimental subjects who watched live filmed slapstick antics were informed of the bodily experiences to be expected from their epinephrine injections; others were not.) It was concluded that the expectations of the individual are decisive in experiencing emotional qualities induced by epinephrine. Injection is not necessarily equivalent to normal physiological excitement, however; and adrenaline has central effects (in the brain and spinal cord) as well as peripheral influence.

INSTINCT AND BEHAVIOUR

Instinct theories. Darwin's view that emotionally expressive activities are evolutionary remnants and James's notion that emotion is organic sensation supported efforts to interpret these functions as instinctive. Instincts are defined as genetically fixed patterns of activity, and it is critical to differentiate them from acquired behaviour. While British psychologist William McDougall (1871–1938) attempted to specify an appropriate emotion for every instinct, his list has shrunk considerably in modern versions. One of these reduces the array to the following: incorporation, rejection, destruction, protection, reproduction, deprivation, orientation, and exploration. In this model, emotional intensity is said to vary within these eight qualitative dimensions. Such speculative models neglect the quantitative effects of cognitive, energizing factors in emotion. For similar reasons the concept of instinct came on hard times in the first half of the 20th century.

About that time a distinct research movement (ethology) was developing in Europe. Avoiding the speculative aspects of earlier instinct theories, ethologists provided evidence for instinctive behaviour among animals in the wild. Their strict empirical data pointed to instinctive functions not clearly manifested under rigid laboratory conditions. A territorial instinct, for example, seems well supported and could help account for such emotional behaviour as aggression, fighting, courage, cowardice, and pride. Austrian ethologist Konrad Lorenz assumed a close relationship between emotion and instinct, seeking evidence of human instincts in verbal reports of emotional experience. Uncertainty in such an undertaking stems from evidence that instinctive behaviour in humans is greatly modified by learning. Ethologists seem to face the formidable problem of distinguishing these two factors in every individual case. Nevertheless, unless all human behaviour results strictly from learning, instinctive tendencies merit attention.

Psychoanalytic theorists (S. Freud, C.G. Jung) assumed a close relationship between instincts and emotions, but failed to provide a systematically complete account. Their influential, but indirect, effect on research efforts is discussed under MOTIVATION.

The behavioral approach to the study of emotion. Such extreme behaviourists as U.S. psychologist John B. Watson (1878–1958) held that subjective aspects are to be disregarded; that only objective emotional behaviour is measurable and, therefore, significant. Instinctive behaviour was viewed as an unconditioned chain of activity, fundamental emotional reactions being described as unconditioned reflexes. An impressive finding of behaviouristic research was that emotional behaviour could be conditioned to appear on signal. Watson named fear, rage, and love as the basic emotions, and showed how readily they could be elicited from children, rabbits, or other animals by a conditioned (initially neutral) auditory or visual stimulus.

Emotions as conditionable reflexes

Among more sophisticated, later representatives of behaviourism, emotional intensity gained a central position. The U.S. neobehaviourist B.F. Skinner in 1953 wrote: "We define an emotion—insofar as we wish to do so—as a particular state of strength or weakness in one or more responses induced by any one of a class of operations."

MOTIVATION AND EMOTION

By the 1930s the term instinct was largely being replaced by the language of drive and motivation, so much so that it was being asked whether what was coming in the back door had not just been thrown out the window. With the word motivation, psychologists indeed faced some of the earlier difficulties they had with the notion of instinct. Laboratory studies indicated that stimulus and response were separated by a relatively unobservable process that itself is a condition of learning. This intervening process (represented by such words as hunger, thirst, fear, or sexual desire) was not considered a special problem by classical reflexologists who saw it in terms of complex, unconditioned chains of instinctive reflexes. As a subjective emotional experience, it was held by behaviourists to be unworthy of attention. Yet, when instrumental or operant conditioning to rewards and punishments came to be widely studied, what had been called instinct seemed to loom as central problems once more.

The concept of drive. Drive (motivation is a more general synonym) was defined as a characteristic state of excitement that appears under physioloigcal (homeostatic) imbalance. An organism deficient in some nutrient, for example, shows specific food-seeking activities. Drive states were defined as intervening variables (or hypothetical constructs) to be inferred from observable antecedents (*e.g.*, food deprivation) and consequences (*e.g.*, food seeking). Consequences were held to be systematically manipulable through antecedents, and the critical factor in instrumental learning (or reinforcement) was believed to have been found in drive reduction. Evidence was offered to show that originally indifferent stimuli, if consistently presented when reinforcement occurs, can become reinforcing and drive-inducing (incentive, cue) factors themselves. Attention was thus focussed on the role of learning in establishing central motivational states that influence behaviour generally. The motivational process was conceived in the 1970s as a complex of neural and humoral changes that reflect learned tendencies; the concept seemed to be widening to include specific anatomic–functional endowments of the central nervous system. This superficial description is intended to aid in understanding a growing debate that began in the mid-1930s with the question of emotions, appeared to reach a peak in the early 1950s, but still continued on into the 1970s.

Drive states

Energy and emotion. To some psychologists most of the empirical and assumed characteristics of motivation coincided with those of emotion. Elizabeth Duffy, a pioneer in motivational interpretations in the U.S., argued convincingly that the concept of emotion was no longer justified in a scientific context. She held that definitions of emotion describe nothing qualitatively different from other activities or response patterns, and that so-called emotional activity was simply a quantitative change in the energy level of the organism. Duffy interpreted the associated subjective experiences merely as the conscious aspect of adaptive function, noting that "emotion" is: ". . . the individual's response to situations which promise well or ill for the attainment of his goals." She held that relatively slow, mild, and consistent

Emotion as change in energy level

changes in energy level produce subjective states corresponding to the old notion of "mood," and that rapid, intense changes generate "emotion." She noted that energy level (including its introspective aspects) varies along a continuum so that there is no fixed point at which a "nonemotional" level suddenly becomes "emotional."

The last notion seems most significant and enduring. Essentially, it denies that emotion is an independent psychic entity, suggesting that it is not even a special quality that appears at a definite level of physiological intensity. On the whole, Duffy convinced many investigators that "emotion" is not to be separated from motivation; that the objective, experimental approachability of the latter should provide a sufficient basis for dispensing with the concept of emotion.

It is difficult, however, to understand Duffy's discussion of the conscious aspects of adaptative responses: "I think it will be agreed that the individual does not experience 'emotion' except in situations which are of significance to him." This apparently means that, if the environment does not create "significant" energy change, there is no "emotion." But it can be argued that the energy factor is not a state but an active process, even if it tends toward equilibrium; and that, consequently, in its integrated form, an emotional base is always given. It is difficult to understand how people can indifferently experience stimuli that do not generate perceptible equilibrium changes, since such changes should not be perceived at all. If perceptible changes are imagined that do not have a motivating consequence (*e.g.*, what William James called "cold" perceptions), Duffy's postulated continuum is broken, and emotional quality could be said to emerge only at a special point in the continuum. At any rate, Duffy gave no definite answers to such questions. The problem they present remained unsolved by the efforts of U.S. psychologist Robert W. Leeper to challenge the traditional belief that emotional activity is disruptive.

Emotion as disruptive

Organizing and disorganizing emotional effects. Leeper's work created stormy debates, since it seemed to contradict folk wisdom and obvious, everyday evidence of disorganizing effects on behaviour during fright or blind anger. Leeper argued that the essential function of emotion is to be sought in its *organizing* effects. Many of his critics held that emotion can be either organizing or disorganizing, depending on the situation. They seemed to be rapping on an open door, however, since Leeper himself already had considered the reorganization necessary for adaptation to be a simultaneous disorganization of earlier, inadequate behaviour. He emphasized that disorganization is not the *essence* of emotion. His critics seemed to miss the point, since Leeper's argument was with any approach that ripped subjective activity apart by offering cognitive, affective, and conative distinctions. The interpretation of emotions as disruptive endorses the view that human emotion (affective) is only a weakness, deriving from a brute animal heritage that is pitted against rationality (cognitive) and will (conative).

Despite the attractiveness of arguments of the sort offered by Leeper, experimental efforts to apply the motivational interpretation of emotions seem to be returning to the traditional view that was under challenge. Early in his work Leeper himself emphasized so-called higher, secondary, generally emotional factors, rather than motives deriving from homeostatic biological processes; his later writings showed this tendency even more.

In a brilliant psychological analysis of Sir Ernest Henry Shackleton's South Pole expedition, he argued similarly in favour of the decisive superiority of higher motivations (*e.g.*, achievement) against primary biological drives. He offered the critical conclusion that ". . . emotional processes basically and fundamentally are perceptual processes."

The richness and differentiation of perceptual function indeed sheds light on the variety of flexibility of human emotions. What is of theoretical concern is that the perceptual conception offered by Leeper seems inconsistent with the emotional-motivational continuum he (and Duffy) so unflaggingly defended. When he wrote of "motivationally neutral perception," he seemed to be arguing for discontinuity; clearly, he endorsed the possibility of unmotivated or emotionless perception. The possibility that this may be a return of the concept of "cold" (William James) highlights the critical and consistently recurring problem of perceptual function in emotion.

PERCEPTION AND EMOTION

Objections to the James-Lange theory. Although his emergency theory supported James's hypothesis that emotion is organic sensation, Cannon later categorically rejected that hypothesis: "(1) Total separation of the viscera from the central nervous system does not impair emotional behavior. (2) The same visceral changes occur in very different emotional states and in nonemotional states. (3) The viscera are relatively insensitive structures. (4) Visceral changes are too slow to be a source of emotional feeling. (5) Artificial induction of the visceral changes typical of strong emotions does not produce them."

Though it seemed that these critical statements of Cannon's would end the debate on the James–Lange theory, James's prophecy that it would be as difficult to reject as to verify his hypothesis was to prove stubbornly viable. British Nobelist C.S. Sherrington (1857–1952) had presented impressive experimental support for Cannon's first, apparently most telling, objection. Despite such measures as surgical severing of the spinal cord (at the neck) and of the parasympathetic vagus nerve, ostensibly "isolated" heads of laboratory animals consistently continued to show a full range of emotional expression.

It was argued that laboratory animals cannot give introspective reports. People with severed spinal cords, however, report a full range of emotional experience.

The impact, however, is considerably reduced when it is realized that visceral impulses still can reach the brain via intact cranial nerves. Humoral (chemical) information also continues to be carried by blood from lower parts of the body (although this represents a relatively sluggish process).

Cannon's principal objection to the James–Lange theory is that animals without peripheral feedback will respond emotionally only to stimuli to which they had earlier learned to respond. Theoretically, once emotional expression is conditioned, visceral impulses are no longer necessary; but this assumption has defied experimental verification.

Even if there is no absolute correlation between visceral–somatic activity patterns and specifically experienced emotional qualities, the possibility of some relationship cannot be excluded with certainty. In summary, most of Cannon's arguments seem questionable; however, consider now his statement that visceral alterations are too sluggish to be an immediate source of subjective emotional experience.

At the turn of the century, attempts to measure the speed of visceral responses indeed indicated that they proceeded too slowly to be held responsible for subjectively "felt" emotions. But there was considerable reluctance to reject the James–Lange theory on these grounds since the accuracy of introspective accounts was not trusted. There is, however, one consideration of the time factor that avoids this difficulty, once and for all showing that peripheral theories of the genesis of emotions have become untenable.

"Cold" and emotionally "coloured" perception

James's abstract description of neural mechanisms tends to oversimplify and harden emotional processes into states. If the events he described be reconsidered, this time with special attention to the time factor, the environmental stimulus that initiates definite changes in the brain cortex then becomes the first event. This cortical activity represents what James called a "cold" perception, subjectively an awareness of the stimulus devoid of affect, though the existence of such a discrete state has been challenged. Next, the pattern of neural excitation was held to flow from the cortex to other parts of the body (the periphery) to produce an appropriate bodily (somatic) response. Secondary impulses feeding back

from somatic receptors to reach initially excited parts of the cortex were said to lend emotional "colouring." James was asserting separate perceptions rather than an essentially unified process: a "cold" perception that reflects external stimulus relationships, followed by an emotionally "coloured" perception.

In itself, the assumption of a "cold" perception contradicts James's repeated statements, in other contexts, that perceiving cannot be deprived of emotional elements. The contradiction might be resolved by asserting that the final form of perception develops in two phases. A much more serious objection arises, however, when it is asked if the environmental stimulus excites an initially indifferent, quiescent cortex. Clearly, the cortex is affected by a constant stream of impulses coming from other parts of the body. It can justly be asked whether this continuing process does not "colour" the effect of every incoming environmental impulse. This is possible, but obviously would not fulfill the requirements of James's hypothesis, since it involves visceral responses to other, earlier stimuli. It could also be imagined that any residual cortical excitement would always be erased by the immediate environmental stimulus, but this would wipe out traces of any equally significant earlier stimulus, change the experience to a "cold" perception, and leave "emotional gaps" in the perceptual process. By this assumption, emotional components could not develop during any rapid sequence of incoming stimuli. The only alternative to this would seem to be that the environmental impulse is emotionally "coloured" immediately on arrival, perhaps because the incoming impulse interacts with earlier cortical processes that reflect continuing visceral–somatic states, or because the environmental impulse on its way to its specific cortical projection area may stimulate brain tissues that represent the viscera. It can be assumed that the initial impulse can travel either toward the cortex or toward the periphery. The peripheral flow would elicit characteristic visceral–somatic responses while the other would stimulate the cortex with minimal delay to lend "colouring" to the already activated sensory projection area. Even if this makes two perceptions plausible, both would be emotionally "coloured." The dominant bodily activity at the moment the environmental impulse directly stimulates the cortex "colours" the effect of that impulse, while the second lends "colour" by altering the peripheral response. This happens much faster than James assumed, since the development of peripheral reverberations *de novo* is not needed. These assumptions coincide with empirical measurements that show subjective emotional response to precede induced visceral change.

Cannon's thalamic theory. Many points of the interpretation just given have been experimentally verified. The interpretation failed to gain explicit expression in Cannon's time when much less was known about expressive emotional patterns and about visceral representation in the central nervous system. Nevertheless, Cannon's (1927) thalamic theory of emotion rested on similar considerations. (The thalamus is the prominent structure of the diencephalon; a major relay centre, it integrates the periphery, other subcortical regions, and the cortex.)

Experimental and clinical observations had led to the conclusion that the integration of fundamental emotional processes is mediated by subcortical brain structures. By the end of the 19th century it had been found that animals surgically deprived of cerebral cortex can show marked emotional symptoms (*e.g.*, anger, attack). Particular attention had been drawn to the human thalamus by emotional aberrations associated with circulatory disturbances or tumours there. Unilateral thalamic injuries were known to intensify sensitivity on the same side of the body, heightening both pleasant and unpleasant sensory experiences and disturbing emotional expression (*e.g.*, producing involuntary crying and laughing). In view of such evidence, the anatomical position of the thalamus and its extensive cortical connections persuaded Cannon to concentrate on this centre in seeking the central integration of emotion. Specifically, he posited that environmentally produced impulses moving toward the cortex simultaneously excite appropriate thalamic regions. Those resulting thalamic impulses that flowed to the cortex were said to provide the perceptual "colouring," the other thalamic impulses serving to excite visceral and somatic expressive emotional patterns. A significant element of Cannon's theory (especially attractive to psychologists) is that of inhibition of the thalamus by the cortex to control the emotional component (unless followed by an appropriately intense external stimulus).

Criticism of Cannon's theory

In addition to enthusiastic support, Cannon drew criticism that largely objected to selecting the thalamus as the centre of emotional activity. Critics observed that many expressive symptoms attributed to the thalamus (*e.g.*, compulsive crying and laughing) can be produced by injuries to other points in the nervous system. Besides, most people with such injuries reported no subjective emotional experiences appropriate to the symptoms; for example, one might cry without feeling sad. The existence of a portion of the neural pathways postulated by Cannon was also contested. Accurate on many points, these objections were not so much to Cannon's general theory as to the anatomical bases he specified. Cannon's basic hypothesis can be considered a precursor of activation theory in which subjective and expressed emotions clearly were held to be integrated by the central nervous system. In this sense, Cannon's theory continues to be endorsed.

ACTIVATION THEORY

Pathologies such as brainstem lesions in human encephalitis (producing disorders of wakefulness and sleep) and emotional symptoms associated with circulatory disorders and subcortical tumours served to draw attention to the integrative role of the midbrain. Improved stereotactic devices for accurately placing electrodes and cannulas deep into the brain made it possible to study these regions in freely active animals. Permitting focal electrical and chemical stimulation, circumscribed tissue destruction, and detection of spontaneous, local electrical activity in the brain, these methods in the 1930s began an avalanche of discoveries.

Sleep and wakefulness were found to have readily distinguishable spontaneous electrical patterns in the brain. People in deep sleep, for example, tend to show a high-voltage, slow pattern; during relaxed wakefulness lower amplitude (or voltage) and more rhythmic, higher frequency is characteristic. Among cats with their cerebrospinal axes severed at different points, such electroencephalographic indicators show that intact pathways connecting the mesencephalic brainstem to anterior brain structures are indispensable for wakefulness.

Electroencephalography

Local electrical stimulation has established that the brainstem's central, netlike structure of gray matter (reticular formation) is responsible for wakefulness (rather than sensory pathways on the edges of the mesencephalon). Stimulation of the reticular formation in slightly anesthetized animals induces throughout the cortex fast, low-voltage waves (so-called desynchronized activity) characteristic of wakefulness. Selective destruction of the reticular region results in enduring sleeplike coma.

Reticular electrical activity indicates that many pathways to the cortex send collateral fibres into the reticular formation. Collateral impulses in the netlike structure contribute to diffuse reticular excitement that passes via the thalamus to activate the whole cortex. The abbreviation ARAS (for ascending reticular activating system) has taken firm root in the investigative literature.

Wakefulness thus arises not only from direct cortical stimulation by specific sense receptors but also from reticular function that integrates receptor effects to influence the entire cortex. On this basis inferences have been made about emotion: (1) The functional continuum associated with the ARAS includes all organized psychological processes from sleeping to extreme states of excitement; consequently, it includes emotion. (2) The basic functions of the ARAS (activation, arousal) coincide with those considered most typical of emotion in motivational theories. Consequently, this continuum of activation also

is a motivational-emotional continuum. (3) Impulses arising from both the internal and external environment act to excite the ARAS, and the activating effect of the ARAS is a precondition for the integrative function of the cortex. It follows that even the most elementary perceptual effect is emotionally "coloured" from its very inception.

These three considerations are the basis of the activation theory of emotion developed by U.S. psychologist Donald B. Lindsley (1951), according to whom activation is a continuum along which degrees of intensity may be measured alternatively as visceral–somatic responses (peripheral effects) or as central electroencephalographic responses. At one end of the intensity continuum are the first arousing, orienting reactions of attention; the other extreme is characterized by major signs of emotion.

Activation theory has been well-supported and extended by experimental evidence. For example, the activation effect, once believed to be restricted to the reticular formation, involves other parts of the brain, including several structures of the limbic system (see below *The limbic system*).

Autonomic function in homeostasis

Anatomically the activating mechanism embraces the autonomic nervous system (also called the vegetative system), characterized as controlling the stability (homeostasis) of the internal milieu. Autonomic function is under the influence of neural impulses from the viscera and also of humoral factors such as metabolic waste products and hormones. Beyond homeostasis, these factors have general activating, arousing effects as well. At the same time, autonomic function affects the readiness of skeletal muscles to react. Complicated stimulating and inhibiting effects on this musculature arise chiefly from a specific portion of the spinal outflow (the so-called gamma-fibre loop). Apart from this, sharp patterns of emotional expression can be elicited from the activation system, particularly the hypothalamus.

Postulated effects of activation on discriminative cortical function have been verified. Mild reticular electrical stimulation reduces the time required by monkeys to discriminate among briefly flashed pictures. Stimulation of specific points on the cortex can result in diffuse reticular activation, although the cortex on the whole tends to inhibit reticular function. The latter mechanism amounts to a negative feedback loop from the reticular system to the cortex and back. This regulatory effect indeed could correspond to the cortical inhibition postulated by Cannon.

Objections to the activation theory of emotion stem from considerations already known to its author when he wrote: ". . . activation theory appears to account for the extremes, but leaves intermediate and mixed states relatively unexplained as yet." Activation theory indeed fails to explain individual emotional qualities, tending to clarify negative rather than positive emotions. The notion of moderate activation as pleasant and of intense activation as unpleasant seems simplistic since both negative and positive emotions show extremely wide ranges in level of excitement.

Despite criticisms of activation theory, it is still argued that the basis for individual emotional qualities should be sought in the details of the activation system. Nevertheless, some psychologists assign primary, determining emotional function to cognitive factors.

REGULATORY THEORY: EMOTION AS RELATED TO HOMEOSTATIC MECHANISMS

Hypothalamic activity. Below the thalamus, the hypothalamus forms part of the walls and floor of a brain cavity, the third ventricle. This relatively small hypothalamic structure is a complex receptor area, remarkably reflecting the state of the internal milieu by reacting to humoral and neural stimuli. It functions homeostatically through its dangling pituitary (the so-called master endocrine gland) and through its efferent neural connections; *e.g.*, helping to regulate blood circulation, body temperature, fat metabolism, and sexual function. Hypothalamic impairment can severely and chronically disturb such functions. This bit of brain also is a substrate for marked emotional expression (*e.g.*, fear, anger). The question of a connection between hypothalamic regulatory and emotional function seems fundamental.

The striking evidence is that every apparently distinct regulatory function is mediated by a structurally discrete, functionally contrasting pair of hypothalamic subcentres. This parallels lower brainstem organization (as in opposing effects on respiration and blood pressure). Stimulation and destruction of disparate hypothalamic centres have contradictory effects. Frontal hypothalamic stimulation primarily yields parasympathetic activity called trophotrop (constructive or protective) visceral functions. These include temperature regulation (heat-reducing mechanisms). Posterior hypothalamic stimulation tends to mobilize sympathetic or ergotrop (energy-consuming) visceral effects such as heat production. Destruction of any of these centres disturbs or abolishes their function; with frontal hypothalamic injury, animals are defenseless against overheating and, with posterior damage, against chilling.

Trophotrop and ergotrop effects

Lateral hypothalamic stimulation induces even well-fed animals to eat voraciously; stimulation of the ventromedial region will make the hungriest animal stop eating or chronically reject food. Reciprocally, lateral electrosurgical destruction abolishes feeding to the point of death from starvation. Ventromedial destruction has produced overeating (hyperphagia), in animals such as rats, cats, or monkeys, which become grotesquely obese.

The hypothesis that emotional qualities are interdependent with these regulating processes is supported by evidence that the posterior hypothalamus also is essential in integrating manifestations of anger and aggression. Injury to many cortical and subcortical areas (*e.g.*, the amygdala) can produce similar phenomena; however, the central significance of the hypothalamus clearly has been confirmed.

"Sham rage"

Anger reactions following hypothalamic injury or stimulation once were labelled "sham rage" (implying transient expressive movement devoid of subjective emotional quality). Yet, animals with such symptoms make goal-directed attacks on appropriate environmental stimuli. It is further demonstrable that hypothalamic stimulation promotes potentially goal-oriented reflex activity (*e.g.*, biting) triggered only by appropriate receptors (*e.g.*, when the region of the mouth served by the trigeminal nerve is denervated, no biting accompanies the other aggressive symptoms).

Manifestations of escape, defense, anger, and attack have been elicited from a contiguous region running from the amygdala through the hypothalamus to the mesencephalon; one worker concluded that brain tissues responsible for escape form a concentric zone that surrounds the region of tissue that produces signs of attack. This spatial configuration suggests a reciprocal, functional relationship that is strikingly demonstrable. When movement is restrained during electrical stimulation of an animal's escape zone, the original effect is transformed into clearly contrasting signs of angry aggression.

The fundamental emotional role of the human hypothalamus emerges during neurosurgery and in hypothalamic disease. Mechanical manipulation or electrical stimulation of the patient's hypothalamus generates violently aggressive or euphoric outbursts.

Recall definitions of drive and motivation as intervening variables inferrable from antecedents and consequences. Yet electrical stimulation of the hypothalamus has identical motivational effects to those ordinarily observed. In a genuine sense identification of the neurological substrate lends such intervening variables more than mere hypothetical validity. Instrumental conditioning by direct brain stimulation does effect enduring motivational changes; they are not simply transient expressive movements.

The motivational effects of such stimulation have been known since the 1950s. Reinforcement in learning escape and avoidance reactions is provided by allowing animals to switch off escape-inducing brain stimulation. After a few successful trials, animals learn to manipulate the

switch in an instrumental response to a warning stimulus or to brain stimulation itself. Animals that have been conditioned only by warning signals will execute learned avoidance and feeding responses in a coordinated fashion when the hypothalamus alone is stimulated, without the warning (conditioned) stimuli.

Brain stimulation as reinforcement

The phenomenon of self-stimulation through electrodes implanted in the brain was discovered in work originally designed to investigate motivational effects. By pressing a bar or pedal, a rat can stimulate implanted parts of its own brain. For some brain loci, continuous, enthusiastic pressing is observed and for other locations, systematic avoidance of the pedal is noted, suggesting emotionally rewarding and punishing effects. The method seems to hold promise for quantitatively mapping (in terms of pressing rate) the two fundamental emotional qualities in the brain. (For example, the number of times an animal presses to interrupt a continuous punishing stimulus can be recorded.)

Reward and punishment centres in the brain

It already has been shown that rewarding and punishing effects are mediated through relatively well-defined and distinct regions. Reward primarily seems to involve the medial forebrain bundle running from the lateral hypothalamus to the brainstem, punishment being associated with the medial hypothalamus; these effects seem to overlap in some regions. Both effects can be elicited elsewhere (amygdala, hippocampus, thalamus, mesencephalic tegmentum), although data for precise localization often have been contradictory.

Despite this evidence for central emotional mechanisms, unanswered questions remain. If pedal pressing in self-stimulation really is equivalent to fundamental positive and negative emotional qualities, it is not clear how this is related to more common forms of emotional expression (*e.g.*, eating, sex). Nor is it known whether all of them are aspects of the same process or are intrinsically independent functions that simply overlap anatomically. A general underlying process seems to be at work in evidence that self-stimulation quantitatively changes with traditional efforts to manipulate drive level (*e.g.*, administration of hormones, castration, increased feeding or food deprivation). Yet, while self-stimulation or consummatory behaviour (*e.g.*, feeding) can be elicited from identical brain loci, the thresholds for these effects seldom coincide. Surprisingly, positive self-stimulation can be elicited from the same brain loci that induce marked anger reactions; the signs of anger do not interfere with self-stimulation.

Such data neither exclude nor endorse the possibility that reward and punishment in self-stimulation correspond to the excitement and inhibition of homeostasis. Though it might, for example, be implied that the motivation for food is emotionally positive and that food avoidance is emotionally negative (or vice versa); this would seem to be improbable since injury of the ventromedial hypothalamus increases food intake (with eventual obesity) only if food is easily obtained. When food-getting requires great effort, the animal will fast. In later stages following such injury, experimental animals become quite fastidious, chronically rejecting tasteless food and losing weight. Observations of this sort suggest that destruction of a subcentre assumed responsible for inhibiting food intake (a centre of satiation) does not increase but reduces the motivation of the animal to eat. (Perhaps the initially excessive consumption of easily accessible food is mediated by some feeding mechanism at a lower neural level.) A fundamental contradiction also seems to emerge from evidence that rewarding self-stimulation more or less coincides (at times is identical) with loci that prompt the animal to eat; on the other hand, areas that inhibit food intake (coinciding with satiation centres) have negative effects on self-stimulation. The subjectively critical problem is that the animal tends to stimulate a brain area that induces signs of hunger (from an introspective human point of view, a negative emotional state) rather than stimulating regions connected with satiation. It is not known why an animal will stimulate itself at a point in the brain that unfailingly produces signs of anger.

These discrepancies suggest that emotional-motivational qualities are not exclusive products of functionally independent subcentres representing contrasting effects; they seem to be integrated resultants of some kind of interaction among neural structures. The following data seem to support this interpretation.

Behavioral effects indicate that the initial, rewarding effect of stimulation of some hypothalamic regions gradually becomes punitive if stimulation is prolonged. Thus, and depending on the predetermined duration of stimulation, animals can be taught systematically both to connect and to disconnect themselves with two opposed pedals. Whatever the mechanism for such changes from positive to negative reinforcement, it decisively involves the intensity of stimulation. Under relatively mild hypothalamic stimulation animals rapidly learn to avoid a vicinity that will switch them off; with high intensity, approach to the switch-off zone also is learned quickly. Animals behave as if they want to sustain mild excitement and to stop intense stimulation. Mild stimulation seems to be emotionally positive, while intense excitement appears negative. While this adds little to the observation that activation theory fails to account for positive emotional qualities of great intensity, what follows may hold genuine promise.

In the experiments just summarized, bodily movements from brain stimulation were recorded simultaneously with electroencephalographic data. Positive or negative emotional effects were found to be related both to the direction of induced movement and to electrical activity in the hippocampus. (Functionally, the cortical region called the hippocampus has a close and ancient evolutionary relationship with the hypothalamus.) Bodily movement toward the side opposite the stimulated hemisphere, associated with slow-wave activity in the hippocampus, signalled an effect the animal appeared reluctant to interrupt; that is, seemed to experience as emotionally positive. Movement or withdrawal toward the side of the stimulated hemisphere, combined with fast activity of a low amplitude (desynchronization) in the hippocampus, characterized an effect the animal consistently switched off, apparently experiencing it as emotionally negative.

Hippocampal indicators of emotional quality

Bodily movement and hippocampal electroencephalographic activity jointly provide an exceptionally flexible indicator of motivational-emotional effects. Systematic variation of the intensity of brain stimulation reveals that an initially positive emotional effect will increase with stimulus intensity or with time only to a given limit; it then progressively decreases and even may reverse. Analysis of movements and electroencephalographic responses invariably has suggested that an effect becomes emotionally negative in direct ratio to the degree with which the originally produced movement tendency is restrained or inhibited. This is the basis for the generalization that, in terms of its mechanism, negative emotion is synonymous with such inhibition. In this sense the process of negative emotion is not a homogeneous neural mechanism but a composite of simultaneous excitement and inhibition. In psychological terms, this corresponds literally to tension, as clearly is indicated by the aftereffects. At the moment brain stimulation is interrupted, the function that was progressively inhibited during stimulation erupts as an aftereffect. This brief but intense effect (in neurophysiology it is called a rebound response) is equivalent to emotionally positive activity according to the movement–hippocampal signs. When the rebound is repeatedly experienced in the same part of the laboratory, the animal will learn to approach that particular place. This evidence that sudden release of inhibited function tends to facilitate learning touches on a controversial aspect of the relationship between motivation and learning.

Excitement and inhibition

At any rate, these data illuminate the question of how the mechanism that generates intense, positive emotional activity is to be conceived: it seems to be one of liberation from inhibition. The movement–hippocampal indicators of this positive emotional quality in rebound are just as intense as the preceding signs of negative emotion, an observation that seems to relieve the inadequacy of

activation theory of accounting only for weak positive emotions.

It may be objected that the signs of positive emotion in rebound are either short-lived or contingent on previously existing negative emotional activity. It is also debatable whether the simple notion of liberation from inhibition does full justice to subjective experience of intense positive emotion. There is, however, evidence on which an account of the development of prolonged, intense, positive activity may be based. It has been established that long-lasting homeostatic conditions ("tuning states") of internal equilibrium can result from humoral and neuronal autonomic mechanisms (sympathetic or parasympathetic). It seems quite possible that the intensity and cyclic characteristics of homeostatic tuning correspond to introspective emotional qualities. Prolonged equilibrium seems to be an appropriate correlate of moods; rapid fluctuation being associated with such classically labelled emotional qualities as fright, surprise, and anger.

By this conception of emotion as a product of regulatory neurophysiology, the isolation of fundamental emotional qualities is to be sought in the dynamics of excitement and inhibition. Efforts to specify regulatory changes that produce any given emotional quality face the technical problem of concomitant measurement and identification of appropriate behaviour and of relevant neural activity.

Determinants of emotional quality

Perhaps emotional qualities also are determined by the adaptive purpose served by different regulatory functions. Although homeostatic processes that support life rest on identical principles, emotionally they are introspected as qualitatively different. The same regulatory mechanism (in identical phase and intensity) may mediate one's approach to a loved person or to a dish of favourite food; yet, romantic and epicurean appetites are perceived differently. Their distinction clearly rests on perceptual learning, itself already shown to depend on regulatory dynamics. Interaction and overlap of systems that regulate such behaviour as feeding and mating further complicate the study of emotional qualities since underlying mechanisms may not be separable.

At any rate, research into the regulation of motivation has simply provided an endorsement of the old, fundamentally two-dimensional (positive–negative) hedonic theory, all other qualities being held derivative. Nevertheless, the motivational approach has delineated indicators that seem to resolve once-puzzling discrepancies.

It bears repeating that in research limited to such peripheral autonomic indicators as vascular and hormonal changes, only two emotional qualities have been consistently separable. These indicators lump fear, anxiety, and similar states; then they paradoxically group extreme pleasure (joy) with a set of hardly agreeable subjective experiences such as anger, rage, and resentment. (Efforts to distinguish joy from anger in facial expression also produce confusion.) Added to this is the apparently paradoxical observation that an anger-producing hypothalamic locus also encourages self-stimulation.

Joy (emotionally positive) and such experiences as anger (traditionally regarded as negative) could be empirically confused through inadequate objective indicators if their underlying mechanisms overlapped too much. Indeed, overlap in terms of compressed sequence over time is observed during electrical stimulation of the hypothalamus. Evoked anger and attack consistently are preceded by signs of fear that last from a fraction of a second to a few seconds (depending on the stimulation). Somatic symptoms of fear are accompanied by hippocampal activity characteristic of negative emotion. Signs of anger and attack appearing as rebound (at the moment stimulation is switched off) or as an eruptive transition during stimulation are consistently associated with the slow, high-amplitude hippocampal activity (theta waves) characteristic of *positive* emotional states. According to the hippocampal indicator, rage and attack are to be interpreted as *positive* emotional reactions that arise as liberations from inhibition. This interpretation is buttressed by the observation that rage and attack have a typical motor pattern of approach (in contrast to withdrawal in fear and defense).

Thus, it seems less paradoxical that self-stimulation and anger go together. These experiments serve to illustrate the postulate that regulatory changes at different levels of intensity correspond to different emotional qualities. In these terms, rage is a positive counterpart of panic-like fear, emerging during a period of relatively intense homeostatic-motivational activity.

Rage and attack as positive emotions

Negative emotional-motivational neural signs among laboratory animals generally arise in situations that interfere with effort or tendency to approach a goal. The psychological terms frustration and conflict coincide by definition with such neuronal interpretations of negative emotions; namely, they also signify inhibition of the approach tendency. On the other hand, aggressive behaviour seems emotionally positive on the basis of regulatory indicators. Aggressive human personality changes, among the most common consequences of conflict and frustration, can also be conceived as the enduring symptoms of liberation from inhibition. Negative emotional aspects of what are called repression and regression might be interpreted as activities in which inhibition has become dominant.

The limbic system. Hypothalamic mechanisms fall far short of exhausting the emotional substrate. The hypothalamus mediates the most fundamental homoeostatic bases for survival of individual and species (*e.g.*, internally influenced metabolic function). These primitive, built-in, motivational-regulatory mechanisms permit wider possibilities for classifying emotion than seemed possible when activation theory first was developed. They also draw attention as archaic evolutionary sources of emotion. Learning processes, however, also become crucial in regulating motivation. Through learning, initially indifferent external stimuli become motivating and provide a basis for new learning. The individual gradually learns a complicated hierarchy of cognitive signs and relationships that have emotional significance beyond that of processes that serve physiological survival. Clearly, the hypothalamus alone is structurally insufficient for evaluating complex cognitive emotional influences or for integrating their ramifications in visceral response. Nevertheless, it is unlikely that complicated cognitive emotional processes are independent of or fundamentally different from hypothalamic mechanisms.

The influence of cognitive learning

Similar considerations led to a theory in the 1930s that attributed cognitive aspects of emotion to higher forebrain structures in close anatomic relationship with the hypothalamus. These were depicted primarily as perceiving emotion, contrasting the hypothalamus as a structure for its expression. The forebrain locus central to the theory is the cingulate gyrus (ridge) on the medial cortical surface. It is older in evolution than the brain's neocortex and simpler in structure. Through intermediate relay centres (hippocampus, mammillary bodies, anterior thalamus) the gyrus connects the hypothalamus with higher structures for emotional representation. This speculative emphasis on the cingulate gyrus directed attention to the integrative complex of brain structures now known as the limbic system. (Comparative anatomy also provides important impetus.) The term limbic system is used loosely to include all rhinencephalic structures once identified with the sense of smell (olfaction); *e.g.*, the hippocampus, amygdala, septum, cingulate gyrus, portions of the temporal lobe, and medial nuclei of the thalamus. The name rhinencephalon ("nose brain") now is used as a synonym for limbic system, since it has been found to have at most a secondary relationship to the primary olfactory brain.

In contrast to the neocortex, limbic structures have prominent vegetative (autonomic) influence. This is powerfully consistent with evidence that emotional processes are mediated by neural systems that also regulate the internal milieu. Such evidence prompted the suggestion of the term visceral brain (in lieu of rhinencephalon or limbic system), stressing the role of hippocampal structures and of the amygdala over that of the cingulate gyrus. (The amygdala is a complex of subcortical nuclei that

The limbic system or the visceral brain

connect with the temporal neocortex.) It was theorized that these structures are agencies for sensory and motor emotional function (analogous to neocortical sensory and motor projections that interact with the environment).

Emotional–cognitive unity clearly emerges in the frequent finding that hippocampal injury (laboratory animals and humans) also produces serious disorders in learning and remembering. Presumably this hippocampal mechanism participates in analyzing environmental data related to physiological survival, integrating them with the regulatory activity of the hypothalamus.

Injuries to various parts of the limbic system have a significant influence on emotional lability. Massive damage involving temporal lobe and amygdala tends to tame aggressive animals, producing oral and hypersexual tendencies. Such injury damps the effects of frightening stimuli and tends seriously to reduce readiness to react socially. More selective lesions, however, may increase aggressive tendencies. At earlier stages of their evolution (as in many extant species) limbic structures are closely related to olfactory function. It has been observed that the sense of smell among ancestors of man must have been critical (*e.g.*, in acquiring food, in detecting an enemy or sexual partner).

Such motivational polarities of the hypothalamus as starvation–gluttony appear for the limbic system but are much less clear and barely definable anatomically. Perhaps the emotional functions of the limbic system will become clearer after its regulatory functions have been more thoroughly explored.

The frontal lobes. Even the frontal lobes, last part of the human neocortex to evolve, seem to be special structures for emotional integration. Once called the "silent" lobes, they now are known to have substantial visceral–autonomic representation, transmitting visceral impulses through several subcortical systems (*e.g.*, the hypothalamus).

Lobectomy and lobotomy

Among lower primates, it long has been known that experimentally produced signs of fear and anxiety disappear with removal of the frontal lobes. Innovators adapted such findings to humans for the relief of otherwise intractable anxieties and psychotic depression. Instead of removing the lobes (lobectomy), surgeons cut the connections between the frontal cortex and subcortical structures; the operation is called a lobotomy. With encouraging initial successes, lobotomies were performed on thousands of people. In some cases the method was modified to destroy small bits of cortex or thalamic relay centres for the frontal lobe.

This therapeutic method largely belongs to the past; effects of lobotomy tend to be transient and difficult to predict; psychotropic drugs are more reliable and entail less risk; and lobotomy has produced more or less serious psychological damage (*e.g.*, distractibility, intellectual deficit). However, lobotomy can reduce or abolish emotional concomitants of severely painful bodily conditions. After frontal pathways are interrupted, the patient can sense pain but no longer reacts with signs of emotional distress.

This phenomenon clearly bears on emotional function and raises the question of how painful experiences arise. It may be observed (in terms of regulatory theory) that brain structures substantially regulate the impulses they receive from so-called pain receptors. Classically, the experience called pain was considered as an elementary sensory function tied to such receptors. More recent work suggests that the experiences involve at least two interacting regulatory mechanisms. One is a so-called gate-control mechanism of the spinal cord that modulates (inhibits and facilitates) receptor impulses directed toward the brain. Such impulses may evoke painful experience if they are realized in the brain. They are assumed to be realized only if a second mechanism in the hierarchy (that regulates motivational intensity as discussed above in relation to hypothalamic function) functions in appropriate phase. Theoretically, pain relief through frontal-lobe removal could be related to such a motivational control mechanism.

Injury to the frontal lobe most consistently results in disinhibition; rage manifested by decorticate dogs long has been interpreted as liberation of subcortical centres from inhibition by the cortex. Hierarchically arranged inhibiting functions are found on every level of the nervous system, and the activity at any level may be inferred to be the result of integrated interaction with structures farther along the neuraxis. Cortical tissues seem to integrate and interact with subcortical and spinal structures. In this way the organism can make appropriately selective somatic and visceral responses as it learns the complexities of the environment. This continuing process of selection and shifting at higher levels modifies, retards, and inhibits the responses of the more primitive regulatory apparatus.

Disturbances that accompany injuries at the widest variety of levels of neural structures that integrate emotion-motivation demonstrate that all human reactions (from primitive homeostatic responses to cognitive judgments) are emotional. Destruction or impairment of structures that mediate the emotional component does not result in "cold" cognitive function but in impaired cognitive function. The cognitive (as well as visceral) interests of the organism seem always to be emotionally expressed; indeed, such a unity of function seems to have served the very survival of the human species.

BIBLIOGRAPHY. P.T. YOUNG, *Motivation and Emotion: A Survey of the Determinants of Human and Animal Activity* (1961), an original endeavour to explain behaviour from the viewpoint of motivation and emotion; M.B. ARNOLD, *Emotion and Personality*, vol. 1, *Psychological Aspects*, vol. 2, *Neurological and Physiological Aspects* (1960), an attempt to integrate manifold sources of evidence and theories on emotion, and (ed.), *The Nature of Emotion: Selected Readings* (1968); D.O. HEBB, "Emotion in Man and Animal: An Analysis of the Intuitive Process of Recognition," *Psychol. Rev.* 53:88–106 (1946), a lucid, readable paper on the genesis of emotion from one of the most influential theoreticians in contemporary psychology; E. GELLHORN and G.N. LOOFBOURROW, *Emotions and Emotional Disorders* (1963), a thorough review of central and visceral events and disorders that accompany emotions; D.K. CANDLAND (ed.), *Emotion: Bodily Change, an Enduring Problem in Psychology: Selected Readings* (1962).

(E.Gr.)

Empiricism

In its most general sense, Empiricism is the name applied to the attitude that beliefs are to be accepted and acted upon only if they first have been confirmed by actual experience—a definition that accords with the derivation of the name from the Greek word *empeiria*, "experience."

More specifically, Empiricism comprises a pair of closely related, but still distinct, philosophical doctrines—one pertaining to concepts and the other to propositions:

Empiricism of concepts and of beliefs

1. The first of these doctrines, a theory of meaning, holds that words (*e.g.*, the word substance) can be understood or the concepts requisite for any articulate thought possessed only if they are connected by their users with things that they have experienced or could experience (*e.g.*, pieces of wood, or the gases in a gasoline engine).

2. The second doctrine, a philosophical theory of knowledge, views beliefs, or at least some vital classes of beliefs (*e.g.*, that Jane is kind), as depending ultimately and necessarily on experience for justification (Jane is seen performing acts of kindness).

It is not obvious, however, that either of these two doctrines strictly implies the other. Several recognized Empiricists have admitted that there are a priori propositions but have denied that there are a priori concepts. The reverse disconnection between the two forms of Empiricism, however, has no obvious exponents, since there are hardly any philosophers who totally deny a priori propositions and certainly none who would at the same time accept a priori concepts.

As stressing experience, Empiricism is thus opposed to the claims of authority, intuition, imaginative conjecture, and abstract, theoretical, or systematic reasoning as sources of reliable belief. Its most fundamental antithesis is with the latter (*i.e.*, with Rationalism, also called intel-

lectualism or apriorism). A Rationalist theory of meaning asserts that there are concepts not derived from or correlated with experienceable features of the world, such as "cause," "identity," or "perfect circle," and that these concepts are a priori (Latin "from the former") in the traditional sense of being part of the mind's innate or natural equipment—as opposed to being a posteriori (Latin "from the latter"), or grounded in the experience of facts. On the other hand, a Rationalist theory of knowledge holds that there are beliefs that are a priori (*i.e.*, that depend for their justification upon thought alone), such as the belief that everything must have a sufficient reason or that a process cannot exist by itself but must occur within some substance. Such beliefs can arise either from intellectual intuition, the direct apprehension of self-evident truth, or from purely deductive reasoning.

VARIOUS MEANINGS OF EMPIRICISM

Broader senses. In both everyday attitudes and philosophical theories, the experiences referred to are principally those arising from stimulation of the sense organs, in particular those of sight and touch. Most philosophical Empiricists, however, have maintained that sensation is not the only provider of experience, admitting as empirical the awareness of mental states in introspection or reflection, such as feelings of pain or of fear, often metaphorically described as present to the "inner sense." It is a controversial question whether still further types of experience, such as moral, aesthetic, or religious experience, ought to be acknowledged as empirical.

Two other viewpoints related to but not the same as Empiricism are the Pragmatism of the American philosopher and psychologist William James, an aspect of which was Radical Empiricism, and Logical Positivism, also called Logical Empiricism (see POSITIVISM AND LOGICAL EMPIRICISM). Though these philosophies are, indeed, empirical, each has a distinctive focus that warrants its treatment as a separate movement. Pragmatism stresses the involvement of ideas in practical experience and action, whereas Logical Empiricism is more concerned with scientific experience.

Popular and medical Empiricism

When describing an everyday attitude, the word Empiricism sometimes conveys an unfavourable implication of ignorance of or indifference to relevant theory. Thus, to call a doctor an "Empiric" has been to call him a quack —a usage traceable to a sect of medical men who were opposed to the elaborate medical, and in some views metaphysical, theories of Galen, a prominent Greek physician of the 2nd century AD, theories which dominated medicine until the 17th century. The medical Empiricists opposed to Galen preferred to rely on treatments of observed clinical effectiveness, without inquiring into the mechanisms sought by therapeutic theory. But "Empiricism," detached from this medical association, may also be used, more favourably, to describe a hard-headed refusal to be swayed by anything but the facts that the thinker has observed for himself, a blunt resistance to received opinion or precarious chains of abstract reasoning.

Stricter senses. As a more strictly defined movement, Empiricism reflects certain fundamental distinctions and occurs in varying degrees.

Fundamental distinctions. If the blurring of the distinction between concepts and propositions has confused discussions of Empiricism, another influence at least equally vexing is that which, embodied in the traditional terminology of the debate, contrasts the empirical not with the a priori but with the innate. Since logical problems are easily confused with psychological problems, it is difficult to disentangle the question of the causal origin of man's concepts and beliefs from that of their meaning and justification.

Innate versus a priori

A concept, such as "Five," is said to be innate if a man's possession of it is causally independent of his experience —*e.g.*, of groupings of five objects. A concept, such as "Ought," is a priori, on the other hand, if the logical conditions of its application (or those of the word expressing it) do not include any reference to experienceable states of affairs (thus "Ought" cannot be defined in terms of facts alone). Similarly, a belief is innate if its acceptance is causally independent of the believer's experience; and it is a priori if its justification is logically independent of experience. Propositions could be innate without being a priori: for example, the baby's empirical belief that its mother's breast will nourish it.

Some colour is lent to the confusion of the a priori and the innate by the fact that most empirical concepts are actually acquired through the technique of ostensive definition, in which a concept (such as "long") is conveyed by introducing instances of it (several long pencils) into the experience of the learner. But it is not in virtue of their mode of acquisition that concepts are empirical; it is the way in which they are applied, once they are possessed, that qualifies them as such. Even if a man were born with an instinctive capacity to use the word blue and never had to learn how to use it, it would still be an empirical concept if the occasion for its use were always his sense experience of a blue object. Furthermore, the fact that men learn, postnatally, to use the word cause does not prove that it expresses an empirical concept if its application to something always implies more than the sense which experience of that kind of thing presents to the mind.

Another supposedly identical but in fact more or less irrelevant property of concepts and beliefs is that of the universality of their possession or acceptance—that a priori or innate concepts must be the common possession of all men and that such beliefs must be accepted by everyone. There may be, in fact, some basis for inferring universality from innateness, since many innate characteristics such as the fear of loud noises appear to be common to the whole species.

Formal versus categorial

Two main kinds of concept have been held to be a priori and thus nonempirical. First, there are certain formal concepts of logic and of mathematics that reflect the basic structure of discourse: "not," "and," "or," "if," "all," "some," "existence," "unity," "number," "successor," "infinity." Secondly, there are categorial concepts, such as "substance," "cause," "mind," and "God," so-called after the "categories of thought" as listed by Aristotle and Kant, which the mind imposes upon the given data of experience.

A very large variety of different types of proposition has been held to be a priori. Few would deny this status to such definitional truisms or obvious tautologies as "all hairless heads are bald" or "a rose is a rose." There are also the truths of logic, of mathematics, and of metaphysics—whether transcendent, such as the existence of God or things-in-themselves (lying behind appearances), or immanent and thus discernible within reality, such as the principles (presupposed by much natural science) of conservation, causality, and sufficient reason. Some have held that the basic principles of ethics or the causal laws of nature are a priori. Empiricism maintains, however, that some of these are a priori.

Degrees of Empiricism. Empiricism, whether concerned with meaning or knowledge, can be held with varying degrees of strength. On this basis, absolute, substantive, and partial Empiricisms can be distinguished.

Absolute Empiricists hold that there are no a priori concepts, either formal or categorial, and no a priori propositions. Absolute Empiricism about knowledge is less common than that about concepts, and nearly all philosophers admit that at least obvious tautologies and definitional truisms are a priori; but many would add that these represent a degenerate case.

Substantive Empiricism

A more moderate form of Empiricism is that of the substantive Empiricists, who are unconvinced by attempts that have been made to interpret formal concepts empirically and who therefore concede that formal concepts are a priori but deny that categorial concepts, such as "substance," "cause," and "God," are a priori. In this view, formal concepts would be no longer semantical, pertaining to the relation of words to things; they would be, instead, merely descriptive or purely syntactical, pertaining to the relations between ideas. On this basis

"God," for example, would not be an entity alongside other entities but a device for arranging a man's factual beliefs about the world; the concept "God" would thus play a structural and not an informative role.

The parallel point of view about knowledge assumes that the truth of logical and mathematical propositions is determined, as is that of definitional truisms, by the relationships between meanings that are established prior to experience. The truth often espoused by moralists, for example, that one is truly obliged to rescue a man from drowning only if it is possible to do so is a matter of meanings and not of facts about the world. On this view, all propositions that, in contrast to the foregoing example, are in any way substantially informative about the world are empirical. Even if there are a priori propositions, they are all formal or verbal or conceptual in nature, and their necessary truth derives simply from the meanings that man has attached to the words he uses. A priori knowledge is useful because it makes explicit the hidden implications of substantive, factual assertions. But a priori propositions do not themselves express genuinely new knowledge about the world; they are factually empty, "true in all possible worlds." Thus "All bachelors are unmarried" merely gives explicit recognition to the commitment to describe as unmarried anyone who has been described as a bachelor; but it does not add anything new.

Substantive Empiricism about knowledge, which is fundamental to most contemporary Analytical philosophy, regards all a priori propositions as being more-or-less concealed tautologies, or, in Kantian terms, as being analytic. If a person's "duty" is thus defined as that which he should always do, the statement "A person should always do his duty" then becomes "A person should always do what he should always do." Deductive reasoning is conceived accordingly as a way of bringing this concealed tautological status to light. That such extrication is nearly always required means that a priori knowledge is far from trivial.

For the substantive Empiricist, truisms and the propositions of logic and mathematics exhaust the domain of the a priori. Science, on the other hand—from the fundamental assumptions about the structure of the universe to the singular items of evidence used to confirm its theories—is regarded as empirical throughout. The propositions of ethics and those of metaphysics, which deals with the nature of Being as such (for example, "Only that which is not subject to change is real"), are either disguised tautologies or empirical statements or only pseudo-propositions; *i.e.*, combinations of words that, despite their grammatical respectability, cannot be taken as true or false assertions at all.

Partial Empiricism

The least thoroughgoing type of Empiricism here distinguished, ranking third in degree, can be termed partial Empiricism. Many philosophers hold that other concepts besides the formal are a priori and that there are substantially informative propositions about the world that are nevertheless not empirical. The theses of transcendental, or Kantian, metaphysics and those of theology, the general scientific principles of conservation and causality, the basic principles of morality, and the causal laws of nature have all been held to be at once synthetic and a priori—substantial and yet establishable by reasoning alone without recourse to experience. In all versions of this view, however, there remain a great many straightforwardly empirical concepts and propositions: ordinary singular propositions about matters of fact and the concepts that figure in them are held to fall in the empirical domain.

HISTORY OF EMPIRICISM

In ancient philosophy. So-called common sense is inarticulately empiricist; and philosophy, in seeking to correct it, has to start from a rationalistic position. Philosophical Empiricism is thus always critical, a resistance to the pretensions of a more speculative philosophy. The ground was prepared for Plato, the greatest of Rationalist philosophers, by three earlier bodies of thought: the Ionian cosmologies of the 6th century BC—so-called from their concentration along the western coast of Asia Minor—with their distinction between sensible appearance and a reality accessible only to pure reason; the philosophy of Parmenides (early 5th century BC), the important early monist, in which purely rational argument is used to prove that the world is really an unchanging unity; and Pythagoreanism (*q.v.*), which, holding that the world is really made of numbers, took mathematics to be the repository of ultimate truth.

The first Empiricists in Western philosophy were the Sophists, who rejected such Rationalist speculation about the world as a whole and took man and society to be the proper objects of philosophical inquiry. Invoking skeptical arguments to undermine the claims of pure reason, they posed a challenge that invited the reaction that comprised Plato's philosophy (see below, *Criticism and evaluation*).

Plato and to a lesser extent Aristotle were both Rationalists. But Aristotle's successors in the ancient Greek schools of Stoicism and Epicureanism (*qq.v.*) advanced an explicitly Empiricist account of the formation of man's concepts or ideas. For the Stoics the human mind is at birth a clean slate, which comes to be stocked with ideas by the sensory impingement of the material world upon it. Yet they also held that there are some ideas or beliefs, the "common notions," present to the minds of all men; and these soon came to be conceived in a nonempirical way. The Empiricism of the Epicureans, however, was more pronounced and consistent. For them man's concepts are memory images, the mental residues of previous sense experience; and knowledge is as empirical as the ideas of which it is composed.

In medieval philosophy. Most medieval philosophers after St. Augustine (see below, *Criticism and Evaluation*) took an Empiricist position, at least about concepts, even if they recognized much substantial but nonempirical knowledge. The standard formulation of this age was: "There is nothing in the intellect that was not previously in the senses." Thus St. Thomas Aquinas (1225–74) altogether rejected innate ideas. Both soul and body participate in perception, and all of man's ideas are abstracted by the intellect from what is given to the senses. Man's ideas of unseen things, like God and angels, are derived by analogy from the seen.

The 13th-century scientist Roger Bacon emphasized empirical knowledge of the natural world and anticipated the polymath Renaissance philosopher of science Francis Bacon (1561–1626) in preferring observation to deductive reasoning as a source of knowledge. The Empiricism of the 14th-century Franciscan Nominalist William of Ockham was more systematic. All knowledge of what exists in nature, he held, comes from the senses, though there is, to be sure, "abstractive knowledge" of necessary truths; but this is hypothetical and does not imply the existence of anything. His more extreme followers extended his line of reasoning toward a radical Empiricism, in which causation is not a rationally intelligible connection but merely an observed regular sequence.

The Empiricism of William of Ockham

In modern philosophy. In the earlier and unsystematically speculative phases of Renaissance philosophy, the claims of Aristotelian logic to yield substantial knowledge were attacked by several 16th-century logicians, and, in the same century, the role of observation was stressed. One mildly skeptical Christian thinker, Pierre Gassendi (1592–1655), advanced a deliberate revival of the empirical doctrines of Epicurus. But the most important defender of Empiricism was Francis Bacon, who, though he did not deny the existence of a priori knowledge, claimed that, in effect, the only knowledge that is worth having (as contributing to the relief of man's estate) is empirically based knowledge of the natural world, which should be pursued by the systematic, indeed almost mechanical, arrangement of the findings of observation and is best undertaken in the cooperative and impersonal style of modern scientific research. Bacon was, indeed, the first to formulate the principles of scientific induction.

A Materialist and Nominalist, Thomas Hobbes (1588–1679), combined an extreme Empiricism about concepts, which he saw as the outcome of material impacts on the

bodily senses, with an extreme Rationalism about knowledge, of which, like Plato, he took geometry to be the paradigm. For him all genuine knowledge is a priori, a matter of rigorous deduction from definitions. The senses provide ideas; but all knowledge comes from "reckoning," from deductive calculations carried out on the names that the thinker has assigned to them. True knowledge is thus not merely a priori but also analytic. Yet it all concerns material and sensible existences: everything that exists is a body.

Locke's Empiricism

The most elaborate and influential presentation of Empiricism was made by John Locke (1632–1704), an early Enlightenment philosopher, in the first two books of his *Essay Concerning Human Understanding* (1690). All knowledge, he held, comes from sensation or from reflection, by which he meant the introspective awareness of the workings of man's own mind. Locke confused the two issues of the nature of concepts and the justification of beliefs. His Book I, though titled "Innate Ideas," is largely devoted to refuting innate knowledge. And even so, he later admitted that much substantial knowledge—in particular, that of mathematics and morals—is a priori. He argued that infants know nothing; that if men are said to know innately what they are capable of coming to know, then all knowledge is, trivially, innate; and that no beliefs whatever are universally accepted. Locke was more consistent about the empirical character of all man's concepts and displayed in detail the ways in which simple ideas can be combined to form complex ideas of what has not in fact been experienced. One group of dubiously empirical concepts—those of unity, existence, and number—he took to be derived both from sensation and from reflection. But he allowed one a priori concept—that of substance—which the mind adds, seemingly from its own resources, to its conception of any regularly associated group of perceptible qualities.

Bishop George Berkeley (1685–1753), a theistic Idealist and opponent of Materialism, applied Locke's Empiricism about concepts to refute Locke's account of man's knowledge of the external world. He drew and embraced the inevitable conclusion that material things are simply collections of perceived ideas, a position that ultimately leads to phenomenalism; *i.e.*, to the view that reality is nothing but sensations. He accounted for the continuity and orderliness of the world by supposing that its reality is upheld in the perceptions of an unsleeping God. The theory of spiritual substance involved in Berkeley's position seems to be vulnerable, however, to most of the same objections as those that he posed against Locke.

The Scottish Skeptical philosopher David Hume (1711–76) fully elaborated Locke's Empiricism and used it reductively to argue that there can be no more to man's concepts of body, mind, and causal connection than what occurs in the experiences that he has of them. For Hume all necessary truth is formal or conceptual, determined by the relations of identity and exclusion that hold between ideas.

Voltaire imported Locke's philosophy into France; and its Empiricism, in a very stark form, is the basis of sensationalism, in which all of the constituents of human mental life are analyzed in terms of sensations alone.

A genuinely original and clarifying attempt to resolve the controversy between Empiricists and their opponents was made in the critical philosophy of Immanuel Kant (1724–1804), drawing upon Leibniz and Hume. With the dictum that, although all knowledge begins with experience it does not all arise from experience, he established a clear distinction between the innate and the a priori. He held that there are a priori concepts, or categories—substance and cause being the most important—and also substantial or synthetic a priori truths. Although not derived from experience, the latter apply to experience. A priori concepts and propositions do not relate to a reality that transcends experience; they reflect, instead, the mind's way of organizing the amorphous mass of sense impressions that flow in upon it.

Lockean Empiricism prevailed in 19th-century England until the turn to Hegel occurred in the last quarter of the century. To be sure, the Scottish philosophers who followed Hume but avoided his Skeptical conclusions insisted that man does have substantial a priori knowledge. But the philosophy of John Stuart Mill (1806–73), logician, economist, and Utilitarian moralist, is thoroughly Empiricist. He held that all knowledge worth having, including mathematics, is empirical. The apparent necessity of mathematics, according to Mill, is the result of the unique massiveness of its empirical confirmation. All real knowledge for Mill is inductive and empirical; and deduction is sterile. On similar lines, the philosopher of evolution Herbert Spencer (1820–1903) offered another explanation of the apparent necessity of some of man's beliefs: they are the well-attested empirical beliefs of his ancestors from whom he has inherited them, an evolutionary revival of the doctrine of innateness. Two important mathematicians and pioneers in the philosophy of modern physics, W.K. Clifford (1845–79) and Karl Pearson (1857–1936), defended radically Empiricist philosophies of science, anticipating the Logical Empiricism of the 20th century.

J.S. Mill and Positivism

In contemporary philosophy. The most influential Empiricist of the 20th century was the great British philosopher and logician Bertrand Russell (1872–1970), who at first was Lockean in his theory of knowledge—admitting both synthetic a priori knowledge and concepts of unobservable entities. Ludwig Wittgenstein (1889–1951), the influential pioneer of the school of Linguistic Analysis, convinced Russell that the truths of logic and mathematics are analytic; and Russell then came to believe, with Hume, that the task of philosophy is to analyze all concepts in terms of what can be directly present to the senses. In this spirit, he tried to show that even the concepts of formal logic are ultimately empirical though the experience that supplies them may be introspective instead of sensory.

Doctrines developed through the collaboration of Russell and Wittgenstein yielded the Logical Positivism of the German philosopher Rudolf Carnap (1891–1970) and of the Vienna Circle, a discussion group in which that philosophy was worked out. The Empiricism of Logical Positivism is especially evident in its restatement of the fundamental thesis of Hume's philosophy in a form known as "the verification principle," which recognizes as meaningful and synthetic only those sentences that are in principle verifiable by reference to sense experience.

CRITICISM AND EVALUATION

The earliest expressions of Empiricism in ancient Greek philosophy were those of the Sophists. In reaction to them Plato presented the Rationalistic view that man has only "opinion" about changing, perceptible, existing things in space and time; that "knowledge" can be had only of timeless, necessary truths; and that the objects of knowledge—the unchanging and imperceptible forms or universals (such as "Bed," "Man," etc.)—are the truly real. The circles and triangles of geometrical "knowledge," in this view, are quite different in their perfect exactness from the approximately circular and triangular things present to man's senses. In the *Phaedo*, Plato expounded a theory of literally innate ideas; man, for example, has a conception of exact "equality," which, since it could not have been supplied by the senses, must have been acquired by the soul before it was embodied.

The critique of Plato and Aristotle

Aristotle agreed with Plato that knowledge is of the universal but held that such universal forms should not be conceived as "separated" from the matter embodying them. This belief does not make Aristotle an Empiricist, although he was certainly a less extreme Rationalist than Plato. Aristotle took the Rationalist view that every science or body of knowledge must resemble Euclidean geometry in consisting of deductions from first principles that are self-evidently and necessarily true and that, although the senses acquaint man with the sensible forms of things, he cannot have knowledge of them unless reason is brought into play to acquaint him with their intelligible forms as well.

The Stoic view of "common notions," or beliefs that are held by all men—a Rationalistic element in an otherwise

Empirical school of thought—was expanded in the early medieval period by the Christian Platonist St. Augustine (AD 354–430), a thoroughgoing Rationalist. The Stoic common notions, he held, are truths that God has implanted in the human mind by direct illumination.

Though the early modern expression of Empiricism in the 17th century by Francis Bacon heralded the scientific age, its influence was lessened by his failure to appreciate the revolutionary use of mathematics that comprised the genius of Galileo's new physics and, even more fundamentally, by his underestimation of the need for imaginative conjecture in the formation of scientific hypotheses to restrict the overwhelming number of facts that would otherwise have to be handled. In contrast to the views of Bacon, his contemporary, the philosopher and mathematician René Descartes (1596–1650), one of the principal founders of modern thought, developed a comprehensive Rationalism that was more immediately influential. For him all clear and distinct ideas, and in particular those of philosophy and of geometrical physics, are innate; sense experience is at most the agency that elicits ideas already present in the mind; in principle, all knowledge is a priori and demonstrable by pure reasoning, but in practice, because of man's finite intellect, it is necessary to rely on experience to confirm propositions for which rational proof is beyond reach. In England, innate ideas and knowledge were defended by Edward Lord Herbert of Cherbury (1582–1648), precursor of Deism, and by a school of Puritan Humanists known as the Cambridge Platonists. The case for innate ideas, however, is hard to establish; there can be in the nature of the case little actual evidence that one can possess concepts before he has had some relevant experience.

In the second half of the 17th century, the Empiricist views of John Locke were similarly controverted by a systematic thinker and man of affairs, G.W. Leibniz, who examined Locke's views in minute detail in his book *Nouveaux essais sur l'entendement humain* (1704, published 1765; Eng. trans., *New Essays*, 1916), arguing that ideas can be virtually innate in a less trivial sense than Locke allowed. Interpreting Locke's notion of reflection as reasoning rather than as introspection, Leibniz supposed that Locke was more of a Rationalist than he really was.

In contemporary philosophy, there are thinkers who, though broadly sympathetic to Positivism, have voiced reservations about its more specifically Empiricist elements. One important philosopher of science, Karl Popper (1902–), has rejected the inductivism that views the growth of empirical knowledge as the result of a mechanical routine of generalization. To him it is falsifiability by experience that makes a statement empirical. An American philosopher and logician, W.V. Quine (1908–), critical of the Logical Empiricists' frequent recourse to the concept of meaning, has rejected the sharp distinction that has been made between analytic and factual truths, on which most of contemporary Empiricism rests. For him, both man's concepts and his beliefs are the joint outcome of experience and conventional decision, and he denies that the role of the two factors can be readily distinguished.

The theory of knowledge has been the central discipline in philosophy since the 17th century, and its most basic issue is that between Empiricism and Rationalism, an issue that is still being actively debated. On the one hand, the idea that science rests on substantial but nonempirical presuppositions has been put in question by the fact that in some areas it seems to get along without them: without conservation in cosmology, without causality in quantum physics. On the other hand, the traditional theory of the innate powers of the mind has been reanimated by the considerations underlying the theory of language offered by Noam Chomsky, a generative grammarian, who holds that language learning is too rapid and too universal to be attributed entirely to an empirical process of conditioning. The basic strength of Empiricism consists in its recognition that man's concepts and beliefs apply to a world outside himself, and that it is by way of his senses that this world acts upon him. The question, however, of just how much the mind itself contributes to the task of processing its sensory input is one that no simple argument can answer.

BIBLIOGRAPHY

Classic texts: JOHN LOCKE, *An Essay Concerning Human Understanding*, 2 vol. (1690); DAVID HUME, *A Treatise of Human Nature*, book 1, pt. 1 (1739); IMMANUEL KANT, *Kritik der reinen Vernunft* (1781; Eng. trans., *Critique of Pure Reason*, 1929); JOHN STUART MILL, *A System of Logic, Ratiocinative and Inductive*, books 1 and 2 (1843).

General surveys: W.H. WALSH, *Reason and Experience* (1947); H.H. PRICE, *Thinking and Experience*, 2nd ed. (1969), original contributions to the Empiricist tradition. For a comprehensive selection of standard works from Locke to J.S. Mill, see A.J. AYER and R. WINCH (eds.), *The British Empirical Philosophers* (1952).

Modern works: BERTRAND RUSSELL, *Human Knowledge* (1948); W.T. STACE, *Theory of Knowledge and Existence* (1932); RUDOLF CARNAP, *Der logische Aufbau der Welt* (1928; Eng. trans., *The Logical Structure of the World*, 1967); A.J. AYER, *Language, Truth and Logic*, 2nd ed. (1946), a short exposition of the extreme Empiricist position.

(A.M.Q.)

Employment and Unemployment

Work satisfies many needs. For the individual it satisfies the need to exercise his faculties and to participate in the collective work of society. It also gives him a claim upon the social product, enabling him to support himself and his family. From the standpoint of the community work is necessary for survival and civilization.

Throughout most of history, man has had to wrest a living from the soil, for the longest period with little more than his bare hands. Without technology, the earth can be made to yield only a meagre existence, and the population dependent upon it is inexorably limited by starvation. Everything man consumes comes from the efforts he has made to transform nature; even the simplest foods, such as wheat, potatoes, or tree fruits, have been created by artificial selection and must constantly be defended from the competition of natural plants. It has been estimated that if mankind were to disappear from the face of the earth, wheat would give way in less than a quarter of a century to various wild grasses and the same fate would befall other crops, fruit trees, and domestic animals.

CHANGING CHARACTER OF THE LABOUR FORCE

In traditional societies, in which the productivity of agricultural labour is very low, virtually the entire population must be employed in farming. As a country develops, it becomes able to feed its people more easily, so that a smaller proportion need be employed in farming. When productivity reaches a certain point the demand for primary goods (agricultural products) drops in relation to the demand for other goods such as clothing, shelter, and manufactured products. The production of these so-called secondary goods ultimately becomes organized in factories and expands dramatically. As the demand for other manufactured goods grows and remains high, employment in the secondary sector remains high as well. There is also a third, or tertiary, sector of employment comprising the service trades, teaching, administration, scientific research, medical care, art, tourism, and other pursuits that are not carried on in factories. As technical progress in this sector is slight, compared to that in the primary and secondary sectors, its activities require a large number of people. In countries with high living standards, demand for products of the tertiary sector keeps increasing; hence employment in that sector increases more rapidly than in the others.

Effects of industrialization on the labour force

The radical nature of the employment changes that occur in the course of economic development can be seen from the statistics of employment in the primary, secondary, and tertiary sectors. It is estimated that in France in 1800 some 80 percent of the working population were employed in agriculture, whereas in 1970 the figure was only about 13 percent. In the United States in 1820 about 72 percent of the working population were employed in agriculture; by 1970 the proportion had fallen

to 4 percent. This meant, of course, that a great shift occurred in the labour force, from primary production into secondary and tertiary production. The most remarkable expansion (in terms of employment) took place in the tertiary sector. In France in 1800 that sector employed only 10 percent of the workers, while in 1970 it accounted for about 47 percent. Tertiary employment in the United States rose from 15 percent of the working population in 1820 to 65 percent in 1970. The examples in the Table show how other countries compare with

Historical Changes in the Distribution of Employment Among the Three Sectors of Production
(percentage of total labour force)

	primary*	secondary†	tertiary‡
France			
1800	80	10	10
1856	53	24	23
1901	42	30	28
1936	37	30	33
1959	24	38	38
1967	16	39	45
1970	13.3	39.7	47
United States			
1820	73	12	15
1850	64	18	18
1900	38	28	34
1940	18	34	48
1960	9.1	34.2	56.7
1967	5.5	33.4	61.1
1970	4	31	65
Turkey			
1927	82	8	10
1935	82	8	10
1950	85	7	8
1968	73	11	16
Sweden			
1965	12	42	46
1969	8.8	40.3	50.9
India			
1901	72	12	16
1967	72	14	14
Japan			
1920	54	25	21
1940	44	30	26
1950	48	27	25
1967	23	34	43
Great Britain			
1881	13	50	37
1921	10	47	43
1938	5	44	51
1950	4	50	46
1967	5	54	41
1969	1.7	47.0	51.3

*Agricultural and pastoral employment. †Employment in manufacturing. ‡Employment in service trades, teaching, administration, and other activities not included in the primary and secondary sectors.
Sources: National statistical yearbooks.

France and the U.S. in this respect. In India or Turkey today, as in France or the United States in the early 1800s, it is still necessary for three people out of four to work in agriculture in order to maintain the population at a low level of subsistence. This does not mean that there is no room for a tertiary sector even in very poor countries; there are always festivals and holidays (in India, funerals) that call for activities not related to feeding the population; there are the arts; and, of course, there are the irreducible needs of administration and law enforcement.

TECHNOLOGY AND UNEMPLOYMENT

The great shrinkage of employment in the primary sector is one of the most important phenomena of modern history. Men who abandon the soil must change not only their means of livelihood but their residence and their way of life as well, and all in the course of a single lifetime. This migration from a peasant culture to an urban industrial one has involved many millions of people. For a long time the migration went chiefly toward the factories; as recently as 1930 it was still possible to imagine that the world of the future would be one immense factory. Now it is clear that although the output of the factories has continued to increase, factory employment has not. Employment in the secondary sector in the United States reached its peak in the period from 1920 to 1970, at about one-third of the labour force, and since then the expansion has been in the tertiary sector. Technological progress would have led to unemployment, or to much-reduced working hours, had it not been for the expanding tertiary sector.

Ever since the time of the Industrial Revolution, men have feared that new machinery would lead to unemployment. Such a fear is a natural reflex of the man who sees a machine replacing him in his work. This often led to violent attacks upon machines, as in England in the 18th and early 19th centuries. It now appears that the fear of technological unemployment is exaggerated.

Two kinds of technological progress

To understand the effect of technological progress on employment it is necessary to distinguish between two kinds of such progress. There is, first, what may be called intensive progress, involving an increase in the efficiency with which man exploits nature: in agriculture, for example, all of the techniques that make it possible to extract products from the soil in a shorter period of time per worker. Intensive progress obviously results in a reduced need for workers. But there is another kind of progress that one may call extensive progress, which enlarges the exploitation of nature. Thus, the discovery of raw materials enlarges the framework of the economy, and the industrial development based upon it brings about the creation of new kinds of employment. The invention of a new product, such as the automobile or the radio, has the same effect. In France it has been estimated that more than 1,000,000 persons make their livings from the automobile, whether as factory workers, garage mechanics, or drivers.

Intensive progress allows man to satisfy his existing needs with less labour; extensive progress, on the other hand, satisfies needs that either did not exist or were not satisfied previously, and, consequently, creates new jobs. But the two kinds of progress do not automatically balance each other. In an economy that is not highly developed, technological progress in agriculture may allow the population to consume more food. In a more developed economy, one in which standards of living are already high, the same progress will not lead to increased food consumption; the end result may be glut, with lower farm prices and unemployed agricultural workers. In a period of general technological progress, employment is likely to be unstable because of the necessity for production to adapt itself to the changing demands of the market.

The consumption of primary, or agricultural, goods eventually reaches a point of saturation. The consumption of manufactured products passes through a phase of increase and then one of relative decline, except when new products appear on the market that have the extensive effect described above. It is the tertiary sector, however, that absorbs most of the manpower freed by technological progress in the other sectors. Employment as a whole does not, over the long run, decline because of technological progress. It is easy to show that in wealthy countries neither the size of the work force nor the length of the work week has had any tendency to decline in the last 50 years. There have been economic crises resulting in unemployment, but the unemployment was eventually absorbed. A dynamic economy requires that labour be mobile enough to move out of sectors in which technological progress has reduced the need for manpower and into sectors in which labour is in short supply. This migration is not pleasant for those who are caught up in it, since it is inevitably accompanied by some degree of unemployment or underemployment.

MEASURING UNEMPLOYMENT

The condition of unemployment seems simple enough: it is being without a job. But for the statistician unemployment is difficult to define precisely. The definition used in compiling the statistics of the United States and other de-

veloped countries treats as unemployed anyone who is capable of working and is actively seeking work but is unable to find it. Before a person can be unemployed in this sense he must be an active member of the labour force in search of a job. Students and housewives perform work, but they are not considered to be employed unless they are paid; by the same token they are not considered unemployed unless they are actively seeking gainful jobs. Any person who, because of handicap or discouragement, has not tried to find work—even if only to the extent of registering at a public employment office—will not be counted in the statistics of unemployment. There is a category of young people leaving schools or universities, who do not find jobs commensurate with their education; in some countries the government recognizes them as unemployed and tries to assist them.

Underemployment. Outside of the very clear category of unemployment, there is a whole gamut of degrees of nonemployment and underemployment. Until recently, unemployment was not distinguished from nonemployment with any degree of clarity. Numerous persons were not employed or, more often, were underemployed—that is, working here and there at one thing or another for a space of hours or days or months without a precise contract. A period of inclement weather would put day labourers out of work, and they received no aid of any kind. In the rural sectors of many countries today there appear to be considerable numbers of persons who may be called underemployed because they do not have full-time work even though they are part of the labour force.

Measurement of part-time unemployment is not too difficult and is done in a number of countries. In 1964, when the Swedish unemployment rate was 1.6 percent, involuntary part-time unemployment accounted for another 1.4 percent. Italy classifies as underemployed those persons who have worked fewer than 33 hours in the survey week. When the Italian unemployment rate was 3.4 percent in 1964, such persons accounted for another 1.7 percent. In West Germany, persons working up to 25 hours per week who desire more hours are classified as unemployed. In the United States, involuntary part-time workers averaged 2,443,000 in 1970 as compared with an average of 4,088,000 fully unemployed. In Japan only 590,000 persons (about one percent of the labour force) were counted as unemployed in 1968, but 1,496,000 employed wished to find part-time employment and 1,768,000 were seeking regular jobs.

Concealed unemployment

In modern industrial societies the statistics of employment sometimes conceal substantial amounts of unemployment or underemployment. This is because the unemployed may be brought into factories where they are not needed and have insufficient work to do. If the economy is a competitive one, each firm or enterprise will have an incentive to keep its operations efficient enough so that it makes a profit, and this will tend to prevent the hiring and maintaining of surplus workers. But where enterprises are subsidized or owned by the state, it is necessary to develop administrative and political controls to assure that work is carried on efficiently and in the interest of the public. These controls do not work as well in enforcing efficiency as those of a profit-making economy; they are often biased, erratic, or subject to the interference of pressure groups, politicians, or bureaucrats. In the Communist countries of Europe there has been much discussion, since the 1950s, of the problem of running socialized enterprises efficiently; the planners and economists of those countries have conceded that enterprises that are not subject to the pressures of the market have more difficulty than others in adjusting their production to the needs of consumers, and consequently in working out optimal plans and raising the efficiency of their labour force. It is possible to reduce unemployment and underemployment on a national scale, even to the point of an apparent shortage of labour, while at the same time permitting a considerable amount of misemployment and its accompanying inefficiency.

Rates of unemployment. During the Great Depression of the 1930s, unemployment in the Western capitalist countries reached very high rates. In the United States, the proportion of unemployed as counted by the government statistics reached 25 percent of the labour force in 1933. In Sweden unemployment reached 25 percent twice—in 1921 and 1931. In Great Britain the rate was above 15 percent in those two years. Since World War II, unemployment rates have been far below catastrophic levels. They have been higher in the United States than in other industrial countries: in the 1960s unemployment in the U.S. ranged between a high point of 6.7 percent in 1961 and 3.5 percent in 1969. In Sweden, from 1960 to 1970 unemployment remained below 2.2 percent. In Great Britain, from 1957 to 1970, the highest figure was 2.5 percent. In France, in the same period, unemployment exceeded 2 percent only once. (The figures are not precisely comparable as between countries because of differences in the way in which unemployment is measured.) In the Federal Republic of Germany and in Switzerland unemployment was so low as to be negligible, and the three countries imported large quantities of manpower, net immigration in 10 years amounting to about 3,000,000 persons in West Germany, 2,500,000 in France, and 500,000 in Switzerland. In Japan, unemployment moved downward from 1.8 percent in 1955 to less than one percent throughout most of the 1960s.

Varying incidences of unemployment

Even when the overall rate of unemployment is low, it remains a serious social problem because its incidence is not the same for all categories of workers. While some may be only temporarily out of work or passing from one job to another, others may be unemployed for long periods of time. The statistics published by most countries do not distinguish among categories of unemployed workers. The United States Bureau of Labor Statistics, however, has studied the incidence of unemployment quite thoroughly. Its figures show that women have more unemployment than men, young people more than adults, and persons without education or special skills more than educated persons. Thus, in 1964, when the overall rate of unemployment was 5.2 percent, the rate for men with less than five years of schooling was 9.4 percent. The difference in the rates had been much less in 1950, suggesting that employment opportunities for those with little education had declined. Blacks in the United States suffer more severely from unemployment than do whites. In 1970 the jobless rate for black teenagers was 30.1 percent as compared to 13.6 percent for white teen-agers; the rate for black men was 7.3 percent as compared to 4.0 percent for white men; and the rate for black women was 9.3 percent as compared to 5.4 percent for white women.

The most damaging unemployment, from a personal or family point of view, is that which persists for a long time. In 1970, about 16 percent of all U.S. unemployed had been out of work 15 weeks or more, and 6 percent for 27 weeks or more. Middle-aged workers and blacks were disproportionately represented among the long-term unemployed.

In all of the industrially developed countries, governments now collect increasing quantities of information on the labour force. Concise statistics, however, cannot adequately portray the whole spectrum of employment problems. Some workers remain in the same occupation, the same business, and even the same firm for decades, while others change jobs frequently. Some work at more than one job at a time. Some have seasonal jobs. Some work only a few hours a month, while others may work 60 or 65 hours a week. Some pass through a number of jobs searching for the one that is agreeable and pays well. A substantial number of persons cannot be easily classified as inside or outside of the labour force. Many who are outside the active labour force might be drawn in if jobs were obtainable or if help were available to assist them in qualifying for jobs. This sort of hidden unemployment may be more serious than overt unemployment if personal and social handicaps are combined with a shortage of work opportunity.

GOVERNMENT EFFORTS TO INCREASE EMPLOYMENT

Since World War II the governments of most industrial countries have become committed to a program of re-

ducing unemployment and underemployment. The constitution of France explicitly charges the state with assuring full employment. Similar goals have been affirmed with varying degrees of concreteness in other countries. The Charter of the United Nations makes full employment a major objective of its members.

Approaches to employment policy

Governments have followed various policies in pursuit of full employment. One general approach is to try to improve the supply of manpower. Another is to alleviate the effects of unemployment and underemployment. A third approach seeks to maintain economic activity at a high level through monetary and fiscal policies. Finally, there is the method of economic planning tried out in France, where the government planning commission sets targets for the various sectors of the economy that are linked to forecasts of manpower, in the belief that the difficult and complex problems of employment and unemployment cannot be separated from other problems of economic and social development (see ECONOMIC PLANNING).

Manpower policies. The first consistent efforts to deal with the instability of employment caused by technical and economic progress involved providing information on the state of the labour market and on the qualifications of those seeking work. This was done in the hope of directing job seekers more efficiently to existing jobs or of helping them to prepare themselves for the occupations and careers in which manpower seemed likely to be needed. Public employment services developed, particularly in the 1930s, that sought to direct applicants more efficiently to existing jobs or to help them prepare for occupations and careers in which manpower seemed likely to be needed. Most large industrial nations now have such agencies, designed to bring together the two sides of the labour market. Information leads to guidance and guidance to education and training, or retraining. The rapid progress of technology in recent times has required a perceptible raising of the qualifications of workers: skilled labourers have increased relatively to the unskilled, technicians relatively to skilled labourers, and engineers relatively to technicians. Governments have therefore endeavoured to provide or subsidize training programs for those who could benefit from them.

Often unemployment is a regional matter. Attempts have been made to mitigate it through regional development programs in a number of countries.

Ameliorative policies. All of the developed countries try to soften the impact of unemployment through unemployment compensation and other forms of social insurance. In some cases, governments seek to induce employers to retrain workers for new jobs rather than lay them off.

The maintenance of aggregate demand. The use of monetary and fiscal policies to keep the economy functioning at a high level of employment has been undertaken in a number of countries since the Depression of the 1930s (see FISCAL AND MONETARY POLICY). In periods of recession or of growing unemployment the government may increase aggregate demand by expanding the money supply or by increasing its own spending. This approach had some success in the United States during the 1960s, when a major income tax reduction, tax incentives for business investment, and a great increase in federal expenditures for both war and nonwar purposes brought the unemployment rate down to 3.5 percent in 1969. In the years that followed, however, the country was faced with serious inflation; in the effort to stabilize prices it had to incur rising unemployment.

National economic planning. The problems of unemployment and underemployment cannot be divorced from other problems of economic and social development. In the 1950s and 1960s the French government undertook national planning on a broad scale and succeeded in maintaining rapid economic growth, full employment, and a stable currency. The French experiment compared favourably with experience in Britain, where fiscal and monetary measures failed to overcome a persistent stagnation of the economy, and with the U.S. experience of inflation and recurring unemployment. It seemed likely that employment policies in most industrial countries would be oriented in the future toward a context much broader than that of placement bureaus, labour mobility, and fiscal and monetary policy.

Indications of a new, more comprehensive approach to problems of unemployment might be seen in the studies of the United States economic commissions and of the Organization for Economic Co-operation and Development. A report by the OECD, *The Growth of Production 1960–1980: Experience, Perspectives and Problems of Economic Policy* (1970), made estimates of employment for a number of countries in 1980, based on forecasts of general economic development. It is likely that this approach to economic planning will become the general tendency in all industrial countries in the future. It will also, one may hope, cease to be tied to a specialized discipline of econometrics. The speed with which society is evolving, the variety of possible ways of life, and the unforeseeable changes in the human condition make it increasingly necessary to view economic phenomena as only part of civilization, inseparable from the rest.

BIBLIOGRAPHY. Among the most important general works on employment and unemployment in modern industrial society are: JOHN MAYNARD KEYNES, *The General Theory of Employment, Interest and Money* (1936); and WILLIAM HENRY BEVERIDGE, *Full Employment in a Free Society* (1944). The distinctive French approach to employment forecasting is set forth in the *Rapports généraux de la commission de la main-d'oeuvre* of the COMMISSARIAT GENERAL DU PLAN, available only in French.

The specialized literature is enormous. Notable examples include: the publications of the INTERNATIONAL LABOUR ORGANIZATION, particularly its *Manpower Aspects of Recent Economic Developments in Europe* (1969); the publications of the ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT in the "Reviews of Manpower and Social Policies" series; and in *Employment and Earnings*, a monthly report issued by the UNITED STATES BUREAU OF LABOR STATISTICS.

Some recent works on the changing structure of the labour force include: SOLOMON BARKIN (ed.), *Technical Change and Manpower Planning: Coordination at Enterprise Level* (1967), consisting of 28 case studies dealing with 8 countries; VICTOR FUCHS, *The Service Economy* (1968), an analysis of the expanding tertiary sector in the United States; A.J. JAFFE and J.N. FROOMKIN, *Technology and Jobs: Automation in Perspective* (1968); and REID HUNTER and FRANCIS M. BODDY, *Labour Problems of Technological Change* (1970), a thorough study of the British situation.

(J.Fo.)

Enamelwork

Enamel, in art, is a vitreous glaze or combination of vitreous glazes fused on a metallic surface. The general term enamels is applied to metal objects that have this material as the principal decoration.

Enamels have been used to decorate the surface of metal objects, perhaps originally as a substitute for the more costly process of inlaying with precious or semiprecious stones but later as a decorative medium in their own right. Whereas painting on metal has a short life and, even when new, is overshadowed by the brilliance of the polished metal, enamelling provided the craftsman with a means of giving to the surface of metal a durable, coloured, decorative finish. With the painted enamels of the Renaissance and the portrait miniatures of the 17th century, the technique reached its most ambitious and artistic form; for here the craftsman attempted to create his version of an oil painting, using a metal sheet instead of a canvas and enamels instead of oil paints. While this medium undoubtedly has its limitations—few painted-enamel plaques of the Renaissance, for example, are much more than one foot square—it may be claimed that, whereas oil paints on canvas will fade and darken, the colours of enamels are permanent. Nevertheless, relatively few creative artists of distinction have chosen to work in this medium; thus, its prime function has tended to be purely decorative.

Few types of metal object have not, at some period, been enriched with enamelled decoration. Throughout history, jewelry has been made more colourful by the application of enamels. Similarly, arms and armour, horse trappings, and even domestic items, such as mirrors

Objects typically decorated with enamelwork

and hanging bowls, were embellished with enamel decoration. Throughout the Middle Ages, both secular and ecclesiastical objects, such as chalices, cups, reliquaries, caskets, crosiers (a staff carried by bishops and abbots as a symbol of office), and spoons, were elaborately enamelled. With the advent of painted enamels in the Renaissance, tableware was completely covered with enamel, and painted-enamel panels were used to decorate the ceilings and walls of rooms in the châteaus of France. Following upon the invention of the domestic table clock and of the watch in the 16th century, enamelling became one of the most popular forms of decoration for the dials and cases; by the 18th century, items of the drawing room, such as snuffboxes, etuis (cases for small articles like scissors and needles), tea caddies, candlesticks, scent bottles, and thimbles, were frequently made of enamel.

Among the objects decorated with enamels in east Asia are vases, incense vessels, teapots, suits of armour, and sliding doors.

This article treats the material, techniques, and history of enamelwork.

MATERIALS AND TECHNIQUES

Enamel is a comparatively soft glass, a compound of flint or sand, red lead, and soda or potash. These materials are melted together, producing an almost clear glass, with a slightly bluish or greenish tinge; this substance is known as flux or frit—or, in France, *fondant*. The degree of hardness of the flux depends on the proportions of the components in the mix. Enamels are termed hard when the temperature required to fuse them is very high; the harder the enamel is, the better it will withstand atmospheric agencies, which in soft enamels first produce a decomposition of the surface and ultimately cause the breakup of the whole enamel. Soft enamels require less heat to fire them and consequently are more convenient to use, but they do not wear so well, especially if subjected to friction.

Clear flux is the base from which coloured enamels are made, the colouring agent being a metallic oxide, which is introduced into the flux when the latter is in a molten state. The brilliance of an enamel depends on the perfect combination of its components and on maintaining an equal temperature throughout its fusion in the crucible. The colour of many enamels is achieved by a change in the proportion of the components of the flux rather than by a change in quantity of the oxide. For example, turquoise-blue enamel can be obtained from the black oxide of copper by using a comparatively high proportion of carbonate of soda; in the same way, a yellowish-green enamel can be obtained from the same black oxide by increasing the proportionate amount of red lead.

Clear flux is also used to make opaque enamels; the addition of calx, a mixture of tin and lead calcined, renders translucent enamels opaque. White enamel is produced by adding stannic and arsenious acids to the flux, the quantity of the acid affecting the density, or opacity, of the enamel.

Preparation of the enamel

The heated enamel, after being thoroughly stirred, is usually poured out onto a slab and allowed to solidify into cakes of approximately four to five inches (10 to 13 centimetres) in diameter. For use, each cake must be pulverized into a fine powder with a pestle and mortar; the powder then has to be subjected to a series of washings in distilled water until all the floury particles are removed. The metal, on which the powdered enamel is to be spread, is cleansed by immersion in acid and water. All trace of the acid is then removed by washing and by drying in warm oak sawdust. After the wet powder has been spread on the metal, it is allowed to dry in front of the furnace before it is carefully introduced into the muffle of the furnace (a compartment protected from the flame), where it is heated to the point at which it fuses and adheres to its metal base. The firing of enamel takes only a few minutes, and the object is then withdrawn and allowed to cool.

The various techniques practiced by craftsmen in the past differ mainly in the methods employed in preparing the metal to receive the powdered enamel.

Cloisonné. In the cloisonné technique, thin strips of metal are bent and curved to follow the outline of a decorative pattern; they are then attached, usually soldered, to the surface of the metal object, forming miniature walls that meet and create little cells between them. Into these cells, the powdered enamel is laid and fused. After it has cooled, the surface can be polished to remove imperfections and to add to the brilliance. The cloisonné technique is particularly suited to objects made of gold, such as jewelry.

Millefiore glass and champlevé enamel on a hanging bowl in the Sutton Hoo burial ship. Anglo-Saxon, *c.* 625–660. In the British Museum. Diameter 6 cm.
By courtesy of the Trustees of the British Museum

Champlevé. This process is the opposite of the cloisonné technique: instead of building up on the surface of the metal object, the surface is gouged away, creating troughs and channels separated by thin ridges of metal that form the outline of the design. The troughs are filled with powdered enamel and fused. The champlevé technique requires a thick metal base and therefore is used on copper and other base metals.

Basse-taille. This technique is a sophisticated extension of the champlevé method, for again the metal surface has to be cut away and filled with enamel, but here there are two major differences. First, within the area that has been cut away to receive the enamel, a design or figural composition is chased (chiselled), or sometimes engraved, in low relief. Because the highest point of the relief is below the general surface of the surrounding metal, the enamel, which is level on its outer surface, lies in varying thicknesses over the modelled surfaces of the low relief. Second, because the coloured enamels used in this technique are translucent, the composition of the low relief shows through; and, since the metal used is normally gold or silver, the light is reflected back through the translucent enamelling, adding a brilliant tonal quality to the enamel, just as sunlight enhances the beauty of a stained-glass window. The effect of the reflected light varies according to the thickness of the enamel lying over the undulating surfaces of the low relief; consequently, an impression of plasticity and of three-dimensional modelling is created by the subtle variations in tonal strengths of the enamel colours, which range from bright highlights to the rich tones of the deep recesses.

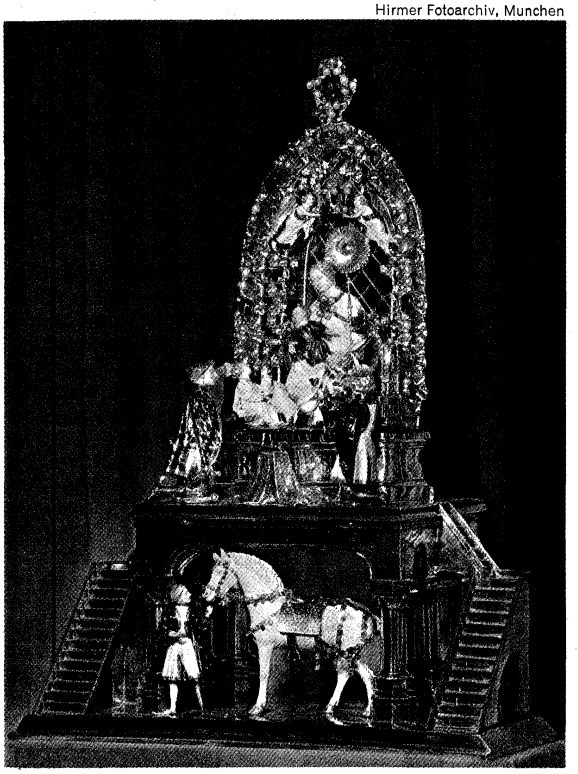
Stained-glass window in miniature

Plique-à-jour. The plique-à-jour technique is designed to produce an effect of a stained glass window in miniature through the use of translucent enamels. The technique is exactly the same as cloisonné enamelling except that the strips of metal forming the cells are only temporarily attached—not soldered—to a metal base to which the enamel will not stick. After the enamel is fused and sufficiently annealed, the metal sheet, usually aluminum–bronze, is removed with a few light taps,

leaving a network of metal strips filled with enamel "windows." The enamels can be carefully polished to enhance their appearance.

Encrusted enamelling (émail en ronde bosse). Encrusted enamelling is the term used to describe the technique of enamelling the irregular surfaces of objects or figures in the round or in very high relief. Both opaque and translucent enamels are applied to these small-scale sculptural objects, which are usually made of gold. The great technical problem is to devise methods of supporting and protecting these objects during the firing. Frequently, plaster of paris is used to envelop parts of the object, leaving exposed only those parts on which enamel is to be applied and fused.

Hirmer Fotoarchiv, Munchen



"Goldenes Rössel," gold shrine with encrusted enamelling, French, *c.* 1403. Given to Charles VI of France by his queen, Isabella of Bavaria. In the pilgrimage church of Altötting, West Germany. Height 62 cm.

Painted enamels. This technique differs fundamentally from the preceding five in that the various coloured enamels are not separated from each other by metal strips or ridges. Although these enamels are still applied in their wet, powdered state, the adjacent patch of coloured enamel is first allowed to dry to avoid one running into the other and so blurring the outline between them.

The metal generally used in this technique is copper. It is cut with shears into a plate of the size required and slightly domed with a burnisher or hammer, after which it is cleaned with acid and water. The enamel is laid equally over the whole surface both back and front, and then the object is fired. The first coat of enamel being fixed, the design is delineated by drawing with a needle through a layer of wet white enamel or any other that is opaque and most advantageous for subsequent coloration.

Grisaille enamels

In the case of grisaille enamels, the white is mixed with water, turpentine, spike oil of lavender, or essential oil of petroleum and painted over a dark-enamel ground. Light areas of the design are painted thickly; gray areas, thinly to allow the dark ground to tone the white pigment. The technique creates a strong contrast between light and shade, creating an impression of low relief. The scenes in grisaille are sometimes rendered more subtly by hatching, executed with a pointed tool or needle to reveal the dark enamel beneath.

In coloured painted enamels, enamel colours are spread over the grisaille treatment; when fired, parts of the surface are heightened by touches of gold, usually painted in thin lines, like hatchings. Other parts can be made more brilliant by the use of foil, over which the transparent enamels are placed and then fired.

HISTORY

Ancient Western. The origins of the art of enamelling are uncertain. While there is archaeological evidence that glass was being made from the 3rd millennium BC in western Asia and from the 15th century BC glass vessels were undoubtedly being made in Egypt, there is no proof that enamelling on metal was practiced in either Asia Minor or Egypt until after the time of Alexander the Great (died 323 BC).

Mycenaean enamelling

Perhaps the origins of the art are to be found on Mycenaean metalwork of the 13th to 11th centuries BC. Six gold rings, excavated from a Mycenaean tomb of the 13th century BC at Kouklia (near Old Paphos), in Cyprus, are decorated with a cloisonné technique that suggests an intermediary stage between inlay and true enamelling. Scientific examination has shown that the different coloured enamels were not in the form of powder when they were inserted into the cloisons before being fired and fused together; rather they were in the form of fragments of coloured glass. Unfortunately, no report exists of any scientific examination of a more accomplished example of Mycenaean enamelling—the decoration on the gold sceptre found in a royal tomb at Kourion Kaloriziki, in Cyprus—but it is generally believed that this is true enamelling and datable to the 11th century BC.

If true enamelling existed in Mycenaean work, it would be reasonable to expect the technique to have been inherited by the Greeks and transmitted by them to the rest of Europe, perhaps by way of the colonies on the north shore of the Black Sea and in the south of Italy. Unfortunately, however, there is a long gap between the Mycenaean enamels and the Greek gold jewelry of the 6th–3rd centuries BC, which is sparingly enamelled, often having no more than touches of blue and white enamel enclosed by thin gold wire openwork (filigree).

Until recently the most ancient examples of enamelling outside Mycenaean art were said to be on ornaments discovered in a cemetery in the Kuban, close to the Caucasus, variously dated between the 9th and 7th centuries BC; but the most important of these Kuban enamels, the famous Maikop belt buckle (the Hermitage, Leningrad) depicting a griffin attacking a horse, is now regarded by Russian experts as a forgery. Consequently, the earliest enamelling from south Russia may date from the 3rd or 2nd century BC.

Celtic enamel decoration

A slightly earlier date is given to a number of excavated bronze objects of western European origin, which are said to bear the remains of cloisonné enamel decoration. Until this early Celtic material has been scientifically examined and proved to be true enamel as distinct from inlaid coral, cut stone (chiefly lapis), or coloured glass applied cold, theories about it remain open to question. At the present time it is a matter of conjecture what link, if any, may have existed between the enamellers in south Russia and those Celtic craftsmen who by the 3rd century BC, if not earlier, were using red enamel in place of coral inlay.

During the Roman period, enamelling—both cloisonné and champlevé on bronze—was carried on almost entirely in those old Celtic areas that had become the northern provinces of the Roman Empire. It may well be to these provincial works that a passage from the works of the Greek philosopher Philostratus, 2nd century AD, refers. The author, describing a boar hunt at which the riders appear with horse trappings ornamented in bright colours, writes:

> It is said that the barbarians in the ocean [*i.e.*, the Celtic tribes] pour these colours into bronze moulds, that the colours become as hard as stone, preserving the designs.

This is a fair description of the process of champlevé enamel and suggests that the technique, in use in the British Isles, was not practiced at the time in Greece or

Italy. Enamelled horse trappings such as Philostratus describes have been found in many places in the British Isles. This type of Celtic enamelling of the Roman period lived on in northwest Europe, particularly in Ireland, until as late as the 12th century. Some of its more striking effects seem to be derived from Roman glassmaking practices, particularly its use of millefiori glass, a mosaic of very thin glass rods of different colours and shapes fused together and then cut into thin sections, which the Celtic craftsmen fused into a ground of coloured enamel.

Medieval. *Byzantine.* The most dramatic development in the history of enamelling took place in the Byzantine Empire between the 6th and 12th centuries, a period during which only the cloisonné technique—almost exclusively executed on gold—was in use. At their zenith in the 10–11th centuries, Byzantine enamellers created delicate, highly expressive miniature scenes in a great range of colours that shine like jewels. The masterpiece of this period is the altar screen "Pala d'Oro" in St. Mark's, Venice, believed to have been brought from Constantinople to Venice about 1105. The quality of Byzantine enamelling began to decline in the late 12th century.

Islāmic. There is no direct evidence that enamelling on metal was practiced at any Islāmic centre in western Asia. Those scholars who would argue that the technique of Byzantine gold-cloisonné enamelling originated in Syria before the 7th century AD can point to just one object, which, only on inconclusive stylistic considerations, may be associated with Umayyad Syria rather than with Byzantium. Only one other enamelled object has survived with strong Islāmic connections: a dish with an Arabic inscription referring to an Artuqid Prince, who reigned AD 1114–44. The enamel technique is cloisonné, but with bronze wires soldered onto a copper base. As no other examples have been found and as the inscription in Arabic indicates an imperfect knowledge of the language, it may be the work of a Byzantine craftsman working in the Artuqid kingdom.

Western European. As early as the 7th century, according to some scholars, Byzantine work was being copied by Lombard craftsmen in northern Italy; later it was imitated in Sicily and other parts of Italy—even perhaps in England, where the famous Alfred Jewel, made to the order of the English king Alfred the Great in the 9th century AD, shows strong Byzantine influence. In the Ottonian period (AD 936–1002), gold-cloisonné enamelling seems to have flourished in eastern France, and in the Rhineland, particularly among the goldsmiths working at Essen and in the workshops of Archbishop Egbert (AD 937–993) at Trier.

In western Europe cloisonné enamelling was abandoned in the 12th century, in favour of the champlevé technique executed on a base metal such as copper or bronze. This revival may have taken place first in Spain, in the valleys of the Rhine and the Meuse, or in France at Limoges; but, by the middle of the century, expert craftsmen in these centres—and in England—had established it as one of the foremost mediums for artistic expression in the Romanesque style. In the Mosan school, the famous 12th-century enamellers Godefroid de Claire at Liège and Nicholas of Verdun created champlevé enamelwork of unprecedented merit. The best work from Limoges was executed at the turn of the 12–13th century; thereafter, the output was commercialized and standards fell steadily throughout the 13th and 14th centuries.

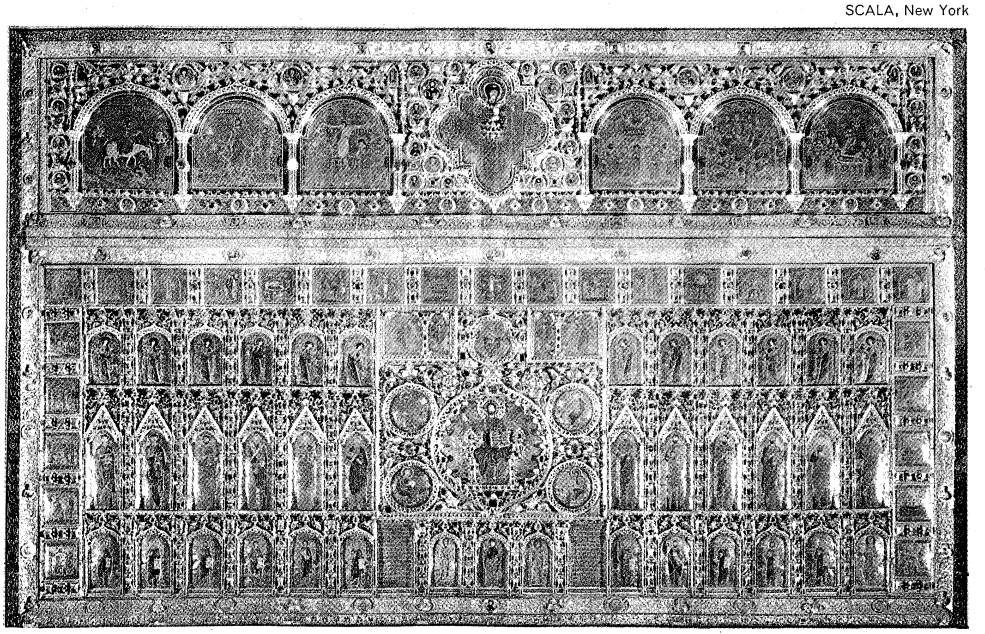
Basse-taille enamelling

In the late 13th century, gold and silver objects were again decorated with enamel but in a new technique, basse-taille enamelling. The earliest surviving dated example was made in Italy in 1290. Throughout the following century, Italian goldsmiths, particularly from Siena and Florence, produced pictorial masterpieces in this medium. The technique was especially favoured in Spain and France. No more accomplished example has survived than the "Royal Gold Cup" (British Museum), commissioned by the brother of the French king Charles V about 1380. The sides and the cover have scenes depicting the life and martyrdom of St. Agnes in the most glowing rich colours and elegant draftsmanship of the period. The great era of basse-taille enamelling ended with the Renaissance, though it remained popular in Spain and southern Germany, chiefly in Augsburg, to the middle of the 17th century.

Plique-à-jour and encrusted enamelling

15th century to the present: European. Under the patronage of the courts of France and Burgundy in the late 14th and first half of the 15th centuries, goldsmiths devised new and more audacious methods of enamelling. Using translucent coloured enamels, they created the effect of stained-glass windows in miniature by the technique known as plique-à-jour. One of the loveliest pieces is the silver-gilt Merode beaker of Flemish or Burgundian origin, probably *c.* 1430–40, decorated with two bands of enamels set in tiny windows with Gothic tracery (Victoria and Albert Museum, London). Employing another technique, encrusted enamelling, they created both large-scale three-dimensional compositions and miniature work to be worn as jewelry. Among the finest and earliest surviving examples is the Reliquary of the Holy Thorn (in the Waddesdon bequest in the British Museum): the Holy Thorn, set in a gem, is surrounded by the Last

SCALA, New York



"Pala d'Oro," altar screen of gold cloisonné enamel, Byzantine, 10th–12th centuries, reassembled with later additions in a Gothic frame in 1342–45. In St. Mark's Cathedral, Venice.

Copper dish with cloisonne enamelling in turquoise, cobalt-blue, red, yellow, and white; the Arabic inscription refers to the Ortokid prince of Amida and Hisn Kaifa, Da'ud ibn Suqmān (reigned AD 1114-44). In the Tiroler Landesmuseum Ferdinandeum, Innsbruck, Austria. Diameter 23.11 cm.

By courtesy of the Tiroler Landesmuseum Ferdinandeum, Innsbruck, Austria; photograph, A. Demanega

Judgment scene, in which all the figures (20) are enamelled, many of them being executed wholly in the round. The taste for this type of enamelled goldsmith work spread to all the courts of Europe; and, although the style changed several times, first from Gothic to Renaissance and then to Baroque, the essential extravagant toy-like quality remained. Of all the Renaissance goldsmiths who moved from court to court, and, through these movements, helping to create an international style, only Benvenuto Cellini wrote (*c.* 1560) a technical treatise on the subject.

Although the technique of painted enamels was probably first evolved by Flemish craftsmen about 1425–50 for the Burgundian court and perhaps developed by Venetian and north Italian enamellers between 1450 and 1500, the supremacy of the Limoges workshops was established by the beginning of the 16th century. For the next 100 years, French Mannerist art found talented expression in this medium, and, enjoying court patronage, the best Limoges enamellers strove to compete with other artists in decorating the rooms of royal palaces. Painting in grisaille was finally introduced at Limoges by about 1530–40.

Miniature painting in enamel colours

A new dimension was given to painted enamelwork about 1620–30 by a French goldsmith, Jean I Toutin of Chateaudun, and some rival craftsmen in Blois. Their achievement was to invent a highly skillful method for fine miniature painting in enamel colours on a white-enamel ground. Since the technique was admirably suited to the current enthusiasm for portrait miniatures, artists of distinction, such as Jean Pettitot, were employed by Charles I of England and the French kings to work in this medium.

With equal artistic skill, other French enamellers decorated items of jewelry, especially watchcases; and, by the second half of the 17th century, this craft had become centred on Geneva, where it continued to flourish into the 19th century. In England, particularly in the Midlands, the Continental style of painted enamelled "toys" was copied and produced on a large scale, but the technique of transfer printing on enamel was invented in England and brought to perfection at the Battersea (London) factory during 1753–56. The design was applied to the white-enamel ground by transferring to paper, and then to the surface to be decorated, an impression from an engraved metal plate that had been brushed with enamel colours. Throughout the 17th and 18th centuries, enamellers used the technique of fine miniature painting in enamels in Germany, Holland, England, and Russia in order to produce the "toys" of the fashionable world of society.

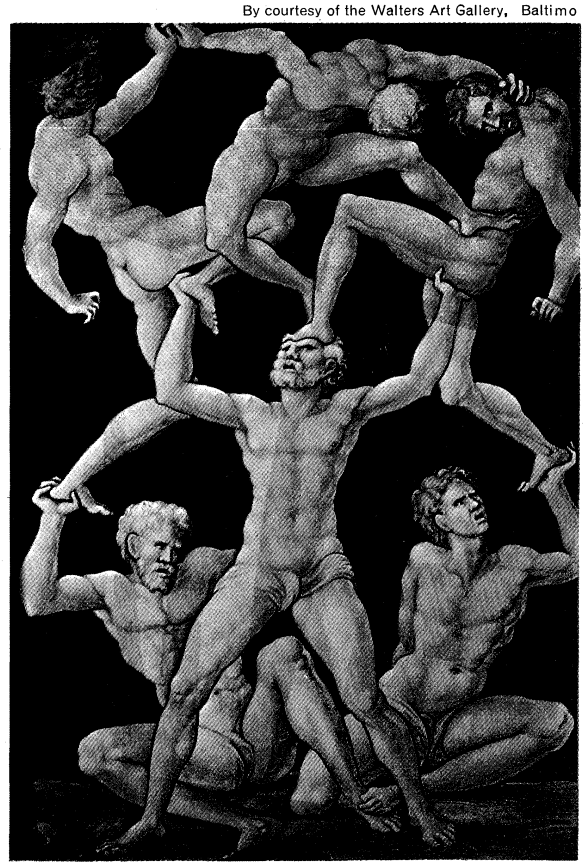
The technique called *en résille sur verre* flourished for only about 40 years (*c.* 1600–40), and few examples have survived. Yet it required an exceptional degree of skill. The technique consists of cutting the design in a medallion of glass, usually coloured, lining the incisions with gold and filling them with variously coloured enamels. The exponents of this kind of enamelling were mainly French.

Although surviving examples are rare, there is a distinctive group of brass objects, mainly candlesticks and andirons, which have green, blue, or white opaque enamelling. These objects were made in 17th-century England (perhaps in Sussex).

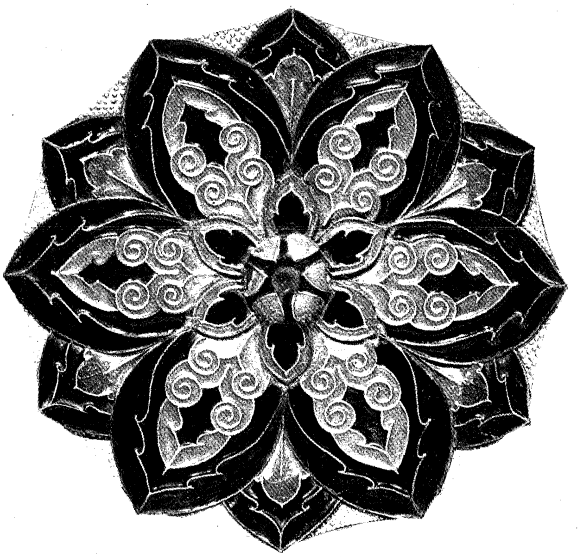
Almost all of the early enamelling techniques have continued to be used by goldsmiths in modern times—from the Parisian makers of gold snuffboxes in the 18th century to Carl Fabergé at the beginning of the 20th century. Art Nouveau jewellers, such as René Lalique, and modern artists, such as Georges Braque, Georges Rouault, and Gerda Flockinger, have kept alive the craft of enamelling and added to the multiplicity of its ingenious effects. (H.Ta.)

China. Enamels do not appear to have reached China until long after they were found throughout Europe. All authorities are agreed as to the Western origin of the art, which in all probability was introduced into China by traders or by travelling craftsmen. Although by the 5th century AD the Chinese were informed as to the production of glass—an essential material for the making of enamels—and were already highly skilled in the working of bronzes and other metals, there is no evidence that the art of enamelling was practiced before the T'ang dynasty (618–907). There is in the Shōsō-in (principal storehouse) at Nara, in Japan, a silver mirror, the back of which is decorated in cloisonné. It is generally agreed that the mirror is of Chinese origin, dating from the T'ang dynasty, as is certainly the case with many other

By courtesy of the Walters Art Gallery, Baltimo



"The Acrobats," painted enamel plaque in grisaille with flesh tints applied as washes or brushed over in serrated thin lines of red, after a design of Juste de Juste, a sculptor working at Fontainebleau. The work was executed at Limoges, *c.* 1550. In the Walters Art Gallery, Baltimore. 43.5 cm × 30.48 cm.



Mirror back, cloisonné enamel on silver, Chinese, T'ang dynasty 9th century. In the Shōsō-in, Nara, Japan. Diameter 18.7 cm.
By courtesy of the Shoso-in, Nara, Japan

Earliest known enamelwork

objects in the collection. At present, this is the only known Chinese enamelled ware made before the 14th century; but it can be safely assumed from this piece that the art of cloisonné was developed to a respectable height in the T'ang dynasty. It appears that cloisonné work was well established in China at the end of the 14th century and that Byzantine work of similar character was also so well known as to invite comparison with the native product. The former may well have served as an example for Chinese craftsmen. As one scholar points out:

> The workmanship presents occasionally . . . striking resemblances with certain enamels of the Byzantine school; the mixture of different enamels inside the wall of the same cell, the employment of gold incrustations in the treatment of the fingers and the hands, etc.

Active trade and cultural intercourse between the Near East and China during the Yüan (Mongol) dynasty must have been the reason for this revival of enamelwork, which then flourished through the Ming and Ch'ing dynasties (1368–1644 and 1644–1911, respectively).

Chinese enamels fall into three categories—cloisonné, champlevé, and painted. In none does the technique vary appreciably from that employed in Western countries.

Cloisonné. The earliest example of cloisonné enamel that can be authentically associated with east Asia is the silver mirror in the Shōsō-in mentioned above. Its cloisonné back is decorated in a design of a six-petalled blossom in three layers, the tips of the outer rows of petals forming 12 points of the mirror. The piece is regarded as a T'ang dynasty work. Apart from this, the sequence of known Chinese enamels begins in the Yüan period, and the earliest recorded marks belong to the reign of the last emperor of that dynasty (1333–67). The great period of production is certainly that of the Ming dynasty, which followed.

Ming enamels

The mark most commonly found within this period is that of the Ching-t'ai reign (1449–57). The Ming enamels, bold in design with fine depth and purity of colour, were never surpassed in later epochs. The two shades of blue, a dark-lapis-lazuli tone and a pale sky blue with a very slight tinge of green, are particularly excellent. The red is of dark-coral tint, and the yellow full-bodied and pure. Greens derived from copper are sparingly used. Black and white are the least successful, the former shallow and dull, the latter clouded and muddy. As fine as the Ming enamels are, however, there is an imperfection of technique, close examination revealing minute pitting in the enamels, which was caused by inadequate packing of the material, and some lack of polish in the surface. These technical defects, however, do not appreciably detract from the great artistic value of the Ming enamels.

A great revival of art industries took place under the patronage of the emperor K'ang-hsi (1661–1722), who, in 1680, established a series of imperial factories. He commissioned sets of incense vessels of cloisonné enamel for presentation to the numerous Buddhist temples founded under his auspices in the neighbourhood of Peking, as well as other objects for the honorific gifts that were characteristic of his enlightened reign. The enamels produced during his time are marked by an improvement in technical quality as compared with those of the Ming period; to a considerable extent they also retained the finer qualities of the Ming wares. In many cases the forms of ancient bronze vessels were revived and enriched with enamels.

The style of this reign persisted during that of K'ang-hsi's successor, Yung-cheng (1723–35), while the long rule of Ch'ien-lung (1736–96) was marked, in enamel as in the case of many other industrial arts, by a further perfection of technique but by a loss of much of the vigour of design and breadth of execution that distinguished the products of earlier periods. Modern enamels, although they are primarily imitations of older work, are more hurriedly made and therefore not so well finished as the older work.

Champlevé. Some of the most ancient enamel examples extant belong to this class, and examples in which both champlevé and cloisonné are employed are not uncommon.

Although the opaque enamels were more common, Chinese artisans occasionally used translucent enamels on a silver or gold base. The cloisonné back of the silver mirror in the Shōsō-in, for example, is decorated with transparent enamels, but important pieces such as this are rare. This technique more often appears in Chinese jewelry.

Painted enamels. The painted enamels of China, generally known, from the principal seat of their manufacture, as Canton enamels, are practically identical in technique with the Limoges and other painted enamels of

Canton enamels

Ming dynasty vase, cloisonné enamel, Chinese, *c.* 1500. In the British Museum. Height 41.5 cm.
By courtesy of the trustees of the British Museum

Europe. Specimens of the latter are known to have been taken to China by the missionaries of the late 17th and 18th centuries; they not only exercised direct influence on the Chinese ware but also, in some cases, were copied. Representations of European subjects, copies of engravings and armorial decorations, are also found. Painted enamels are termed by the Chinese *yang-tz'u* ("foreign porcelain"), the palette of colours used being the same as with enamelled porcelain, whose decoration under foreign influence is called *yang-ts'ai* ("foreign colours"). A ground of opaque enamel, generally white, is laid on the copper, and on this the colours are superimposed and fired. The best period of this art was the 18th century. Although imitations have continued to be made, nothing of real quality in this style was produced after the termination of the reign of Ch'ien-lung in 1796. The method has always been looked upon by the Chinese as alien in taste; in fact, a great part of the Canton enamels were made for export, not only on European commission but also for clients in India, Persia, and several other Asian countries.

Japan. The art of enamelling in Japan may date back to the 7th century. One enamelled metal piece, discovered in a tomb near Nara, which probably dates from the 7th century, seems to be of Japanese origin. The civic Taihō code, compiled in the 8th century, provides one official to be in charge of founding metals and "painted glass decoration." Subsequently, however, this art seems almost to die out until the 17th century. When Dōnin Hirata I (1591–1646) succeeded in making enamelled wares, having learned the technique from Koreans, his art was highly appreciated by Tokugawa Ieyasu, then the shogun of Japan, under whose patronage Hirata worked in Kyōto. There is a suit of armour with enamelled metal fittings ascribed to him, as well as enamelled metal fittings decorating sliding doors and lintels in the Katsura Palace, Kyōto. His family continued the trade until the late 19th century, making use, on a small scale, of both the cloisonné and champlevé methods. There was no further development of importance until Kaji Tsunekichi (1803–83) and his pupils established in Nagoya an important and successful manufacture of cloisonné, which obtained a considerable vogue, especially among foreigners.

While Kaji used brass cloisons and opaque enamel colours only, his successors used silver cloisons and succeeded in making both transparent and translucent enamels. They further modified the cloisonné process with remarkable ingenuity and produced work of great interest. First, they reproduced in minutely detailed cloisonné work realistic pictures, of trees and flowers, for example. This effort led them in the 1880s to produce lineless cloisonné enamels, which have all the beauty and brilliance of true cloisonné, with thick layers of enamel colours, but which, showing no trace of cloisons, permit a gradation of colours as in the less clear and brilliant painted enamels. The effect was achieved by taking off the cloisons before each firing, the process being repeated at least three times. The artists working for the factory of Namikawa Sōsuke of Tokyo in the late 19th century were most successful in this technique. Another Namikawa of Kyōto worked in true cloisonné. The factory of Jubei Ando of Nagoya has produced more variations. These developments have carried the art of enamel very far from the old traditions; while the skill and ingenuity of technique they evince may be appreciated, there is a danger of losing the artistry peculiar to the art of enamelling.

By courtesy of the Tokyo National Museum

Panel of Mt. Fuji seen through clouds, done in lineless cloisonné by Namikawa Sōsuke, 1893. In the Tokyo National Museum. 63 cm x 113.6 cm.

Lineless cloisonné

BIBLIOGRAPHY

Materials and techniques: KENNETH FRANCIS BATES, *Enameling: Principles and Practice* (1951); HERBERT MARYON, *Metalwork and Enamelling*, 4th ed. (1959), contains a useful bibliography; MARGARET SEELER, *The Art of Enameling* (1969).

Periods and centres of production: KLAUS WESSEL, *Byzantine Enamels from the 5th to the 13th Century* (1968); LUIGI MALLE, *Cloisonnés bizantini* (1970); SHALVA AMIRANASHVILI, *Medieval Georgian Enamels of Russia* (1964); MARY CHAMOT, *English Mediaeval Enamels* (1930); W.L. HILDBURGH, *Medieval Spanish Enamels and Their Relation to the Origin and the Development of Copper Champlevé Enamels of the Twelfth and Thirteenth Centuries* (1936); MARIE-MADELEINE GAUTNIER, *Ernaux limousins champlevés des XII^e, XIII^e, and XIV^e siècles* (1950); J.J. MARQUET DE VASSELOT, *Les Crosses limousines du XIII^e siècle* (1941); PAUL THOBY, *Les Croix limousines de la fin du XII^e siècle au debut du XIV^e siècle* (1953); KATIA GUTH-DREYFUS, *Transluzides Email in der ersten Hälfte des 14 Jahrhunderts am Ober-, Mittel-, und Niederrhein* (1954); PHILIPPE VERDIER, *Catalogue of the Painted Enamels of the Renaissance in the Walters Art Gallery* (1967), with an excellent bibliography; HENRI CLOUZOT, *Dictionnaire des miniaturistes sur émail* (1924) and *La Miniature sur émail en France* (1928); PIERRE F. SCHNEEBERGER, *Les Peintres sur émail genevois au XVII^e et au XVIII^e siècle* (1958); CHARLES BEARD, "Bavarian Enamels of the Seventeenth Century," *Connoisseur*, 97:267–271 (1936); EDWARD DILLON, "English Enamels on Brass of the Seventeenth Century," *Burlington Magazine*, 16:261 (1910); EGAN MEW, *Battersea Enamels* (1926), out of date in some respects; SANDOR MIHALIK, *Emailkunst im alten Ungarn* (1961; Eng. trans., *Old Hungarian Enamels*, 1961); HARRY GARNER, *Chinese and Japanese Cloisonné Enamels* (1962), with a full bibliography.

(Ed.)

Encyclopaedia

For over 2,000 years encyclopaedias have existed as summaries of extant scholarship in forms comprehensible to their readers. The word encyclopaedia, of Greek origin (*enkyklopaideia*), at first meant a circle or a complete system of learning; that is, an all-round education. When Rabelais used the term in French for the first time in *Pantagruel* (chapter 20), he was still talking of education. Paul Scalich, a German writer and compiler, was the first to use the word to describe a book in the title of his *Encyclopaedia; seu, orbis disciplinarum, tam sacrarum quam prophanum epistemon* . . . ("Encyclopaedia; or Knowledge of the World of Disciplines, Not Only Sacred but Profane . . ."), issued at Basel in 1559. The many encyclopaedias that had been published prior to this time had either been given fanciful titles (*Hortus deliciarum*, "Garden of Delights") or had been simply called "dictionary." The word dictionary has been widely used as a name for encyclopaedias, and Scalich's pioneer use of "encyclopaedia" did not find general acceptance until Denis Diderot made it fashionable with his historic French encyclopaedia, although "cyclopaedia" was then becoming fairly popular as an alternative term. Even today a modern encyclopaedia still may be called a dictionary, but no good dictionary has ever been called an encyclopaedia.

Encyclopaedia as a complete system of learning

The meaning of the word encyclopaedia has changed considerably during its long history. Today most people envisage an encyclopaedia as a multivolume compendium of all available knowledge, complete with maps and a very detailed index, as well as numerous adjuncts such as bibliographies, illustrations, diagrams, lists of abbreviations and foreign expressions, gazetteers, and so on. They expect it to include biographies of the great men and women of the present as well as those of the

past, and they take it for granted that the alphabetically arranged contents will have been written in their own language by many people and will have been edited by a highly skilled and scholarly staff. Yet not one of these ingredients has remained the same throughout the ages. Encyclopaedias have come in all sizes from a single 200-page volume written by one man to giant sets of 100 volumes or more. The degree of coverage of knowledge has varied according to the time and country of publication. Illustrations, atlases, and bibliographies have been omitted from many encyclopaedias, and for a long time it was not thought fitting to include biographies of living persons. Indexes are a late addition, and most of the early ones were useless. Alphabetical arrangement was as strongly opposed as the use of any language but Latin, at least in the first 1,000 years of publication in the West, and skilled group editorship has a history of scarcely 200 years.

In this article the word encyclopaedia has been taken to include not only the great general encyclopaedias of the past and the present but all types of works that claim to provide in an orderly arrangement the essence of "all that is known" on a subject or a group of subjects. This includes dictionaries of philosophy and of American history as well as volumes such as *The World Almanac*, which is really a kind of encyclopaedia of current information.

Encyclopaedias as reflections of the development of scholarship

An outline of the scope and history of encyclopaedias is essentially a guide to the story of the development of scholarship, for encyclopaedias stand out as landmarks throughout the centuries, recording much of what was known at the time of publication. Many homes have no encyclopaedia at all, very few have more than one, yet in the past 2,000 years at least 2,000 encyclopaedias have been issued in various parts of the world, and some of these have had many editions. No library has copies of them all; if it were possible to collect them they would occupy some two miles of shelf space. But they are worth preserving—even those that appear to be hopelessly out of date—for they contain many contributions by a large number of the world's leaders and scholars.

This article is outlined as follows:

I. The nature of encyclopaedias
II. Encyclopaedias in general
    The role of encyclopaedias
    Editing and publishing
III. The kinds of encyclopaedias
    General encyclopaedias
    Encyclopaedic dictionaries
    The modern encyclopaedia
    Encyclopaedias for special interests
    Children's encyclopaedias
    Specialized encyclopaedias
    Encyclopaedias of countries and regions
IV. History of encyclopaedias
    Encyclopaedias in the West
    Encyclopaedias in the East
    Aims and effects of encyclopaedias

I. The nature of encyclopaedias

In the *Speculum majus* ("The Greater Mirror"; completed 1244), one of the most important of all encyclopaedias, the French medieval scholar Vincent of Beauvais maintained not only that his work should be perused but that the ideas it recorded should be taken to heart and imitated. Alluding to a secondary sense of the word *speculum* ("mirror"), he implied that his book showed the world what it is and what it should become. This theme, that encyclopaedias can contribute significantly to the improvement of mankind, recurs constantly throughout their long history. A Catalan ecclesiastic and scholastic philosopher, Ramon Llull, regarded the 13th-century encyclopaedias, together with language and grammar, as instruments for the pursuit of truth. Domenico Bandini, an Italian humanist, planned his *Fons memorabilium universi* ("The Source of Noteworthy Facts of the Universe") at the beginning of the 15th century to provide accurate information on any subject to educated men who lacked books and to give edifying lessons to guide them in their lives. Francis Bacon believed that the intellect of the 17th-century individual could be refined by contact with the intellect of the ideal man. Another Englishman, the poet and critic Samuel Taylor Coleridge, was well aware of this point of view and said in his "Preliminary Treatise on Method" (1817) that in the *Encyclopaedia Metropolitana*, which he proposed to create, "our great objects are to exhibit the Arts and Sciences in their Philosophical harmony; to teach Philosophy in union with Morals; and to sustain Morality by Revealed Religion." He added that he intended to convey methodically "the pure and unsophisticated knowledge of the past . . . to aid the progress of the future." The Society for the Diffusion of Useful Knowledge declared in *The Penny Cyclopaedia* (1833–43) that, although most encyclopaedias attempted to form systems of knowledge, their own would in addition endeavour to "give such general views of all great branches of knowledge, as may help to the formation of just ideas on their extent and relative importance, and to point out the best sources of complete information."

In *De disciplinis* ("On the Disciplines"; 1531) the Spanish humanist Juan Luis Vives emphasized the encyclopaedia's role in the pursuit of truth. In Germany of the early 19th century the encyclopaedia was expected to provide the right or necessary knowledge for good society. Probably the boldest claim was that of Alexander Aitchison, that his new *Encyclopaedia Perthensis* (1796–1806) was intended to supersede the use of all other English books of reference.

Greek and Roman concepts

All these ideas were a far cry from the Greek concept, deriving from Plato, that in order to think better it is necessary to know all, and from the Roman attitude of the advisability of acquiring all useful knowledge in order to carry out one's tasks in life competently. The present concept of the encyclopaedia as an essential starting point from which one can embark on a voyage of discovery, or as a point of basic reference on which one can always rely, is little more than two centuries old.

The prose form has usually been accepted as the only suitable vehicle for the presentation of the text of an encyclopaedia, though the *Mappemonde*, or *L'Image du monde* ("The Image of the World"; 1246)—attributed by some to Gautier de Metz, a French poet and priest, and by others to a Flemish theologian, Gossouin—was written in French octosyllabic verse; and Jean de Magnon planned his never-completed *Science universelle* (1663) to comprise 200,000 lines of heroic verse. Even as late as 1783, the French poet André de Chénier was planning to condense Diderot's *Encyclopédie* into a poem in the Lucretian style. It has also been generally accepted that an encyclopaedia should adopt a straightforward, factual approach. Even so, the Latin writer Martianus Capella wrote his *Liber de nuptiis Mercurii et Philologiae* ("Book Concerning the Nuptials of Mercury and Philology"; *c.* 410–429) in an elaborate allegory in which the seven bridesmaids represented the seven liberal arts. Domenico Bandini similarly divided his encyclopaedia into five parts in honour of Christ's wounds, while Alfonso de la Torre, in his *Visiõ delectable* ("Delightful Vision"; *c.* 1435), adopted the allegorical approach of a child receiving instruction from a series of maidens named Grammar, Logic, Rhetoric, and so on.

The alphabetically arranged encyclopaedia has a history of less than 1,000 years, most of the encyclopaedias issued before the introduction of printing into Europe having been arranged in a methodical or classified form. The early compilers of encyclopaedias held with Samuel Taylor Coleridge that "To call a huge unconnected miscellany of the *omne scibile*, in an arrangement determined by the accident of initial letters, an encyclopaedia, is the impudent ignorance of your Presbyterian bookmakers!" Today several encyclopaedias still retain the classified form of arrangement.

There has never been any general agreement on the way in which the contents of an encyclopaedia should be arranged. In Roman times the approach was usually practical, with everyday topics such as astronomy and geography coming first, while the fine arts were relegated to the end of the work. The Roman statesman and writer

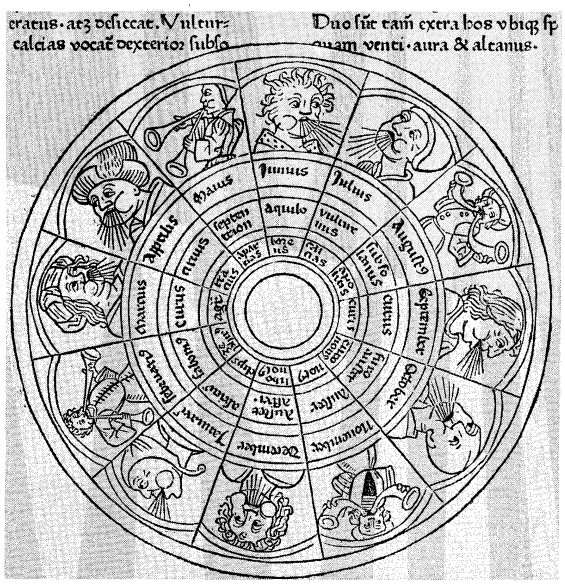


Illustration from the entry on the winds in Isidore of Seville's *Etymologiae*, an edition published in Strasbourg c. 1473.
By courtesy of the trustees of the British Museum; photograph, R.B. Fleming & Co.

Cassiodorus, however, in his 6th-century *Institutiones*, began with the Scriptures and the church and gave only brief attention to such subjects as arithmetic and geometry. St. Isidore of Seville, educated in the classical tradition, redressed the balance in the next century in his *Etymologiarum sive originum libri XX* ("Twenty Books on Origins, or Etymologies"), commonly called *Etymologiae*, giving pride of place to the liberal arts and medicine, the Bible and the church coming later, but still preceding such subjects as agriculture and warfare, shipping and furniture. The earliest recorded Arabic encyclopaedia, compiled by the Arab philologist and historian Ibn Qutayba, had a completely different approach, beginning with power, war, and nobility, and ending with food and women. A later Persian encyclopaedia, compiled in 975–997 by the Persian scholar and statesman al-Khwārizmī, started with jurisprudence and scholastic philosophy, the more practical matters of medicine, geometry, and mechanics being relegated to a second group labelled "foreign knowledge." The general trend in classification in the Middle Ages is exemplified by Vincent of Beauvais's *Speculum majus*, which was arranged in three sections: "Naturale"—God, the creation, man; "Doctrinale"—language, ethics, crafts, medicine; "Historiale"—world history. The encyclopaedists were, however, still uncertain of the logical sequence of subjects, and although there were many who started with theological matters, there were just as many who preferred to put practical topics first.

A turning point came with Francis Bacon's plan for his uncompleted *Instauratio magna* ("Great Instauration"; 1620) in which he eschewed the endless controversies in favour of a three-section structure, including "External Nature" (covering such topics as astronomy, meteorology, geography, and species of minerals, vegetables, and animals), "Man" (covering anatomy, physiology, structure and powers, and actions), and "Man's Action on Nature" (including medicine, chemistry, the visual arts, the senses, the emotions, the intellectual faculties, architecture, transport, printing, agriculture, navigation, arithmetic, and numerous other subjects).

Various arrangements of the contents of encyclopaedias

In his plan Bacon had achieved more than a thoroughly scientific and acceptable arrangement of the contents of an encyclopaedia; he had ensured that the encyclopaedists would have a comprehensive outline of the scope of human knowledge that would operate as a checklist and prevent their omitting whole fields of human thought and endeavour. Bacon had so profoundly altered the editorial policy of encyclopaedists that even 130 years later Diderot gratefully acknowledged his debt in the prospectus (1750) of *L'Encyclopédie*. Because every later encyclopaedia was influenced by Diderot's work, the guidance of Bacon still plays its part today.

Coleridge, who was very much impressed by Bacon's scheme, in 1817 drew up a rather different table of arrangement for the *Encyclopaedia Metropolitana*. It comprised five main classes: Pure Sciences—Formal (philology, logic, mathematics) and Real (metaphysics, morals, theology); Mixed and Applied Sciences—Mixed (mechanics, hydrostatics, pneumatics, optics, astronomy) and Applied (experimental philosophy, the fine arts, the useful arts, natural history, application of natural history); Biographical and Historical, chronologically arranged; Miscellaneous and Lexicographical, a gazetteer, and a philosophical and etymological lexicon. The fifth class was to be an analytical index.

Influence of Coleridge's classification

Although Coleridge's classification was altered by the publisher, and although the *Metropolitana* was an impressive failure, the ideas for it had a lasting influence. Even though nearly all encyclopaedias today are arranged alphabetically, the classifications of Bacon and Coleridge still enable editors to plan their work with regard to an assumed hierarchy of the various branches of human knowledge.

The concept of alphabetical order was well known to both the Greeks and Romans, but the latter made little use of it. Neither the Greeks nor the Romans employed if for encyclopaedia arrangement, with the exception of Sextus Pompeius Festus in his 2nd-century *De verborum significatu* ("On the Meaning of Words"). St. Isidore's encyclopaedia was classified, but it included an alphabetically arranged etymological dictionary. The 10th- or 11th-century encyclopaedic dictionary known as *Suidas* was the first such work to be completely arranged alphabetically, but it had no influence on succeeding encyclopaedias, although glossaries, when included, were so arranged. Bandini's *Fons memorabilium universi* ("The Source of Noteworthy Facts of the Universe"), though classified, used separate alphabetical orders for more

By courtesy of the trustees of the British Museum; photograph, R.B. Fleming & Co.

Engraved title page from the first edition of Francis Bacon's *Instauratio magna*, published in London, 1620.

than a quarter of its sections, and the Italian Domenico Nani Mirabelli's *Polyanthea nova* ("The New Polyanthea"; 1503) was arranged in one alphabetical sequence. These were rare exceptions, however; the real breakthrough came only with the considerable number of encyclopaedic Latin-language dictionaries that appeared in the early 16th century, the best known of which is a series of publications by the French printer Charles Estienne. The last of the great Latin-language encyclopaedias arranged in alphabetical order was *Encyclopaedia* (1630) by the German Protestant theologian and philosopher Johann Heinrich Alsted. The publication of *Le Grand Dictionnaire historique* ("The Great Historical Dictionary"; 1674) of Louis Moréri, a French Roman Catholic priest and scholar, confirmed public preference both for the vernacular and the alphabetically arranged encyclopaedia; this choice was emphasized by the success of the posthumous *Dictionnaire universel* (1690) by the French lexicographer Antoine Furetière.

From time to time important attempts have been made to reestablish the idea of the superiority of the classified encyclopaedia. Coleridge saw the encyclopaedia as a vehicle for enabling man to think methodically. He felt that his philosophical arrangement would "present the circle of knowledge in its harmony" and give a "unity of design and of elucidation." He did agree that his appended gazetteer and English dictionary would best be arranged alphabetically for ease of reference. By then, however, alphabetical arrangement had too strong a hold, and it was not until 1935 that a new major classified encyclopaedia began to appear—the *Encyclopédie française* ("French Encyclopaedia"), founded by Anatole de Monzie. The Dutch *Eerste nederlandse systematisch ingerichte encyclopaedie* ("First Dutch Systematic and Comprehensive Encyclopaedia"; 1946–52) has a classification that is in almost reverse order of that of the *Encyclopédie française*, but it is clear that behind both works lies a philosophical concept of the order and main divisions of knowledge that is influenced by both Bacon and Coleridge. The Spanish *Enciclopedia labor* (1955–60) and the *Oxford Junior Encyclopaedia* (1948–56) follow systems of arrangement that are closer to the French than to the Dutch example.

Role of the trivium and quadrivium

From the earliest times it had been held that the trivium (grammar, logic, rhetoric) and the quadrivium (geometry, arithmetic, astronomy, music) were essential ingredients in any encyclopaedia. Even as late as 1435 Alfonso de la Torre began his *Visiõ delectable* in almost that exact order, and only when he had laid these foundations did he proceed to the problems of science, philosophy, theology, law, and politics. Thus the seven liberal arts were regarded by the early encyclopaedists as the very mathematics of human knowledge, without a knowledge of which it would be foolish to proceed. This idea survived to a certain extent in Coleridge's classification; he stated that grammar and logic provide the rules of speech and reasoning, while mathematics opens mankind to truths that are applicable to external existence.

When Louis Shores became editor in chief of *Collier's Encyclopedia* in 1962, he said that he considered the encyclopaedia to be "one of the few generalizing influences in a world of over-specialization. It serves to recall that knowledge has unity." This echoes the view of the English novelist H.G. Wells, that the encyclopaedia should not be "a miscellany, but a concentration, a clarification and a synthesis." The Austrian sociologist Otto Neurath in the same year suggested that the proposed new international encyclopaedia of unified science should be constructed like an onion, the different layers enclosing the "heart"—comprising in this case the foundations of the unity of science.

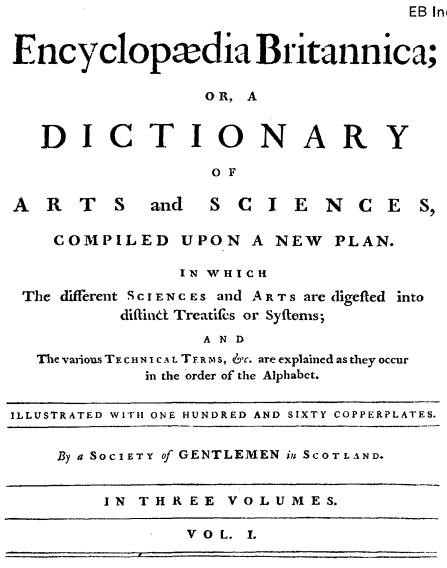
This modern emphasis on the unity of knowledge and the interdependence of its parts shows that encyclopaedists today are keenly aware of their educational role and that although the use of alphabetical arrangement may conceal the framework of classification, that framework is nevertheless an integral part of any good encyclopaedia (displayed in this encyclopaedia in the first volume as the *Outline of Contents*).

II. Encyclopaedias in general

THE ROLE OF ENCYCLOPAEDIAS

Of the various types of reference works—who's whos, dictionaries, atlases, gazetteers, directories, and so forth—the encyclopaedia is the only one that can be termed self-contained. Each of the rest conveys some information concerning every item it deals with, but only the encyclopaedia attempts to give a comprehensive summary of what is known of each topic considered. To this end it employs many features that can help in its task, including illustrations, maps, diagrams, charts, and statistical tables. It also frequently incorporates other types of reference works. Several modern encyclopaedias, from the time of Abraham Rees's *New Cyclopaedia* (1802–20) and the *Encyclopédie méthodique* ("Systematic Encyclopaedia"; 1782–1832) onward, have included a world atlas and a gazetteer, and language dictionaries have been an intermittent feature of encyclopaedias for most of their history.

Most modern encyclopaedias since the *Universal-Lexicon* (1732–50) of the Leipzig bookseller Johann Heinrich Zedler have included biographical material concerning living persons. The first edition of *Encyclopædia Britannica* (1768–1771) had no biographical material at all.



EB Inc.

Encyclopædia Britannica;

OR, A

DICTIONARY

OF

ARTS and SCIENCES,

COMPILED UPON A NEW PLAN.

IN WHICH

The different SCIENCES and ARTS are digested into distinct Treatises or Systems;

AND

The various TECHNICAL TERMS, &c. are explained as they occur in the order of the Alphabet.

ILLUSTRATED WITH ONE HUNDRED AND SIXTY COPPERPLATES.

By a SOCIETY of GENTLEMEN in SCOTLAND.

IN THREE VOLUMES.

VOL. I.

EDINBURGH:

Printed for A. BELL and C. MACFARQUHAR;

And fold by COLIN MACFARQUHAR, at his Printing-office, Nicolfon-ſtreet.

M.DCC.LXXI.

Title page of volume 1 of the first edition of *Encyclopædia Britannica*, published in Edinburgh, 1768–71.

Treatment of biographical material

In their treatment of this kind of information they differ, however, from the form of reference work that limits itself to the provision of salient facts without comment. Similarly, with dictionary material, some encyclopaedias—such as the great Spanish "Espasa" (1905 to date)—provide foreign-language equivalents as well.

An English lexicographer, Henry Watson Fowler, wrote in the preface to the first edition (1911) of *The Concise Oxford Dictionary of Current English* that a dictionary is concerned with the uses of words and phrases and with giving information about the things for which they stand only so far as current use of the words depends upon knowledge of those things. The emphasis in an encyclopaedia is much more on the nature of the things for which the words and phrases stand. Thus the encyclopaedic dictionary, whose history extends as far back as the 10th- or 11th-century *Suidas*, forms a convenient bridge between the dictionary and the encyclopaedia, in that it combines the essential features of both, embellishing them where necessary with pictures or diagrams, at the same time that it reduces most entries to a few lines that can provide a rapid but accurate introduction to the subject.

Interrelations. An encyclopaedia does not come into being by itself. Each encyclopaedia builds on the experience and contents of its predecessors. In many cases the debt is acknowledged: the German publisher Friedrich Arnold Brockhaus bought up the bankrupt encyclopaedia of Gotthelf Renatus Löbel in 1808 and converted it into his famous *Conversations-Lexikon;* but Jesuits adapted Antoine Furetière's *Dictionnaire universel* without acknowledgment in their *Dictionnaire de Trévoux* (1704). Classical writers made many references to their predecessors' efforts and often incorporated whole passages from other encyclopaedias. Of all the many examples, the *Cyclopaedia* (1728) of the English encyclopaedist Ephraim Chambers has been outstanding in its influence, for Diderot's and Rees's encyclopaedias would have been very different if Chambers had not demonstrated what a modern encyclopaedia could be. In turn, the publication of *Encyclopædia Britannica* was stimulated by the issue of *L'Encyclopédie.* Almost every subsequent move in encyclopaedia making is thus directly traceable to Chambers' pioneer work.

Readership. Encyclopaedia makers have usually envisaged the particular public they addressed. Cassiodorus wrote for the "instruction of simple and unpolished brothers"; the Roman statesman Cato wrote for the guidance of his son; Gregor Reisch, prior of the Carthusian monastery of Freiburg, addressed himself to "Ingenuous Youth"; the Franciscan encyclopaedist Bartholomaeus Anglicus wrote for *ordinary* people; the German professor Johann Christoph Wagenseil wrote for children; and Herrad of Landsberg, abbess of Hohenburg, wrote for her nuns. *Encyclopædia Britannica* is designed for the use of the curious and intelligent layman. The editor of *The Columbia Encyclopedia* in 1935 tried to provide a work compact enough and simply enough written to serve as a guide to the "young Abraham Lincoln." The Jesuit Michael Pexenfelder (1670) made his intended audience clear enough by writing his *Apparatus Eruditionis* ("Apparatus of Learning") in the form of a series of conversations between teacher and pupil. St. Isidore addressed himself to the needs not only of his former pupils in the episcopal school but also to all the priests and monks for whom he was responsible. At the same time, he tried to provide the newly converted population of Spain with a national culture that would enable it to hold its own in the Byzantine world.

Contributors. In sympathy with such ends, many scholars have contributed to encyclopaedias. Not all their contributions are known because, until recently, it was not the custom to sign articles. Even today there is what amounts to partial concealment in that articles are often initialled only, and, although a key is provided, few readers look up the writer's identity. It is known, however, that the English encyclopaedist John Harris enlisted the help of such scientists as John Ray and Sir Isaac Newton in his *Lexicon Technicum* (1704) and that Rees's *New Cyclopaedia* (1802–20) included articles on music by the English organist and music historian Charles Burney and on botany by the English botanist Sir J.E. Smith. Illustrious Frenchmen such as Voltaire, Rousseau, Condorcet, Montesquieu, and Georges Boulanger contributed to *L'Encyclopédie;* the writer and statesman Thomas Macaulay, the Russian-born jurist and medieval historian Sir Paul Vinogradoff, and the Czech statesman Tomáš Masaryk to the *Britannica;* the Scottish physicist Sir David Brewster and the Danish physicist Hans Christian Oersted to *The Edinburgh Encyclopaedia* (1808–30); the English astronomer Sir William Herschel and the English mathematician and mechanical genius Charles Babbage to the *Metropolitana;* the Russian Communist leader Lenin to the "Granat" encyclopaedia; and the dictator Benito Mussolini to the *Enciclopedia italiana.*

Language. The language of Western encyclopaedias was almost exclusively Latin up to the time of the first printed works. As with most scholarly writings, the use of Latin was advantageous because it made works available internationally on a wide scale and thus promoted unlimited sharing of information. On the other hand, it made the contents of encyclopaedias inaccessible to the great majority of people. Consequently, there was from the early days on a movement to translate the more important encyclopaedias into various vernaculars. Honorius Inclusus' *Imago mundi* ("Image of the World"; *c.* 1122) was rendered into French, Italian, and Spanish; Bartholomaeus Anglicus' *De proprietatibus rerum* ("On the Characteristics of Things"; 1220–40) into English; the Dominican friar Thomas de Cantimpré's *De natura rerum* ("On the Nature of Things"; *c.* 1228–44) into Flemish and German; and Vincent of Beauvais's *Speculum majus* ("The Greater Mirror") into French, Spanish, German, Dutch, and Catalan. In later years the more successful encyclopaedias were translated from one vernacular into another. Moréri's encyclopaedia, *Le Grand Dictionnaire historique,* was translated into both English and German. The German *Brockhaus* appeared in a Russian translation (1890–1907), and the French *Petit Larousse* had several foreign-language editions. Nevertheless, an encyclopaedia, however successful in its own country, may find acceptance in another country far from easy, because each nation appears to have its own very individual concept of what an encyclopaedia should comprise.

Translations of Latin encyclopaedias

The contemporary world. Encyclopaedias have often reflected fairly accurately the civilization in which they appeared; that this was deliberate is shown by the frequency with which the earlier compilers included such words as *speculum* ("mirror"), *imago* ("image"), and so forth in their titles. Thus as early as the 2nd century the Greek sophist Julius Pollux was already defining current technical terms in his *Onomastikon.* In the 13th century Vincent of Beauvais quoted the ideas of both pagan and Christian philosophers freely and without differentiation, for their statements often agreed on questions of morals. In doing so, he reflected the rapidly widening horizons of a period that saw the founding of so many universities. Bartholomaeus Anglicus devoted a considerable part of his work to psychology and medicine. "Theophilus" (thought to be Roger of Helmarshausen, a Benedictine monk) as early as the 12th century gave a clear and practical account in his *De diversis artibus* ("On Diverse Arts") of contemporary processes used in painting, glassmaking and decoration, metalworking, bone carving, and the working of precious stones, even listing the necessary tools and conditions for successful operations. Pierre Bayle, a French philosopher and critic, showed in his *Dictionnaire historique et critique* ("Historical and Critical Dictionary"; 1697) how the scientific renaissance of the previous 40 years had revolutionized contemporary thought. To every detail he applied a mercilessly scientific and inquiring mind that challenged the assumptions and blind reverence for authority that had characterized most of his predecessors.

Treatment of practical matters in encyclopaedias

At that point in history, much attention was being paid to practical matters: the statesman Jean-Baptiste Colbert himself directed the French Académie des Sciences (1675) to produce a work that eventually appeared as the *Description et perfection des arts et métiers* ("Description and Perfection of the Arts and Crafts"; 1761). The German Meyer's *Grosses Conversations-Lexicon* from the first edition (1840–55) onward paid particular attention to scientific and technical developments, and the *Encyclopedia Americana,* aided by the *Scientific American,* strengthened its coverage in this area from 1911 onward. From its first edition the *Encyclopædia Britannica* included practical articles on topics such as bookkeeping and aspects of farming. Similarly, Abraham Rees had been including articles on subjects such as candlemaking and coachbuilding earlier in the century. The outstanding example of a completely contemporary encyclopaedia was, of course, *L'Encyclopédie,* in which the philosopher Denis Diderot and the mathematician and philosopher Jean Le Rond d'Alembert and their friends set out to reject much of the heritage of the past in favour of the scientific discoveries and the more advanced thought of their own age. Their decision in this respect was both intellectually and commercially successful; since that time every edition of any good ency-

clopaedia has the additional merit of being a valuable source for the thought and attitudes of the world for which it was published.

Encyclopaedias and politics. All great encyclopaedia makers have tried to be truthful and to present a balanced picture of civilization as they knew it, although it is probable that no encyclopaedia is totally unbiassed. A great encyclopaedia is inevitably a sign of national maturity and, as such, will pay tribute to the ideals of its country and its times. The first Hungarian encyclopaedia, János Apáczai Csere's *Magyar encyclopaedia* (1653–55), was mostly a summary of what was available in foreign works, but the *Révai nagy lexikona* ("Révai's Great Lexicon"; 1911–35) was a handsome tribute to Hungary's emergence as a country in its own right, just as the *Enciklopedija Jugoslavije* (1955–) is a prestige work that does full justice to the advances made by Yugoslavia in the mid-20th century. The supreme example of an encyclopaedia that set out to present the best possible image of its people and the wealth and stature of their culture is undoubtedly the *Enciclopedia italiana* (1929–39). Mussolini's contribution of an article on Fascism indicates the extent to which the work might be regarded as an ideological tool, but, in fact, the bulk of its contents is admirably international and objective in approach. The various Soviet encyclopaedias already occupy many feet of shelf space, and the later editions each devote one complete volume to the Soviet Union in all its aspects. Though successive editions have been notable for the obvious political factors that have been responsible for the inclusion and exclusion of entries for famous nationals according to the current state of their acceptance or condemnation by the existing regime, many critics have felt that the newest edition, the first volume of which was issued in 1970, is more mature than any of the others in this regard.

Opposition to encyclopaedias

Political censorship is not the only form of oppression in the world of encyclopaedias. Antoine Furetière, on issuing his prospectus (1675) for his *Dictionnaire universel*, found his privilege to publish cancelled by the French government at the request of the Académie Française, which accused him of plagiarizing their own dictionary. The Leipzig book trade, fearing that publication of Johann Heinrich Zedler's huge *Universal-Lexikon* (1732–50) might put them out of business, made such difficulties that Zedler thought it best to issue his work in Halle. Diderot, the editor, and André-François Le Breton, the publisher, faced such opposition from both church and state in their publication of *L'Encyclopédie* (1751–65) that many of the volumes were secretly printed, and the last ten were issued with a false imprint. In the early part of the 19th century, *Brockhaus* was condemned by the Austrian censor, and in 1950 its 11th edition was branded as reactionary by the East German government.

The reader's needs. People look to encyclopaedias to give them an adequate introduction to a topic that interests them. Many expect the encyclopaedia to omit nothing and to include consideration of all controversial aspects of a subject. Encyclopaedia makers of the past assumed that there was a large public willing to read through an entire encyclopaedia if it was not too large. In the 18th century, for example, there was a good market for pocket-size compendia for the traveller, or for the courtier to browse in as he waited for an audience. Thus, although most encyclopaedias are multivolume works, there are many small works ranging from the *Didascalion* of the scholastic philosopher and mystic theologian Hugh of Saint-Victor (*c.* 1128), through Gregor Reisch's *Margarita philosophica* (1496) and the French writer Pons-Augustin Alletz' *Petite Encyclopédie* (1766), to C.T. Watkins' *Portable Cyclopaedia* (1817). The last was issued by a remarkable publisher, Sir Richard Phillips, who realized the great demand for pocket-size compendia and drove a thriving trade in issuing a number of these; he is thought to have written large sections of these himself.

Royalty and encyclopaedias. Most of the classic Chinese encyclopaedias owe their existence to the patronage of emperors. In the West, the Roman scholar Pliny dedicated his *Historia naturalis* ("Natural History") to the emperor Titus; Julius Pollux dedicated his *Onomastikon* to his former pupil, the Roman emperor Commodus, while the Byzantine philosopher and politician Michael Psellus dedicated his *De omnifaria doctrina* ("On All Sorts of Teaching") to his former pupil the emperor Michael VII Ducas, ruler of the Eastern Roman Empire. Gervase of Tilbury, an English ecclesiastic, compiled his *Otia imperialia* ("Imperial Pastimes") for the Holy Roman emperor Otto IV, and Alfonso de la Torre prepared his *Visiõ delectable* for Prince Carlos of Viana. St. Isidore dedicated his encyclopaedia to the Visigothic king Sisebut, and the French king Louis IX patronized Vincent of Beauvais's *Speculum majus*. Nor did kings eschew the work of compiling encyclopaedias. The emperor Constantine VII of the Eastern Roman Empire was responsible for a series of encyclopaedias, and Alfonso X of Spain organized the making of the *Grande e general estoria* ("Great and General History").

Contents and authority. The extent to which readers have been dependent on editorial decisions concerning not only what to include but also what to exclude has yet to be explored in detail. For example, Vincent of Beauvais rarely mentioned the pagan and Christian legends that were so popular in his day. The anonymous compiler of the scholarly *Compendium philosophiae* ("Compendium of Philosophy"; *c.* 1316) was careful to omit the credulous tales that appeared in contemporary bestiaries. For many centuries it was not considered right to include biographies of men and women who were still alive. And the early Romans, such as Cato the Censor, rejected much of Greek theoretical knowledge, regarding it as a dangerous foreign influence and believing with the Stoics that wisdom consisted in living according to nature's precepts.

Editorial censorship

Whatever the compiler did decide to include had a far-reaching influence. Pliny's vast *Historia naturalis* has survived intact because for so many centuries it symbolized human knowledge, and even the "old wives' tales" it injudiciously included were unquestioningly copied into many later encyclopaedias. The influence of St. Isidore's work can be traced in writings as late as Sir John Mandeville's travels (published in French between 1357 and 1371) and the English poet John Gower's 14th-century *Confessio amantis* ("A Lover's Confession"). Honorius' *Imago mundi* is known to have influenced some of the German medieval chronicles and the Norse saga of Olaf Tryggvason. The main source of classics such as the *Roman de la rose*, the Alexander romances, Archbishop Giovanni da Colonna's *Liber de viris illustribus* ("Book Concerning Illustrious Men"), and the recorded lives of the saints can be traced to the *Speculum majus*. The direct and indirect influence of the critical encyclopaedias of Pierre Bayle and Denis Diderot is, of course, incalculable.

The problem of keeping encyclopaedias up-to-date is very much of a preoccupation of contemporary publishers. The more important encyclopaedias maintain large permanent staffs to scrutinize newspapers and periodicals as well as books in an effort to ensure that the latest theories and discoveries are incorporated and that outmoded information and exploded "facts" give place to more precise and accurate material. The possibilities of utilizing the computer to this end have been studied intensively.

EDITING AND PUBLISHING

One-volume encyclopaedias

There always have been and there still are a number of successful one-volume encyclopaedias. Current outstanding examples include *The Columbia Encyclopedia*, the *Petit Larousse*, and *Hutchinson's New Twentieth Century Encyclopaedia*. Some booksellers and publishers nevertheless confirm that there is, however unreasonably, a certain amount of public prejudice against the single-volume form, and that most people prefer a multivolume work. Throughout the entire history of encyclopaedias there has been much disagreement on the correct number of volumes. Many of the Chinese encyclopaedias have been very much larger than any Western work. Pliny's

Historia naturalis comprised about 2,500 chapters, Zedler's *Universal-Lexicon* was planned for 12 volumes and eventually filled 64; the publishers of *L'Encyclopédie* were faced with a lawsuit (1768–78) for producing a 26-volume encyclopaedia instead of the ten volumes they had promised; and Johann Samuel Ersch and Johann Gottfried Gruber's German *Allgemeine Encyclopädie* ("General Encyclopaedia") had already reached 167 volumes at the time of its discontinuance. Today, although the major Soviet encyclopaedia consists of more than 50 volumes, most encyclopaedias appear to favour a size between 20 and 30 volumes, occupying between three and four feet of shelf space. Thus the modern encyclopaedia appears smaller than its 19th-century counterpart, but, in fact, the content may be greater because the thick matt paper of Victorian times has been replaced by a thinner paper capable of reproducing coloured and black and white halftone illustrations with sharp definition.

The length of encyclopaedia articles is a feature on which there has been a lasting difference of opinion. Most of the great encyclopaedias of the past, and many of the present, devote considerable space to any topic of major importance. As early as the turn of the 19th century, however, it was recognized that an alternative method of treatment would be to break up the larger subjects into a number of specific aspects and distribute them among the appropriate parts of the alphabetical sequence. Thus, an article on broadcasting could be limited to broad aspects, separate entries being made for radio, television, and so on. Supporters of this system argue that cross-references enable such articles to be linked up; opponents say that cross-references are not always followed up and that the fragmentation of the subject enforces much repetition of basic information. Nevertheless, *Brockhaus*, *Meyer*, *Larousse*, and many other encyclopaedias of the *Konversationslexikon* type exist today and have a strong following.

Authorship by one individual

The first encyclopaedia makers had no doubts concerning their ability to compile their works single-handedly. Cassiodorus, Honorius Inclusus (or Solitarius), and Vincent of Beauvais fully justified this attitude, though their task was largely that of the anthologist. Vincent and many other encyclopaedists employed both scribes and scholars to help them in their work, but once the encyclopaedia reached the stage of independent writing it was clear that the editorial task was going to become more complex. Even so, some of the later pocket encyclopaedias—such as the English bookseller John Dunton's mediocre *Ladies' Dictionary* (1694), *An Universal History of Arts and Sciences* (1745) by the French-born Englishman Chevalier Denis de Coëtlogon, and the popular *Allgemeines Lexicon* (1721) by the Prussian scholar Johann Theodor Jablonski—were substantially or almost wholly the work of a single author; such items are, however, negligible.

John Harris, an English theologian and scientist, may have been one of the first to enlist the aid of experts, such as the naturalist John Ray and Sir Isaac Newton, in compiling his *Lexicon Technicum* ("Technical Lexicon"; 1704). Johann Heinrich Zedler, in his *Universal-Lexicon* (1732–50), went further by enlisting the help of two general editors, supported by nine specialist editors, the result being a gigantic work of great accuracy. The French *Encyclopédie*, the largest encyclopaedia issued at that time, inevitably had many contributors, although the French writer Voltaire said that Diderot's collaborator, the Chevalier Louis de Jaucourt (aided by secretaries), contributed about three-quarters of the articles in that work. The pattern for future encyclopaedias was established: for any substantial work it would be necessary not only to have contributions from the experts of the day, but it would also be essential to have subject editors who could supervise the coverage and content in each area of knowledge. With little alteration, this system prevails today.

Encyclopaedia adjuncts. The readers of modern encyclopaedias are rarely aware of the numerous aids that have been provided to make their search for information so easy and efficient. Only when recourse is had to one of the older encyclopaedias does the reader become conscious of the advances that have been made. In former days it was often difficult to distinguish between one article and the next, because distinctive headings or inset titles or the use of boldface was rare. Nor was the necessity for running titles or alphabetical notations at the head of the pages fully appreciated. Even more troublesome was the problem of the arrangement of entries for several persons of the same name; reference to the older encyclopaedias under such headings as "Henry," "Charles," "John," or "Louis"—names held by both princes and religious potentates—will show how little the art of acceptable arrangement was understood.

Arrangement of articles

Cross-references and bibliographies. Cross-references are an essential feature of the modern encyclopaedia; they date back at least as far as Bandini's *Fons memorabilium universi*, but it was Brockhaus who introduced an ingenious system of using arrows instead of the words "see also." *The Columbia Encyclopaedia* achieves the same effect by printing in small capital letters the words under which additional information can be found. Other features of interest in the modern encyclopaedia include the devotion of each volume to a letter of the alphabet (*Compton's Encyclopaedia* formerly followed this system), or the indication of the division between letters by thumb-indexing or by the insertion of a thicker sheet of distinctively coloured paper. The bibliographies in established encyclopaedias for individual articles are usually the result of careful editorial consultation with the writer and with librarians.

Indexes. Undoubtedly the major adjunct of the modern encyclopaedia is its index. As early as 1614 the bishop of Petina, Antonio Zara, included an index of a kind in his *Anatomia ingeniorum et scientiarum* ("Anatomy of Talents and Sciences"). A Greek professor at Basel, Johann Jacob Hoffman, added an index to his *Lexicon universale* of 1677; *L'Encyclopédie* was completed by a two-volume "Table analytique et raisonnée" for the entire 33 volumes of text, supplements, and plates; and the *Britannica* included individual indexes to the lengthier articles in its 2nd edition (1778–84) and provided its first separate index volume for the 7th edition (1830–42). The nature of good indexing was still far from being fully understood, however, and it was only later in the 19th century that really good encyclopaedia indexes were prepared. In the 20-century encyclopaedias that provide indexes—and there are many that do not—the reader is invariably advised to read the guides to their use because the index is a sophisticated tool that, by the aid of a few simple typographical devices and editorial conventions, is able to offer a wealth of information in one alphabetical sequence.

Illustrative material. The use of illustrations in encyclopaedias goes back almost certainly to St. Isidore's time. One of the most beautiful examples of an illustrated encyclopaedia was the abbess Herrad's 12th-century *Hortus deliciarum*. In many earlier encyclopaedias the illustrations were often more decorative than useful, but from the end of the 17th century the better encyclopaedias began to include engraved plates of great accuracy and some of great beauty. *L'Encyclopédie* is particularly distinguished for its superb volume of plates—they have even been reprinted in the 20th century. In modern times the trend has been toward more lavish illustration of encyclopaedias, including elaborate coloured anatomical plates with superimposed layers, and specially inset small coloured halftones, as well as marginal line drawings. Since 1950 a form of encyclopaedia has begun to appear that comprises large numbers of coloured illustrations with a somewhat subordinated text.

Early illustrated encyclopaedias

The level of writing. The American editor Franklin H. Hooper, undaunted by his own lack of scholarship, took a notable part in ensuring that the articles of the 11th edition of *Encyclopædia Britannica* were kept within the mental range of the average man. The problem of the encyclopaedist has always been to strike the right mean between too learned and too simplified an approach. The Roman Cassiodorus wrote his encyclopaedia to pro-

vide a bridge between his unlettered monks and the scholarly books he had preserved for their use. Hugh of Saint-Victor, the theologian and philosopher, achieved one of the best approaches in his charming *Didascalion* (*c.* 1128), in which he used an elegant and simple style that everyone could appreciate. The abbess Herrad, knowing her audience, described in didactic fashion the history of the world (with emphasis on biblical stories) and its content, with commentaries and beautifully coloured miniatures designed to help and edify the simple nuns in her charge. The master of Dante, Brunetto Latini, wanted to reach the Italian cultured and mercantile classes with his *Li livres dou trésor* ("Treasure Books"; *c.* 1264) and therefore used a concise and accurate style that evoked an immediate and general welcome. Gregor Reisch managed to cover the whole university course of the day in his very pleasing and brief *Margarita philosophica*, which correctly interpreted the taste of the younger generation at the end of the 15th century.

Until the 17th century there had been a great many encyclopaedias written by clerics for clerics, and further examples continued to be published. After that time, more popular works began to be published as well, particularly in France, where such palatable compilations as the Sieur Saunier's *Encyclopédie des beaux esprits* ("Encyclopaedia of Great Minds"; 1657) had an immediate success. The philosopher Pierre Bayle in his *Dictionnaire historique et critique* (1697) first introduced the lay reader to the necessity of reading more critically; in this his work constituted a forerunner of *L'Encyclopédie*, with its challenges to so many undiscriminating assumptions about religion and politics, history and government. On the other hand, the contemporary *Dictionnaire universel* of the Jesuit fathers of Trévoux had a popularity among the orthodox that caused it to run through six editions and gradually to increase its size from three to eight volumes between 1704 and 1771.

Encyclopaedias for clerics and lay readers

Supplementary material. The idea of keeping encyclopaedias up-to-date by means of supplements, yearbooks, and so on, is well over 200 years old. In 1753 a two-volume supplement to the 7th edition of Ephraim Chambers' *Cyclopaedia* was compiled by George Lewis Scott, a tutor to the English royal family. Charles-Joseph Panckoucke, a publisher, issued a four-volume supplement to *L'Encyclopédie* (1776–77), in spite of Diderot's refusal to edit it. The *Britannica* included a 200-page appendix in the last volume of the 2nd edition (1784) and issued a two-volume supplement to the 3rd edition (1801; reprinted 1803). *Brockhaus* broke new ground by issuing in monthly parts (1857–64) a yearbook to the 10th edition (1851–55), which, on the commencement of the issue of the 11th edition, changed its name to *Unsere Zeit* ("Our Times") and doubled its frequency (1865–74). In 1907 Larousse began publication of the *Larousse mensuel illustré* ("Monthly Illustrated Larousse"). *The New International Encyclopaedia* issued a yearbook from 1908 (retrospective to 1903), and the *Britannica* issued one yearbook in 1913 and recommenced with the *Britannica Book of the Year* in 1938. The publication of supplements has a much longer history in China, but the system on which they operated was very different from that of the West (see below). Nowadays yearbooks are a common feature of encyclopaedias of standing in the U.S. and other countries. In the main, they are more effective in recording the events and discoveries of each year than they are in keeping the main articles up-to-date, but they perform an essential duty in informing their readers of much that is not reported or that is only inadequately reported in the press; at the same time, they provide a more reasoned assessment and perspective than the daily newspapers and the weekly commentaries can usually achieve.

Some of the leading encyclopaedias offer additional services that are not too widely known. The modern encyclopaedia is a complex work of reference, and the reader needs expert guidance if he is to get the best from its contents. To this end, small subject guides are sometimes issued, which, in narrative form, outline the whole field and bring each topic into perspective, drawing attention to the appropriate articles that will throw further light on the matter. As a means of self-education, some of these outlines maintain a high standard and are an invaluable adjunct to the encyclopaedia. Another feature offered by established encyclopaedias is an inquiry service. An inquiry on a point concerning which the encyclopaedia has not provided the reader with sufficient information gives him access to the files that the encyclopaedia maintains. In addition, the encyclopaedia will refer the more difficult inquiries to its specialist editors or contributors, or it may contact the appropriate specialized libraries, rather than return an inadequate answer to the reader. This particular service is of value to both the encyclopaedia's readers and to its staff, because, on the one hand, it gives the user the reassurance that he has an extra and very lively resource on which to draw for information, and on the other, it keeps the editorial staff in contact with the changing interests of readers and helps to underline omissions and deficiencies in the current edition.

Research services

Other supplementary material sometimes issued by encyclopaedias ranges from ten-year illustrated surveys of events to sets of books that have had a major impact on mankind. Although few publishers include dictionaries as an integral part of their encyclopaedia, they frequently supply a well-known, independently compiled work as part of their service. It is an increasingly common custom, however, for a modern encyclopaedia to incorporate an atlas and a gazetteer, often in the last volume.

Problems of encyclopaedias. *Authorship.* In using a reputable encyclopaedia the reader is inclined to accept the authenticity of any article he happens to read. Subconsciously he is aware that the highly organized body of scholars listed at the beginning of the work must inevitably have ensured the severe scrutiny of all material. Nevertheless, in recent years editors of encyclopaedias have tended more and more to commission signed articles by well-known experts. One of the most famous examples of this was the *Britannica*'s commissioning of articles for its 1922 supplement from some of the most famous men and women of the day: "Belgium" by the Belgian historian Henri Pirenne; "Anton Ivanovich Denikin" by the Russian-born jurist and historian Sir Paul Vinogradoff; "Drama" by St. John Ervine, the Irish playwright and novelist; "Czechoslovakia" by the Czech statesman Tomáš Masaryk; and "Russian Army" by Gen. Yuri Danilov. This created a new dimension in encyclopaedias, for it introduced a personal element on a scale previously seen only in the columns of *L'Encyclopédie*.

There is, in fact, a difference in the treatment of a subject written by a politician such as Masaryk and by an academic historian of distinction. Each writer has something important to offer, and the results will be very different.

Length restrictions. The restrictions imposed by the space available for any particular article are of great consequence. If an expert is asked to contribute an article on his own field, his first reaction is often one of dismay at the comparatively small amount of space that can be allotted to it. Writing articles for encyclopaedias is an art of its own; within a limited space so much must be compressed—nothing important can be omitted, nothing trivial should be included. Most experts would agree that it is easier to write a book than an encyclopaedia article. They would also consider it simpler to write an article for a periodical, because encyclopaedia writing is teamwork in which each article is edited in relation to others closely connected by subject. If a writer makes a statement that is partly qualified or totally contradicted in another article, the contributions of both writers must be scrutinized by the editorial staff, whose job it is to effect some kind of eventual agreement. Truth can be viewed from many standpoints, and references to any controversy may produce problems demanding all the skill and tact of the editors to resolve, particularly when the reputation of the writer is at stake in a signed article.

Revision and updating. The revision and updating of an encyclopaedia is one of the greatest challenges to its

makers, and one to which many ingenious, if admittedly partial, solutions have been found. Louis Moréri set the example in his rapid incorporation of new information in each succeeding issue of his widely used *Grand Dictionnaire historique* ("The Great Historical Dictionary"; 1674). When the German publisher Friedrich Arnold Brockhaus first issued his great encyclopaedia he was forced by an unexpectedly large public demand to issue edition after edition in quick succession (some of them even overlapped). In all of these he took great pride in providing the latest information, personally supervising much of the revision of individual articles. Moreover, he provided special supplements incorporating these revisions for purchasers of each edition, so that they did not have to buy the next in order to keep up-to-date. Modern encyclopaedias of standing use rather different systems, such as yearbooks, to achieve the same objective.

Revision of an article, at first glance, may appear to be a relatively easy task, including such processes as the insertion of death dates for recently deceased persons. A good editor realizes that this is not such a simple matter. On the death of a person of note the press takes the opportunity to reassess his contribution to society. Very often new facts come to light at that time—such as a long but concealed struggle against severe ill health—that materially affect what has been written. Similarly, a single scientific or archaeological discovery may alter to some extent all existing theories on the subject. In such cases the editor cannot achieve revision by the addition of a sentence at the end of an article; rather, he must consider partial or even complete rewriting. With the pace of modern research, this is an ever-present problem with which only a highly organized and scholarly editorial staff can adequately cope.

Conditions for complete revision

Up to about 1930 encyclopaedias achieved revision by issuing new editions every few years and by filling the gaps with supplements of yearbooks. This was a very unsatisfactory procedure that resulted in good editorial teams being largely dispersed when an edition had been issued, with the consequent difficulty of reassembling them when a new edition had been authorized. The insecurity of such a system caused the permanent loss of many very skilled personnel. A turning point came when the *Britannica* introduced the system of continuous revision that is now the practice of most major encyclopaedias in a number of countries. Continuous revision implies that each article is subject to review as new facts come to light and new developments take place. Apart from biographical and historical articles, the major revisions occur more often in scientific and technological articles than in those on the humanities. On the average, a good system of continuous revision results in the revision of about 10 percent of the text of an encyclopaedia in any one year, and probably every article will have been rewritten at least once in ten years. To do this, a permanent editorial staff is necessary. The increasing sales of encyclopaedias and growing public reliance on the accuracy of their contents has enabled publishers to assume this responsibility of continuous revision; an additional advantage has been gained in that a permanent staff gives much thought to further improvements in the course of its work.

The most apparent answer to problems of revision, the looseleaf encyclopaedia, has never proved a publishing success, in spite of the obvious advantages of being able to substitute new for outdated information. *Nelson's Perpetual Loose-Leaf Encyclopaedia* (2nd edition, 1920) was eventually discontinued, but the great *Encyclopédie française* (1935–66) has retained both the looseleaf form and the issue of revised volumes.

Controversy and bias. All major encyclopaedias throughout the ages have been accused of bias; *L'Encyclopédie* was full of such articles, and the early issues of the *Britannica* had many as well, but the views of the writers were well known, and their slanted opinions lent a colour to their words that many a modern encyclopaedia article lacks. Modern editors make every effort to eliminate any hint of bias in the writings of their contributors, but their task is difficult. For example, an account of the Korean War would vary according to whether it was written by a North or South Korean, a Chinese or an American writer.

Efforts to eliminate bias

Similarly, the inclusion of a map showing the frontiers between two or more nations may give rise to vigorous controversy if the nations involved were to dispute any part of the boundaries as shown. The illustration of a painting with an attribution to one artist may draw strong protests from art critics who do not agree with the writer. Controversy today has grown rapidly on many subjects that were not in dispute a few years ago.

Unauthorized copying. One of the problems that beset encyclopaedists is the unauthorized copying of material. Pirated editions of whole encyclopaedias were a normal feature of publishing up to the mid-19th century, but this rarely occurs nowadays, as a result of the international copyright convention, though in the 1950s unauthorized photo-offset versions of the *Britannica* were being printed in pirated editions in Taiwan. The illegal copying of individual articles in whole or in part still happens. Established encyclopaedias maintain a constant watch to avoid the inadvertent "borrowing" or plagiarizing of the contents of other encyclopaedias, and they take special care to avoid "adaptations" from foreign-language encyclopaedias. This is a very different attitude from that of the writer Temple Henry Croker, who, in his preface to *The Complete Dictionary of Arts and Sciences* (1764–66), impudently claimed: "We will venture to promise, that everything valuable in Mr. Chambers' Cyclopedia, and the other works of that Kind, particularly the Encyclopédie . . ., shall be contained in this Dictionary."

Modern encyclopaedias also take legal action in instances in which unscrupulous copying is openly undertaken. Knowledge is, of course, universal property, but its gathering and presentation in the form of an encyclopaedia is the property of its makers, in the same way that any other book is.

III. The kinds of encyclopaedias

GENERAL ENCYCLOPAEDIAS

Within the last 2,000 years, it is possible to perceive a pattern in encyclopaedia production that closely corresponds with the changing social needs of each age. The outstanding circumstances that governed the policy and production of encyclopaedias for the first 15 centuries were that comparatively few people were able to read and, stemming partly from this and partly from the cost of materials and workmanship, that copies of any lengthy work were very expensive. Only when printing was introduced into Europe did the cost of production drop by any large amount; this development in turn helped to stimulate the growth of readership. A notable feature at the time of the early printing press was the sudden growth in the popularity of some of the older encyclopaedias as a result of the tendency to ensure a ready market by printing works of which many manuscript copies were in circulation.

Influence of printing

During the first 16 centuries of their publication the great majority of encyclopaedias comprised great anthologies of the most significant writings on as many subjects as possible. These excerpts were arranged in an order that was constantly varying according to the individual compiler's concept of the hierarchies of human knowledge; some of these classification systems were more suitable than others, but none was completely successful in meeting the tastes of the reading public because there was no general agreement on the essential order of ideas. Although the compiler exercised considerable latitude in his choice of items to include in his encyclopaedia, he often restricted comment to a minimum, so that the reader was free to form his own opinion of what was offered. On the other hand, the informed guidance that the modern established encyclopaedia can offer was, in this way, less available to the reader. In addition, because the compiler selected his material from what had already been written, the gaze of the reader was being turned to the past, and, although he could enjoy the heritage of the preceding cultures, he was not being put in

touch with as much of the contemporary world as he might have desired.

Around the 10th or 11th century, a new type of encyclopaedia began to emerge, probably stimulated by the growing number of language dictionaries that, starting well before printing was used, grew ever more numerous once they could be produced by this method. Many of the early dictionaries were little more than enlarged glossaries, but from the time of *Suidas* onward, there began to appear a type of dictionary—now called encyclopaedic—that added to the definition and etymology of a word a description of the functions of the thing or idea it named. In some dictionaries, such as those of the Estiennes, a French family of bookdealers and printers, this description might in some cases be of considerable length. Thus the compilers of the new form of encyclopaedia that emerged in the 16th and 17th centuries inevitably thought in terms of arranging their entries in alphabetical order, because the dictionaries had already familiarized the reading public with this system. It needed only the success of a single encyclopaedia, such as Moréri's *Grand Dictionnaire historique*, to prove to publishers in general that this was a formula with considerable promise.

The last half of the 18th century was an era of such upheaval in man's concept of the world that the time was ripe for further experiments in the form of the encyclopaedia. The French encyclopaedists Diderot and d'Alembert and their band of contributors broke no new ground in the physical format and arrangement of the encyclopaedia, but their work inspired the intelligentsia of other nations to produce really good encyclopaedias of their own. It is no coincidence that both the German *Brockhaus* and the Scottish *Britannica* appeared with policies so different from all that had gone before that no publisher of any new encyclopaedia could afford to ignore their new pattern. Their formula was so good that the modern encyclopaedia is simply a vastly improved elaboration of their method of arrangement and organization. The compilers of both encyclopaedias had taken the best ideas from the anthologies and miscellanies of the early period of encyclopaedia making and from the later stage of encyclopaedic dictionaries. Realizing that the reading public would not tolerate the omission of some subjects and the unequal treatment of others, they prepared works in which at least a few lines were devoted to almost every conceivable topic, and for more important subjects a full account was provided, written by an expert, if possible.

Formulas of *Brockhaus* and *Britannica*

The three periods of the history of encyclopaedias—(1) to 1600, (2) 1601–1799, and (3) 1800 onward—are very unequal. They are, moreover, to a certain extent misleading, for the different forms of the encyclopaedia overlapped at each turning point for some years, and today there are still some important survivals from the two earlier periods. One can study and compare what each of the three main types of encyclopaedia has had to offer by opening the *Encyclopédie française*, *Webster's Third New International Dictionary*, and the *Encyclopædia Britannica* at the places in which each describes the same subject. The *Encyclopédie française* will provide one or more well-written treatises on the subject by writers of note. This is exactly what the encyclopaedias of the earliest period offered; and in both the old and the contemporary encyclopaedia the reader is left free to form his own opinion after reading what the experts have to say. *Webster's*, a one-volume work, of course provides much less, but it also gives much more, because it adds definitions and, often, explanatory drawings or diagrams to an admirably concise text that tells the reader much in a very few lines. This is exactly what the encyclopaedic dictionaries of Louis Moréri, Antoine Furetière, and others were offering in the 17th and 18th centuries. The *Britannica*'s contribution is distinct from those of the other two in that it provides a synthesis of what is known on the subject to date, and attempts to assess its current position. The comparison of these three examples gives only an approximate idea of the nature of the encyclopaedia in each of its three main stages, but it also helps to remind the reader of today that ideas concerning the form of an encyclopaedia have changed greatly throughout the centuries.

The encyclopaedias of the period before 1600 apparently were designed for a small group of people, who had much the same educational background as well as similar interests and opportunities to pursue them. In general, they had a common outlook on both religious and secular matters. Moreover, although they were citizens of many different countries, they were united by their knowledge and use of Latin, the international language.

Background of readers before 1600

The Eastern Roman emperor Constantine VII Porphyrogenitus (905–959) tried to plant firmly in the hearts of the most worthy of his contemporaries both knowledge and experience of the past. His were troubled times, and he felt justified in using much of his enforced leisure (he came to the throne in 911 but was not allowed to rule until 945) to provide for the administrators and emissaries of his court the most useful extracts from the writings of a very catholic selection of authors, including the patriarch of Constantinople John of Antioch (Johannes Scholasticus), the Roman historian Appian, the Greek historian Polybius, the Greek philosopher Socrates, the 5th-century Byzantine historian Zosimus, and many others. One of the unexpected by-products of this industry was the preservation of a large number of writings, a service that some of the other medieval encyclopaedias also performed.

An advantage of the encyclopaedists of the first period (*i.e.*, before 1600) was that each of them either knew or could visualize his reading public, a point that encouraged a minimum of commentary and moralizing. In a way, they were performing the duties of a personal librarian in that they drew their readers' attention to innumerable passages that they believed might be useful to them in their work or their private lives. The possibility of achieving even more was fully appreciated: the English scholar Alexander Neckham, in his early-13th-century *De naturis rerum* ("On the Natures of Things") hoped that by the imparting of knowledge he might help to lift or lighten man's spirit, and to this end he tried to maintain a simple and admirably clear text. Neckham's near-contemporary Bartholomaeus Anglicus similarly set himself in his *De proprietatibus rerum* ("On the Characteristics of Things") to bring to his readers' attention the nature and properties of the things and ideas on which the early Christian Fathers and the philosophers had expatiated, but he forbore to comment on their writings, leaving his readers to form their own judgments. The anonymous compiler of the *Compendium philosophiae* ("Compendium of Philosophy"; *c.* 1316) believed the knowledge of truth to be the supreme and final perfection of mankind; thus he never moralized on the contents of his encyclopaedia, its cumulative effect thereby being the more impressive.

Within the early period of the history of encyclopaedias a number of stages can be distinguished that make each group of works of great importance in any study of the development of scholarship and the spread of learning throughout the West. Encyclopaedias of classical times reached their culmination in Pliny's *Historia naturalis*, which was issued in the time of the Roman emperor Titus (AD 39–81). Not one of the encyclopaedias of Pliny or his predecessors paid much attention to religion; if it was discussed the approach was antiquarian, the gods of the different nations ruled by Rome being named and described in a dispassionate spirit that reflected both the tolerance and the noninvolvement of the Romans in these matters. The emphasis instead was on government, geography, zoology, medicine, history, and practical matters. The theories of the various philosophers were outlined impartially, no indication being given of any personal preference. This objective approach adopted by the Romans in their encyclopaedias was not achieved again until the 19th century.

Encyclopaedias of classical times

By the time of the Roman philosopher Boethius and the statesman Cassiodorus (*i.e.*, the 5th and 6th centuries AD), the position concerning objectivity had changed. Like Pliny and the Roman statesman Cato, Cassiodorus

had been an administrator, but while his predecessors had been engaged in interpreting and epitomizing the knowledge of the ancient world for the benefit of their own people, Cassiodorus realized the necessity for providing a new interpretation of this knowledge for the Goths, the new masters of Italy. In the next 700 years the impact of Christianity brought a new phase in Western encyclopaedia making, just as the impact of Islām is clearly visible in the Arabic encyclopaedias of the same period. Although religion is not always given pride of place in the encyclopaedias of those times, it pervades the whole of their contents. Thus Cassiodorus' division of his encyclopaedia into two main sections—divine and human—is made even more interesting by his inclusion of cosmography, the liberal arts, and medicine in the first section. Although the compilers of the encyclopaedias of this period could envisage in theory a perfectly logical arrangement for their encyclopaedias by starting with the creation and working downward to the smallest and least significant of God's creations, in practice they found this very difficult to apply, and the result was often only superficially scientific. Moreover, the inclusion of such topics as astrology and magic was surprisingly prevalent and only began to disappear after the publication of *Liber floridus* ("The Flowering Book"; *c.* 1120), by Lambert, a canon of Saint-Omer, a work that discarded practical matters in favour of metaphysical discussion.

Use of the vernacular

The third stage in the development of encyclopaedias came with the introduction of vernacular editions, such as the *Mappemonde* and *Li livres dou trésor*, and the reflection of the impact of Greek philosophical works (in translation) in the middle of the 13th century. In this era there was an increasing number of lay encyclopaedists—*e.g.*, Latini, Bandini, de la Torre—and the subject coverage changed to give more space and importance to the practical matters that interested the rising mercantile class. At the same time, theology no longer dominated the classification schemes. Humanism reached its full expression in the Spanish philosopher Juan Luis Vives' *De disciplinis* (1531), in which all the compiler's arguments were grounded on nature and made no appeal to religious authority. The printing press had eliminated one of the most vexatious problems: the introduction or perpetuation of textual errors by the manuscript copyists. At the same time, the wider circulation of encyclopaedias through the unrestricted sales of printed copies brought about a situation in which the compilers could no longer envisage their reading public and accordingly adjusted their approach to their largely unknown audience.

ENCYCLOPAEDIC DICTIONARIES

The period spanning the 17th and 18th centuries is characterized by the flourishing of the encyclopaedic dictionaries that were pioneered by the Estienne family in France in the 16th century. During these two centuries this form of encyclopaedia adopted two different policies. On the one hand there was the encyclopaedia, such as those of the Germans Johann Theodor Jablonski and Johann Heinrich Zedler, that paid particular attention to the fields of history and biography. On the other, there was a new form of encyclopaedia—if the exception of the 12th-century *De diversis artibus* be set aside—that devoted itself to the arts and sciences. The first type can therefore be said to be retrospective in approach, while the arts and sciences encyclopaedia was clearly identifiable with contemporary matters.

None of these divisions is actually clear-cut, for many traditional encyclopaedias continued to be compiled throughout the period, and not all the historical-biographical encyclopaedias ignored the arts and sciences or contemporary people and events. Nevertheless, the issue of Antoine Furetière's encyclopaedia and the immediate follow-up by *Le Dictionnaire des arts et des sciences* (1694) by the writer Thomas Corneille (the younger brother of the playwright Pierre Corneille) were sufficient to indicate the growing public interest in a more modern form of encyclopaedia. This indication was confirmed by the successful publication of John Harris' *Lexicon Technicum* (1704), which the author described as "an universal English dictionary of arts and sciences: explaining not only the terms of art, but the arts themselves." It is significant that Harris omitted such subjects as theology, biography, and geography. The Englishman Ephraim Chambers went even further in describing his internationally influential *Cyclopaedia* (1728) as

> an universal dictionary of arts and sciences; containing an explication of the terms, and an account of the things signified thereby, in the several arts, both liberal and mechanical, and the several sciences, human and divine, compiled from the best authors.

Ideal contents as viewed in the 18th century

No century has seen more public discussion of the nature of the encyclopaedia than the 18th; at the same time, there was much uncertainty concerning its ideal contents. The fine Italian encyclopaedia of Gianfrancesco Pivati (the secretary of the Academy of Sciences at Venice), the *Nuovo dizionario scientifico e curioso, sacro-profano* ("New Scientific and Curious, Sacred-Profane Dictionary"; 1746–51), avoided the subject of history, whereas the German writer Philipp Balthasar Sinold von Schütz's *Reales Staats- und Zeitungs-Lexicon* ("Lexicon of Government and News"; 1704) concentrated on geography, theology, politics, and contemporary history and had to be supplemented by the German economist Paul Jacob Marperger's *Curieuses Natur-, Kunst-, Berg-, Gewerk- und Handlungslexikon* (1712), which covered the sciences, art, and commerce.

The introduction of the arts and sciences type of encyclopaedia inevitably hastened the use of specialist contributors, for it widened the total subject field considerably. "Hübner" (as Sinold von Schütz's encyclopaedia was known from the writer of the preface) employed many contributors, and it is known from the draft prospectus of the British writer Oliver Goldsmith that an encyclopaedia he projected was to have included comprehensive specialist articles by the lexicographer Samuel Johnson, the statesman Edmund Burke, the portrait painter Sir Joshua Reynolds, the historian Edward Gibbon, the economist Adam Smith, and others. The remarkable progress made in this period can easily be judged when one compares the encyclopaedia *Lucubrationes* (1541), in which the author, Joachim Sterck van Ringelbergh, found it necessary to include a "miscellaneous" section (which he amusingly dubbed "Chaos"), with the approach of Johann Georg Krünitz, a German physician and philosopher, in his highly organized, modern *Oekonomisch-technologische Encyklopädie* ("Economic-Technological Encyclopaedia"; 1773–1858) with its 242 volumes.

THE MODERN ENCYCLOPAEDIA

The period of the encyclopaedic dictionary was brilliant, but it gradually became apparent that, in abandoning the systematic encyclopaedia of the earlier period in favour of the quick reference dictionary form, quite as much had been lost as had been gained. The comparatively brief entries in the encyclopaedic dictionary had, by accident of the alphabet, fragmented knowledge to such an extent that users received only a disjointed knowledge of the things in which they were interested. Nor had the willful and extremely individualistic effort of the French encyclopaedists Diderot and d'Alembert done more than confuse the issue, for they had bent the principles of encyclopaedia making to their own purposes. It is therefore remarkable that the initial solution to the problem should have been found by Andrew Bell (1726–1809), Colin Macfarquhar (*c.* 1745–93), and William Smellie (1740–95), three Scotsmen who were responsible for the first edition (1768–71) of *Encyclopædia Britannica*. Aware of the shortcomings of *L'Encyclopédie*, they devised a new plan. Their encyclopaedia was to include about 45 principal subjects (distinguished by titles printed across the whole page), supported by another 30 lengthy articles, the whole being contained within one alphabetical sequence interspersed with numerous brief entries enhanced by references, where appropriate, to the principal subjects. Some of the princi-

Plan of the first *Encyclopaedia Britannica*

pal articles, notably those on medical subjects, extended to well over 100 pages each. The three collaborators had thus incorporated the comprehensive treatment of important subjects accorded by the earliest form of encyclopaedias and had supplemented this with the attraction of the brief informative notices of minor topics that had been the chief feature of the encyclopaedic dictionary. The key to their success was, however, their retention of the single alphabetical sequence.

Meanwhile, Renatus Gotthelf Löbel was planning to compile an encyclopaedia that could supersede "Hübner." It was Sinold von Schütz who, in the 4th edition of "Hübner," had introduced the word *Conversations-Lexikon* into the title, and it was Löbel who decided to give it pride of place in his new encyclopaedia. The *Konversationslexikon* was designed to provide the rapidly growing German bourgeoisie with the background knowledge considered essential for entry into the polite society of the day. When Brockhaus took over Löbel's bankrupt and incomplete encyclopaedia, he saw the value and appeal of this evocative word and retained it (in various spellings) for many years afterward. Löbel's and Brockhaus' solution to the problem of the form of the modern encyclopaedia was not the same as the *Britannica*'s; it is interesting to note that whereas the *Britannica* model has widely prevailed throughout the English-speaking world, *Brockhaus* has been the model for most of the encyclopaedias prepared in countries in which English is not widely spoken.

Format of the *Konversationslexikon*

Brockhaus, throughout its existence, has faithfully followed a system in which the whole of knowledge has been analyzed into very specific topics. These topics are arranged alphabetically and, under each heading, condensed entries convey the essential information. By ingenious cross-references, entries are linked with other entries under which further information can be found, thus avoiding the inclusion of an index. There is no difficulty in distinguishing encyclopaedias of the *Konversationslexikon* form from encyclopaedic dictionaries. The former are usually of considerable size (*Der grosse Brockhaus*, 1928–35, included 200,000 articles by over 1,000 authors) and possess elaborate cross-reference schemes. Moreover, whenever a really important subject occurs, considerable space is allowed, though the same principle of concentrated text is followed. The *Ready Reference and Index* of this encyclopaedia reflects elements of the *Brockhaus* model but departs from it by offering indexing to longer articles of the circle of learning.

Although the *Britannica* and *Brockhaus* examples eventually became the models for 19th- and 20th-century encyclopaedias, there have been many survivals from the previous periods. Ersch and Gruber's enormous *Allgemeine Encyclopädie* ("General Encyclopaedia"; 1818–89) has been cited as a true example of the medieval "summa"—it is famed for including the longest article in any encyclopaedia, that on Greece, which fills 3,668 pages in volumes 80–87. The *Encyclopédie française* is an even later example of this form and, as Samuel Taylor Coleridge planned it, the *Encyclopaedia Metropolitana* could have proved the supreme example of this type of treatment. The encyclopaedic dictionary has never died; at the very time when *Brockhaus* and the *Britannica* were building their markets, Noah Webster was developing his dictionary's reputation for reliability into a household expression in the U.S.

ENCYCLOPAEDIAS FOR SPECIAL INTERESTS

Most encyclopaedias have been compiled from a purely scholarly point of view and have had no particular ax to grind, though nearly all have been inhibited to a certain extent by the interests and policies of the milieu in which they made their appearance. There are, however, several encyclopaedias that have deliberately been planned for a special purpose. One that is unique and continues to be of the greatest value to historians is the work of the 16th-century Spanish Franciscan Fray Bernardino de Sahagún, who spent much of his life in missionary work in Mexico. Sahagún was ordered to write in Nahuatl the information needed by his colleagues for the conversion of the Indians. The result, the *Historia general de las cosas de Nueva España* ("General History of the Matters of New Spain") is a magnificent record of the Aztec culture as recounted by the Indians of south central Mexico. The arrangement of this work, written in pictorial language as well as in Spanish, follows the familiar medieval pattern and resembles most closely that of Bartholomaeus Anglicus (Sahagún may have been familiar with a recent translation of Bartholomaeus' encyclopaedia). *Historia* is one of the most remarkable encyclopaedias ever compiled, and it has also a special importance for students of language.

Many of both the Arabic and Chinese classical encyclopaedias were compiled with the object of helping civil-service candidates in their studies and of providing administrators with the cultural background needed for their work. Their interest to historians of the two cultures can well be understood, for their arrangement and contents throw useful light on the concepts of administration and justice (to name only two aspects) in the Chinese and Islāmic worlds during the 7th to 15th centuries.

Older Arabic and Chinese encyclopaedias

Of the Western medieval encyclopaedias, the most interesting in this respect is the *De naturis rerum* (*c.* 1228–44) of the Dominican friar Thomas de Cantimpré. His aim was that of St. Augustine: to unite in a single volume the whole of human knowledge concerning the nature of things, particularly the nature of animals, with a view toward utilizing it as an introduction to theology. To achieve this he devoted the first part of his work to mankind, zoology, botany, and geology. He then compiled a much smaller section on cosmology. An additional section on astronomy followed later. Although it was a comparatively unorganized and ill-balanced work, the popular nature of this encyclopaedia, with its stories of magic and allegorical creatures, ensured its widespread success over two centuries.

Religion and politics were the main motives for writing encyclopaedias with a special purpose. Louis Moréri made no secret of his intention to produce an encyclopaedia that would defend the teaching and policies of the Roman Catholic Church. Antoine Furetière and Pierre Bayle, on the other hand, represented the philosophers, and their anticlerical bias was more in tune with the skeptical minds of the age. Nevertheless, there was still a strong orthodox following in France, as the long-continuing demand for the *Dictionnaire universel* of the Jesuit fathers of Trévoux demonstrated, and this encyclopaedia was as firmly in defense of Catholicism as *L'Encyclopédie* was critical of it.

Diderot and d'Alembert's encyclopaedia had originally been intended by its publisher to be no more than an adaptation of Ephraim Chambers' *Cyclopaedia*. The outcome was a giant reference work that criticized the government, satirized the Calvinist clergy of Geneva, championed the Age of Reason, and supported an atheistic materialism. To the more rigid members of the French Establishment, the encyclopaedia was a monster. But the more worldly had no objection to a work whose succeeding volumes were each an audacious source of scandal.

Even Pierre-Athanase Larousse, the French encyclopaedist, was not impartial. His finest encyclopaedia, the *Grand Dictionnaire universel du XIX^e^ siècle* ("Great Universal Dictionary of the 19th Century"; 1865–90), one of the most influential of the century, was deliberately anticlerical in policy. And Herder, in the heart of Catholic Germany, produced a counterweight to the Protestant *Brockhaus* in his *Konversations-Lexikon* (1853–57), which adopted a distinctive Catholic viewpoint. This excellent encyclopaedia was early recognized for its general impartiality, scholarship, and accuracy. In the long run, both "Herder" and *Brockhaus* gradually eliminated their sectarian inclinations, and the current editions are both of high standard.

Objectivity in encyclopaedias

CHILDREN'S ENCYCLOPAEDIAS

Before the 19th century, only Johann Wagenseil (1633–1705) produced an encyclopaedia for children—the *Pera*

Librorum Juvenilium ("Collection of Juvenile Books"; 1695). Larousse issued an interesting *Petite Encyclopédie du jeune âge* ("Small Children's Encyclopaedia") in 1853, but the next, *Encyclopédie Larousse des enfants* ("Larousse Encyclopaedia for Children"), did not appear until 1957. The first of the modern children's encyclopaedias was, however, a long-standing favourite. Prepared by the English writer and editor Arthur Mee (1875–1943), it was called *The Children's Encyclopaedia* (1910) in Great Britain and *The Book of Knowledge* (1912) in the U.S. The contents comprised a series of vividly written and profusely illustrated articles; because the system of article arrangement was obscure, much of the success of the work as a reference tool resulted from its splendidly contrived index, which even now remains a model of its kind. Arthur Mee later produced a completely pictorial encyclopaedia, *I See All* (1928–30), that comprised thousands of small illustrations, each accompanied by only a few words of text. Librarians still treasure it for its reference value, even if it is no longer used by children. In 1917–18 a completely new children's encyclopaedia was published, *The World Book Encyclopedia*, which the title page described as "organized knowledge in story and picture." A success from the start, it issued enlarged editions in quick succession. In 1925 a volume devoted to reading courses and study units was added, but the index was abandoned in favour of innumerable cross-references in the text. Annual supplements were also provided from 1922 onward. In 1961 a Braille edition in 145 volumes was issued; most of the illustrations were eliminated in this, but many of the diagrams and graphs were retained and in many cases the captions for the pictures were incorporated in the text. In 1964 a separate 30-volume set in a special large type was published for the use of the partially blind.

World War I put a halt to the idea of issuing a *Britannica Junior*, and the first edition was not published until 1934. It was based on *Weedon's Modern Encyclopedia*, whose copyright had been bought by Britannica. Renamed *Britannica Junior Encyclopædia* in 1963, it was specifically designed for children under 12. One of its features is its ready-reference index volume, which combines short fact entries with indexing to longer general articles. In 1960 a British *Children's Britannica* was issued, and was subsequently supplemented by yearbooks.

Compton's Encyclopedia

In 1894 Frank E. Compton sold a U.S. school encyclopaedia, the *Students Cyclopedia*, from door to door to pay his way through college. This later became the *New Students Reference Work*, which Compton finally bought. While continuing to publish this, Compton designed a completely new and, for those times, revolutionary work, which first appeared in 1922 as *Compton's Pictured Encyclopedia*. In due course, the system of continuous revision was introduced, close cooperation with educational and library advisers was fostered, and contributions from well-known authors were encouraged. In 1971 *Compton's*, by then published by Encyclopædia Britannica, Inc., introduced a new type of encyclopaedia for younger children just beginning to read, called *Compton's Young Children's Precyclopedia*, with 16 volumes, in which all the illustrations are coloured and supported by a brief but informative text.

Unlike *World Book*, *Compton's*, and the *Britannica Junior Encyclopædia*, the *Oxford Junior Encyclopaedia* (intended for children of 11 upward) is systematically arranged. Each of the 12 text volumes is devoted to a broad subject field: mankind, natural history, the universe, communications, great lives, farming and fisheries, industry and commerce, engineering, recreations, law and order, the home, and the arts. The 13th volume is an index, which includes a special section of ready-reference material. The contents of each volume are arranged alphabetically (with cross-references), and there are many illustrations.

SPECIALIZED ENCYCLOPAEDIAS

The alternative title of the 12th-century *Speculum universale* ("Universal Mirror") of a French preacher, Raoul Ardent (a follower of Gilbert de La Porrée, a French theologian), was the *Summa de vitiis et virtutibus* ("Summa [Exposition] of Faults and Virtues"). The title gives the clue of Raoul's intent, which was to provide a modern authoritative account of the Christian attitude to the world. His plan was different from that of other encyclopaedists, for he limited his work to the discussion (in this order) of theology, Christ and the redemption, the practical and ascetic life, thought, prayer, ethics, the four cardinal virtues, human conduct, and the four senses. This work could, in fact, be termed the first of the specialized encyclopaedias.

Apart from isolated examples such as this, and the technical encyclopaedia of Roger of Helmarshausen, the specialized encyclopaedia did not really make an appearance until the 18th century. The stimulus was probably provided by the increasing number of encyclopaedias that included the arts and sciences to such a point that some of them included little else. In any classified encyclopaedia the individual classes do, of course, constitute a kind of specialized encyclopaedia, but such a work falls far short of the real thing in that it is not sufficiently self-contained to stand on its own. It was inevitable, as the boundaries of knowledge contained in encyclopaedias grew greater, that there would be at least some attempts to produce specialized works of this kind; the French publishers Panckoucke and Agasse did something of this nature in their rearrangement of the contents of Diderot's *Encyclopédie* as a series of 196 volumes of separate dictionaries covering 41 subjects. The idea might have taken hold had it not been for insuperable difficulties caused by the French Revolution.

First specialized encyclopaedia on biography

The first real effort toward a specialized encyclopaedia had come some years earlier, and the subject field that it treated was biography. The *Allgemeines Gelehrten-Lexicon* ("General Scholarly Lexicon"; 1750–51) was compiled by Christian Gottlieb Jöcher, a German biographer, and issued by Gleditsch, the publisher of both "Hübner" and Marperger and the opponent of Zedler's encyclopaedia. Jöcher's work was continued by the German philologist Johann Cristoph Adelung and others and is still of value today. The field of international biography is not a simple one to tackle, and there were only two further efforts of note: J.C.F. Hoefer (1811–78) compiled the *Nouvelle Biographie générale* ("New General Biography"; 1852–66) and J.F. Michaud (1767–1839) was responsible for the *Biographie universelle*. These two great works were to a certain extent competitive; this helped to improve their coverage and content; they are still heavily used in all research libraries. After their publication, the task of recording biographical information on a universal scale reverted to the general encyclopaedias.

Developments in the field of specialized encyclopaedias correspond closely to other developments in the world of scholarship. It is, for example, no accident that so much attention should be paid to the subject of chemistry at a time when L.F.F. von Crell was issuing his series of abstract journals on chemistry. The English scientist and inventor William Nicholson (1753–1815) was first in the field with his *Dictionary of Chemistry* (1795), published by Sir Richard Phillips (who later issued C.T. Watkin's *Portable Cyclopaedia*). On this was based Andrew Ure's *Dictionary of Chemistry*, which was for a long time the standard reference work on the subject in Great Britain. In 1807 the German chemist Martin Heinrich Klaproth issued his *Chemisches Wörterbuch* ("Chemical Dictionary"), but a more important event was the publication of the *Handbuch der theoretischen Chemie* ("Handbook of Theoretical Chemistry"; 1817–19) by the German scientist Leopold Gmelin, a work of such excellence that it still appears in new editions from the Gmelin-Institut. Heinrich Rose, a German chemist, issued his *Ausführliches Handbuch der analytischen Chemie* ("Complete Handbook of Analytic Chemistry") in 1851, and the first edition of the famous Liebig, Poggendorff, and Wöhler's *Handwörterbuch der reinen und angewandten Chemie* ("Handbook of Pure and Applied Chemistry") was issued in 1837; its second edition (1856–65) was expanded to nine volumes. This work was continued by Hermann Fehling's *Neues Handwörterbuch der Chemie* ("New

Pocket Dictionary of Chemistry"; 1871–1930). The French counterpart, C.A. Wurtz's *Dictionnaire de chimie pure et appliquée* ("Dictionary of Pure and Applied Chemistry"; 1869–1908), became the standard work of its day. The Russian-born chemist Friedrich Konrad Beilstein first issued his *Handbuch der organischen Chemie* ("Handbook of Organic Chemistry") in Hamburg, Germany, in 1880–83; it is the most extensive work of its kind today. The French chemist Edmond Frémy's *Encyclopédie chimique* ("Chemical Encyclopaedia") appeared in 1882–99, and *A Dictionary of Applied Chemistry*, edited by Sir Thomas Edward Thorpe, the English chemist, was first issued in 1890–93. Twentieth-century standard works include Fritz Ullmann's *Enzyklopädie der technischen Chemie* ("Encyclopaedia of Applied Chemistry"; 1914–23), Victor Grignard's *Traité de chimie organique* ("Treatise on Organic Chemistry"; 1935), *Elsevier's Encyclopaedia of Organic Chemistry* (1940), the *Encyclopedia of Chemical Technology* (1947–56)—known by the names of its principal editors as "Kirk-Othmer"—Waldemar Koglin's *Kurzes Handbuch der Chemie* ("Short Handbook of Chemistry"; 1951), and the indispensable *Handbook of Chemistry and Physics*.

Specialized encyclopaedias of chemistry

The impressive run of encyclopaedias and handbooks of chemistry over so long a period is paralleled only in the field of music, in which the *Musikalisches Lexikon* ("Musical Lexicon"; 1732) of the German composer and music lexicographer Johann Gottfried Walther began the trend and was supplemented by the very successful *Historisch-biographisches Lexicon der Tonkünstler* ("Historical and Biographical Lexicon of Musicians"; 1790–92) of the German organist and music historian Ernst Ludwig Gerber. The *Biographie universelle des musiciens et bibliographie générale de la musique* ("Universal Biography of Musicians and General Bibliography of Music"; 1835–44) was compiled by the director of the Brussels Conservatoire, the Belgian composer François-Joseph Fétis, almost coinciding with the equally voluminous *Encyklopädie der gesammten musikalischen Wissenschaften* ("Encyclopaedia of Collected Musical Knowledge") of Gustav Schilling, a German lexicographer and historian of music. A pupil of Mendelssohn, Hermann Mendel, founded the *Musikalisches Conversations-Lexikon* (1870), which was completed by August Reissmann, who also edited the musicologist and composer Auguste Gathy's *Musikalisches Conversationslexikon* (1871). The great *Encyclopédie de la musique et dictionnaire du Conservatoire* (1913–31) was begun by the French writer on music Albert Lavignac and continued by Lionel de La Laurencie, but the third part, a dictionary of names and subjects covered in the preceding parts, was never issued. Walter Willson Cobbett compiled the *Cyclopedic Survey of Chamber Music* (1929–30), and the English writer on music Sir George Grove first issued his *Dictionary of Music and Musicians* in 1879–89. The German music historian Hugo Riemann compiled his standard *Musik-lexikon* in 1882, and the comprehensive *Musik in Geschichte und Gegenwart* ("Music of the Past and Present") began publication in 1949.

The publication of the German philosopher G.W.F. Hegel's *Encyklopädie der philosophischen Wissenschaften* ("Encyclopaedia of Philosophical Knowledge"; 1817) was of more than subject importance in that it was a compendium of the author's philosophical system in three parts; Logic, Nature, Mind. It influenced many editors of general encyclopaedias during the rest of the century. The standard work in this field has for many years been the *Dictionary of Philosophy and Psychology* edited by the American psychologist James Mark Baldwin, though the publication of *The Encyclopedia of Philosophy* (1967) provided a substantial work more in line with modern tastes. Other works in this area include the Centro di Studi Filosofici di Gallarate's *Enciclopedia filosofica* (1957), the French philosopher André Lalande's *Vocabulaire technique et critique de la philosophie* ("Technical and Critical Vocabulary of Philosophy", first issued 1902–12), and the Austrian writer Rudolph Eisler's *Wörterbuch der philosophischen Begriffe* ("Dictionary of Philosophical Concepts").

Encyclopaedias of philosophy

The Architectural Publication Society began issuing its *Dictionary of Architecture* as early as 1852, but it took 40 years to complete. A more modern work is *Wasmuths Lexikon der Baukunst* ("Wasmuth's Lexicon of Architecture"; 1929–37). Further material is included in the *Encyclopedia of World Art* (1959–68), the *Reallexikon für Antike und Christentum* ("Encyclopaedia for Antiquity and Christianity"; 1950–), and the *Enciclopedia dell'arte antica, classica e orientale* ("Encyclopaedia of Ancient, Classical, and Oriental Art"; 1958–66).

The words "Pauly-Wissowa" are very familiar to a great number of people. August von Pauly (1796–1845), the German classical philologist, began issuing his *Real-Encyclopädie der classischen Altertumswissenschaft* ("Encyclopaedia of Classical Antiquities") in 1839. The new edition was begun by another German classical philologist, George Wissowa, in 1893. This enormous work on classical studies has no equal in any part of the world, though it can be supplemented in some areas by the encyclopaedic series *Handbuch der Altertumswissenschaft* ("Handbook of Antiquities") begun in 1887.

The Swiss theologian J.J. Herzog (1805–82) gave religion its first great encyclopaedia with his *Real-Encyklopädie für protestantische Theologie und Kirche* ("Encyclopaedia of the Protestant Church and Theology"; 1854–68). Philip Schaff (1819–93), a Swiss-born U.S. church historian, prepared the abridged English edition (1882–84) from which *The New Schaff-Herzog Encyclopedia of Religious Knowledge* stems. James Hastings, a Scottish clergyman, was responsible for no fewer than four encyclopaedic works in this field: *A Dictionary of the Bible* (1898–1904); *A Dictionary of Christ and the Gospels* (1906–08); *Encyclopaedia of Religion and Ethics* (1908–26), still of great importance; and *Dictionary of the Apostolic Church* (1915–18). An even more significant series is the *Encyclopédie des sciences ecclésiastiques* ("Encyclopaedia of the Ecclesiastical Sciences"), which will take many decades to complete. It comprises: *Dictionnaire de la Bible* (1907–); *Dictionnaire de théologie catholique* (1909–); *Dictionnaire d'archéologie chrétienne et de liturgie* (1928–53); *Dictionnaire d'histoire et de géographie ecclésiastiques* (1929–); and *Dictionnaire de droit canonique* ("Dictionary of Canon Law"; 1935–). Other important works are *The Catholic Encyclopedia* (1907–18), which has not been completely superseded by the *New Catholic Encyclopedia* (1967); the finely illustrated *Enciclopedia cattolica* (1948–54); *Die Religion in Geschichte und Gegenwart* ("Religion in the Past and Present"; 1909–13); and the *Lexikon für Theologie und Kirche* ("Lexicon of Theology and the Church"; 1930–38). Two recently published encyclopaedias of religion include *The Encyclopaedia of Islam* (1960–) and the *Encyclopaedia Judaica* (1971–72).

It was not until the 1860s that three of the most useful handbooks now in daily use began to appear. *The Statesman's Year-Book*, important for its statistical and political information, began publication in 1864. In 1868 the English publisher Joseph Whitaker first issued his *Whitaker's Almanack*, and the *World Almanack* started in the same year. The *Chicago Daily News Almanac* appeared from 1885 to 1946, and the *Information Please Almanac* began in 1947. Herder's *Staatslexikon* ("Lexicon of Political Science") was first published in 1889–97; this compendium was soon followed by the *Dictionary of Political Economy* (1894) by the English banker and economist Sir Robert Palgrave. In 1930–35 the *Encyclopaedia of the Social Sciences* was published; an immediate success, it is often referred to as "Seligman" after the name of its chief editor. The new *International Encyclopedia of the Social Sciences* (1968) did not supersede it in every respect. In a similar fashion, the *Handwörterbuch der Sozialwissenschaften* ("Pocket Dictionary of the Social Sciences"; 1952–) supplemented rather than superseded the standard *Handwörterbuch der Staatswissenschaften* ("Pocket Dictionary of Political Science," 4th ed.; 1923–39).

Almanacs and handbooks of social and political science

In the field of literature, if Isaac Disraeli's *Curiosities of Literature* (1791) is ruled out, the first important handbook is the *Dictionary of Phrase and Fable* (1870) by the

English clergyman and schoolmaster Ebenezer Cobham Brewer (1810–97), supplemented with Brewer's *Reader's Handbook* (1879). Other important works include the *Dizionario letterario Bompiani degli autori* ("Bompiani's Literary Dictionary of Authors"; 1956–57), the *Dizionario letterario Bompiani delle opere* ("Bompiani's Literary Dictionary of Works"; 1947–50), *Cassell's Encyclopaedia of Literature* (1953), and the Oxford "companions" to American, English, and French literature. In the last quarter of the 19th century, three major specialized encyclopaedias were issued: *Dictionnaire de botanique* ("Dictionary of Botany"; 1876–92) of the French naturalist and physician Henri Baillon, the *Lexikon der gesamten Technik* ("Lexicon of Collected Technology"; 1894–99) of the German engineer Otto Lueger, and the Berlin Academy's *Enzyklopädie der mathematische Wissenschaften* ("Encyclopaedia of Mathematical Sciences"; 1898–1935). The last was shortly followed by the important but incomplete *Encyclopédie des sciences mathématiques pures et appliquées* ("Encyclopaedia of Theoretical and Applied Mathematical Sciences"; 1904–14).

Physics never received the degree of attention that the encyclopaedists accorded to chemistry and chemical engineering. The standard *Dictionary of Applied Physics* of the English physicist Sir Richard Glazebrook was first issued 1922–23. The *Handbuch der Physik* ("Handbook of Physics") was also issued during those years; the second edition (1955) is often referred to by the name of its editor, Siegfried Flügge. The most recent work is the *Encyclopaedic Dictionary of Physics* (1961–64), edited by James Thewlis.

In medicine the pioneer *British Encyclopaedia of Medical Practice* (1936–39) has been followed by *The Encyclopaedia of General Practice* (1963).

Other important encyclopaedias and handbooks issued in recent years include *The Encyclopedia of Photography* (1949), the superbly illustrated and well-documented *Enciclopedia dello spettacolo* ("Encyclopaedia of the Stage"; 1954–62), which includes all forms of staged entertainment; the *Dictionnaire du cinéma et de la télévision* ("Dictionary of the Cinema and Television"; 1965–); the *McGraw-Hill Encyclopedia of Science and Technology* (1960), and the *Encyclopedia of Library and Information Science* (1968–).

ENCYCLOPAEDIAS OF COUNTRIES AND REGIONS

A special kind of encyclopaedia dealing with a single country or region began to appear in the late 19th century. Sometimes it is possible to distinguish by a subtle form of titling those national encyclopaedias that deal with the world scene from those that concentrate chiefly on their own country. Thus the "Ruritanian Encyclopaedia" can usually be taken to be a work produced in Ruritania that takes a world view, while the "Encyclopaedia of Ruritania" probably deals mainly with Ruritania and the surrounding areas. There are no rules in this matter, and this method of differentiating between the two kinds can only be regarded as an imprecise way of identifying the kind of encyclopaedia in the first instance.

The encyclopaedias of geography are of particular use in this field because they cover in detail many islands, small cities, and other features that are only dealt with in the briefest fashion elsewhere. Of the modern geographical encyclopaedias the following are of especial importance: *Westermanns Lexikon der Geographie* (1968–); *Meyers Kontinente und Meere* ("Meyer's Continents and Seas"; 1968–); the Russian *Kratkaya geograficheskaya entsiklopedya* ("Short Geographical Encyclopaedia"; 1960–66); and the *Länderlexikon* ("Geographical Dictionary"; 1953–60), of which a new edition began to be issued in the early 1970s. These encyclopaedias have an additional value as sources of maps and illustrations that would be difficult to find elsewhere. The main modern encyclopaedias dealing with continents, regions, and countries are:

Encyclopaedias of continents, regions, and countries

Africa: *Encyclopaedia of Southern Africa* (1961); *Deutsches Kolonial-Lexikon* (1920); *Encyclopédie coloniale et maritime* (1941–); *Grande encyclopédie de la Belgique et du Congo* (1938–52).

Australasia: *The Australian Encyclopaedia* (1958); *Encyclopaedia of Australia* (1968); *The Modern Encyclopaedia of Australia and New Zealand* (1964); *An Encyclopaedia of New Zealand* (1966).

Central and South America: *Diccionario enciclopédico de las Américas* (1947); *Verbo: enciclopédia luso-brasileira de cultura* (1963–); *Diccionario enciclopédico del Peru* (1967); *Enciclopedia Yucatanense* (1944–47); *Enciclopedia de México* (1966–); *Diccionario histórico-enciclopédico de la Republica de El Salvador* (1927–); *Diccionario geográfico, estadístico, histórico, de la isla de Cuba* (1863–66); *Gran enciclopédia argentina* (1956–64); *Encyclopedie van de Nederlandse Antillen* (1969); *Encyclopaedie van Nederlandsch West-Indië* (1914–17); *Enciclopédia larense* (1941).

Europe: *Flandria nostra* (1957–60); *Magyar életrajzi lexikon* (1967–69); *Encyclopaedia of Ireland* (1968); *Latvijas PSR mazā enciklopēdija* (1967–); *Latvju enciklopēdija* (1950–55); *Mažoji lietuviškoji tarybinė enciklopedija* (1966); *Norge* (1963); *Enciclopedia României* (1938–43); *Encyclopédie polonaise* (1916–20); *Sverige: land och folk* (1966); *Ukraine: A Concise Encyclopaedia* (1963–); *Narodna enciklopedija srpsko-hrvatsko-slovenačka* (1925–29).

Middle East: *Eretz-Yisra'el: Entziqlopedia Ṭopografit-Hisṭorit* (1946–55).

North America: *Encyclopedia Canadiana* (1957–58).

IV. History of encyclopaedias

ENCYCLOPAEDIAS IN THE WEST

Early development. The first fragments of an encyclopaedia to have survived are the work of Speusippus (died 339/338 BC), a nephew of Plato's. Speusippus conveyed his uncle's ideas in a series of writings on natural history, mathematics, philosophy, and so forth. Aristotle's wide-ranging lectures at the Lyceum were equally influential, and he and Plato appear to have been the originators of the encyclopaedia as a means of providing a comprehensive cultural background.

The Greek approach was to record the spoken word. The Romans, on the other hand, aimed to epitomize existing knowledge in readable form. Their first known effort is the *Praecepta ad filium* ("Advice to His Son"; *c.* 183 BC), a series of letters (now lost) written by the Roman consul Cato the Censor to his son. Cato's intention was to provide a summary of useful information that could help in the process of living and in guiding and helping one's fellow men. A more substantial attempt was made by the learned Latin writer Marcus Terentius Varro in his *Disciplinarum libri IX* ("Nine Books of Disciplines"), his *Rerum divinarum et humanarum antiquitates* ("The Antiquities of Things Divine and Human"), and his *Imagines*, which together covered the liberal arts, human efforts, the gods, and biographies of the Greeks and Romans.

Pliny's *Historia naturalis*

The most important Roman contribution was the *Historia naturalis* of Pliny the Elder, a vast work constituting a kind of classified anthology of information. Although undiscriminating in its record of fact and fancy, it was nevertheless very influential; the Latin grammarian and writer Gaius Julius Solinus drew nearly 90 percent of his 3rd-century *Collectanea rerum memorabilium* ("Collection of Memorabilia") from Pliny, and the *Historia naturalis* served as a major source for other encyclopaedias for at least the next 1,500 years. Even today it is still an important record for details of Roman sculpture and painting.

The statesman Cassiodorus, when he withdrew to the Vivarium in 551, dedicated this monastery to sacred and classical learning. His *Institutiones divinarum et saecularium litterarum* ("Institutes of Divine and Secular Literature") seems to have been designed to preserve knowledge in times that were largely inimical to it. In his encyclopaedia, Cassiodorus drew a clear distinction between the sacred and the profane, but the first Christian encyclopaedia to be compiled for the benefit of the newly converted Spanish population followed a different

scheme. St. Isidore (*c.* 560–636) considered the liberal arts and secular learning to be the true basis of a Christian's education. His *Etymologiae* therefore paid much attention to practical matters and even included an etymological dictionary. This was in line with the thought of St. Jerome—on whose encyclopaedic *Chronicon* and *De viris illustribus* St. Isidore had drawn—who, in common with the early Christian Fathers, was eager to provide a basis for a Christian interpretation and organization of knowledge. This concept was much later to be renewed by the Catalan ecclesiastic Ramon Llull.

The development of the encyclopaedia during the next 500 years, though of social interest, was undistinguished from the point of view of scholarship. Rabanus Maurus (*c.* 776–856), one of the English scholar Alcuin's favourite pupils, compiled *De universo* ("On the Universe"), which, in spite of its being an unintelligent plagiarism of St. Isidore's work, had a lasting popularity and influence throughout the medieval period. A series of encyclopaedias of special subjects—undistinguished anthologies of classical and Christian writings on history, jurisprudence, agriculture, medicine, veterinary surgery, and zoology—was organized by the Byzantine emperor Constantine VII Porphyrogenitus (905–959). Michael Psellus (1018–96), a tutor of a later emperor, contributed a more interesting work, *De omnifaria doctrina*, in the form of questions and answers on both the humanities and science. At this time there was a growing influence on metropolitan and secular learning. In an attempt to counterbalance it, the brief but charming *Didascalion* of Hugh of Saint-Victor (*c.* 1096–1141), which paid much attention to practical matters as well as to the liberal arts, was soundly based on a profound classification of knowledge that influenced many later encyclopaedias. About this time an encyclopaedic dictionary known as *Suidas* broke with tradition by adopting alphabetical order for its contents. This had no effect on the plan of later encyclopaedias, but its contents included so much useful information that it has retained its importance as a source throughout the succeeding centuries.

Adoption of the alphabetical order in *Suidas*

The *Liber floridus* (*c.* 1120) of Lambert of Saint-Omer is an unoriginal miscellany, but it has an interest of its own in that it discards practical matters in favour of metaphysical discussion and pays special attention to such subjects as magic and astrology. Lambert had been influenced by the Pseudo-Dionysian writings, one aspect of the contemporary revival of learning that was to reach its climax in the next century. The greatest achievement of the 12th century was the *Imago mundi* of Honorius Inclusus. Honorius produced his "mirror of the world" for Christian, later abbot of St. Jacob, and drew on a far wider range of authorities than any of his predecessors. The arrangement of the first section on geography, astrology, and astronomy was sound; it started with the creation and worked down to individual countries and cities. This was followed by a "chronicle," and a third section provided a brief list of important events since the fall of Satan. Honorius accurately foresaw his book's fate: innumerable copies, unauthorized plagiarisms, incessant criticism, and incompetent additions for at least 200 years.

Probably the first encyclopaedia to be compiled by a woman, the *Hortus deliciarum* of the abbess Herrad (died 1195), comprised a magnificent illuminated manuscript with 636 miniatures, intended to help and edify the nuns in her charge. Bartholomaeus Anglicus based his *De proprietatibus rerum* (1220–40) on the works of St. Isidore and Pliny. It was designed for ordinary people and became Europe's most popular encyclopaedia for the next three centuries. But the outstanding achievement of the Middle Ages was the *Speculum majus* of Vincent of Beauvais. Vincent was not an original writer but he was industrious, and his work comprised nearly 10,000 chapters in 80 books; no encyclopaedia rivalled it in size until the middle of the 18th century. The work was very well balanced, almost equal space being allotted to the three sections. The "Naturale" dealt with God and man, the creation, and natural history. For this Vincent drew not only on Latin writings but also on Greek, Arabic, and Hebrew, which were at that time (through translations) making a very considerable impact on the thinking of the West. The "Doctrinale" covered practical matters as well as the scholastic heritage of the age. The "Historiale" included a summary of the first two sections and a history of the world from the creation to the times of St. Louis. A fourth section, "Morale," based principally on St. Thomas Aquinas, was added after Vincent's death. The influence of the *Speculum majus* was immediate and lasting. Translations were made into several languages, and complete reprints appeared as late as 1863–79. One of its many values is that it is a source for extracts from many documents of which no other parts have survived. Another is its detailed history of the second quarter of the 13th century.

Vincent's was the last major work of its kind. Later encyclopaedists began to compile for a wider public than the very limited world of religious communities. The first breakaway from Latin came with *Li livres dou trésor* ("Treasure Books") of Brunetto Latini (*c.* 1220–95), the master of Dante, and the Florentine poet and philosopher Guido Cavalcanti. Latini wanted to reach the mercantile and cultured classes of Italy; he therefore used French, their common language. The arrangement of his work was similar to Vincent's but his approach was concise. The language, the brevity, and the accuracy of his encyclopaedia had an immediate and wide appeal. A friend of Petrarch's, Pierre Bersuire, based his *Reductorium, repertorium et dictionarium morale utriusque testamenti* ("Moral Abridgment, Catalogue and Dictionary of Each Testament"; *c.* 1340) on Bartholomaeus' *De proprietatibus rerum*. In contrast to Latini's work, this was a return to the traditional, with its moralizings on the Bible, Ovid's *Metamorphoses*, and natural history, but it had a considerable success when printing was introduced, being issued 12 times by 1526.

Brunetto Latini's breakaway from Latin

One of the most delightful of all encyclopaedias is the little *Margarita philosophica* that Gregor Reisch (died 1525) wrote for young people. In some 200 pages he contrived to cover in a very pleasing style the whole university course of the day, both the trivium and the quadrivium. The arrival of humanism is reflected in the *De disciplinis* of Juan Luis Vives, a pioneer in psychology and philosophical method; Vives grounded all his arguments on nature and made no appeal to religious authority. With the writing of the anonymous *Compendium philosophiae* (*c.* 1300) the concept of the modern scientific encyclopaedia was reached at last. It was the first encyclopaedia to adopt an inquiring and impartial attitude to the things described, and the old wives' tales that had filled so many pages of encyclopaedias from the time of Pliny onward were replaced by the latest scientific discoveries.

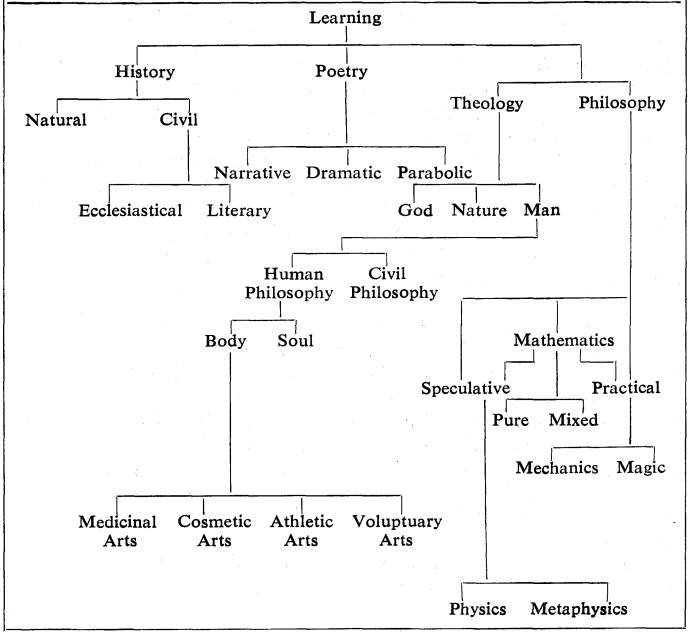
The first indigenous French encyclopaedia, the popular *Dictionarium historicum, geographicum et poeticum* ("Historical, Geographical, and Poetic Dictionary") of Charles Estienne (1504–64) was not published until 1553. For encyclopaedias in their own language, the French still had to rely on translations of the encyclopaedias of other nations, such as *Les diverses leçons* ("The Various Lessons"; 1552) of Pedro Mexia, a mediocre Spanish historian whose haphazard compilation was enormously popular in the 16th and 17th centuries.

The development of the modern encyclopaedia—17th and 18th centuries. Francis Bacon's purpose in writing the *Instauratio magna* was "to commence a total reconstruction of sciences, arts, and all human knowledge, raised upon the proper foundations" in order to restore or cultivate a just and legitimate familiarity between things and the mind. Only a small part of this enormous work was ever completed, but the author had planned 130 sections divided into three main sections: external nature, man, and man's action on nature. From its proposed contents Bacon's intention was clearly to compile an encyclopaedia thoroughly scientific in character—"a thing infinite and beyond the powers of man"—that he himself recognized to be revolutionary in character. His most important contribution was, however, the devising of a new and thoroughly sound classification of knowledge that bears a remarkable resemblance to the classifi-

Bacon's influential classification

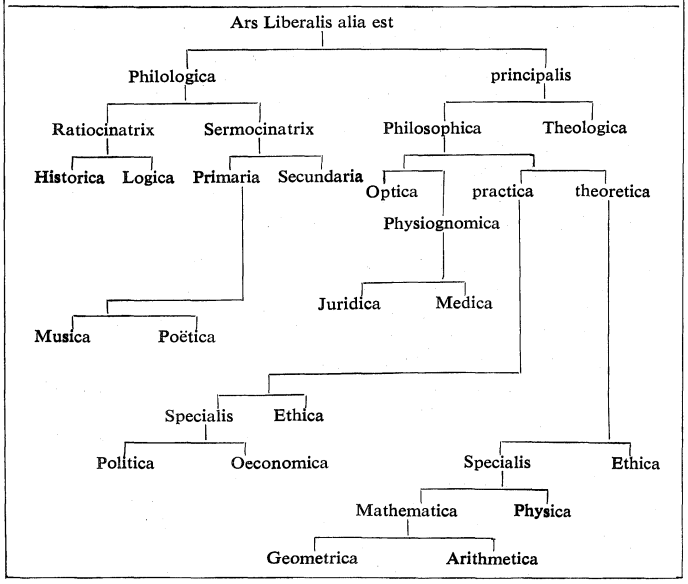
cation put forward by Matthias Martini in his *Idea Methodica* (1606). Although Bacon was apparently unaware of this work, both philosophers were probably working from the same basic Platonic precepts. The results were profound: Diderot made a point of acknowledging the assistance Bacon's analysis of the structure of human knowledge had afforded him in planning the contents of *L'Encyclopédie*, and Samuel Taylor Coleridge hailed "the coinciding precepts of the Athenian Verulam and the British Plato."

Table 1: Francis Bacon's Classification of Knowledge



Only two more Latin encyclopaedias of any importance followed. Antonio Zara, bishop of Petina, compiled the *Anatomia Ingeniorum et Scientiarum* ("Anatomy of Arts and Sciences"; 1614), which was chiefly remarkable for the inclusion of an index. And Johann Heinrich Alsted, who, like Martini, came from Herborn, compiled an *Encyclopaedia* (1630) whose arrangement corresponds broadly to the main outlines of Matthias' classification of human knowledge.

Table 2: Matthias Martini's Classification of Human Knowledge



Zara's and Alsted's encyclopaedias were systematically arranged. The turning point came with Louis Moréri's alphabetically arranged *Grand Dictionnaire historique* (1674), which was especially strong in geographical and biographical material. Its success was immediate; six editions were issued by 1691, each incorporating much new contemporary information. English editions followed in 1694, 1701, and (a supplement) 1705. Other encyclopaedias in England, Germany, Switzerland, and The Netherlands acknowledged its inspiration. The alphabetically arranged encyclopaedia in the vernacular had almost won the day, in spite of the German scholar Daniel George Morhof's modest success with his ill-balanced *Polyhistor Literarius, Philosophicus, et Practicus* ("Literary, Philosophical, and Practical History"; 1688–1708).

Furetière's modern *Dictionnaire universel*

If there was any doubt concerning the more popular form of the encyclopaedia, the issue of Antoine Furetière's *Dictionnaire universel des arts et sciences* (1690) confirmed the true nature of public taste. Furetière not only compiled a fine encyclopaedic dictionary, but he emphasized the arts and the sciences, thus reflecting the rapidly growing public interest in modern culture, science, and technology. If confirmation were still needed, the Académie Française's commissioning of Thomas Corneille to compile *Le Dictionnaire des arts et des sciences* (1694), with its thorough and authoritative treatment of these new encyclopaedic features, demonstrated that even the more conservative scholars were by now keenly aware that a new spirit had arisen. The period of the clerical encyclopaedia had ended, as the Franciscan friar Vincenzo Maria Coronelli found when his *Biblioteca Universale Sacro-Profano* (1701–06) ceased publication at volume 7 of a projected 45.

Pierre Bayle, in his *Dictionnaire historique et critique* (1697), achieved a most remarkable tour de force. Although his encyclopaedia purported to be an updating of the information in Moréri, the entries were largely unexceptionable. The real originality of his work lies in the profuse and scholarly footnotes and the commentaries that at times were an amazing mixture of skepticism, blasphemy, and ribaldry. Bayle challenged orthodox ideas; his brilliant mind spared nothing. This approach heralded that of Diderot, and the distinguished writers who revised later editions—Prosper Marchand and Pierre Desmaizeaux—continued in the same style.

The *Lexicon Technicum* (1704) of John Harris represented the powerful impact of the work of the Royal Society (founded 1660). Here was all the equipment of the modern encyclopaedia: excellent engraved plates, clear practical text, bibliographies appended to the more important articles. So far, England had had to make do with translations of French encyclopaedias. Harris' emphasis on the need to include scientific and technical subjects helped to reverse the trend. This process was completed by the issue of Ephraim Chambers' *Cyclopaedia* (1728). Like Harris, Chambers omitted people in favour of more information on the arts and sciences, and he paid more attention to clear expositions of ancient and modern philosophical systems. His admirably cross-referenced work is universally recognized as the father of the modern encyclopaedia.

The French were well aware of these developments. By 1744 five editions of Chambers' *Cyclopaedia* had been issued. The Paris publisher André Le Breton saw a ready market for a translation. The first proposals were a failure, however, and Diderot was enlisted to plan what at that time was still essentially a translation on a much broader basis. Under the hands of Diderot and d'Alembert the concept changed. *L'Encyclopédie* (1751–65) was a philosophical undertaking carried out on a gigantic scale, and much of the writing was of a high standard. To the orthodox, it appeared that the project had got out of hand, but there were 2,000 subscribers to the first volume, and the subsequent scandals over the irreverent, authority-challenging articles only added to the number of purchasers. The equivocal attitude of high dignitaries in both church and court and the growing public dislike of the encyclopaedia's chief critics—the Jesuits—led to a complex situation in which official disapproval and sub-

Diderot's monumental *Encyclopédie*

stantial private encouragement caused the production and fortunes of *L'Encyclopédie* and its producers to lurch dangerously from one crisis to another. Curiously, Diderot did nothing to further the physical development of the encyclopaedia; his contribution was to fire men's minds with a willful guidance that conformed to the country's increasingly revolutionary spirit. As Voltaire said: "this vast and immortal work seems to reproach mankind's brief lifespan."

The shortcomings of *L'Encyclopédie* were obvious. The essential ingredients of an encyclopaedia, the entries on every conceivable subject, had been sacrificed to make place for lengthy polemics on the controversial topics of the day. The *Encyclopædia Britannica* was intended to improve on this, and, with all its shortcomings, the first edition (1768–71) did exactly that. The achievement of its editors was the more remarkable in that there were already several English encyclopaedias on the market. The Scottish encyclopaedia, however, reflected the taste of the day better than any of its competitors, for it was a completely new work and not just a remaking of Chambers and Harris. There was much to criticize in the first edition, but the second (1777–84; dated 1778–83) was greatly improved, and succeeding issues of the *Britannica* made steady progress.

Meanwhile, Germany, at first largely dependent on translations of foreign encyclopaedias, had produced the scholarly "Hübner" (1704)—the first of the *Konversationslexikon* type—in a form that appealed to the rapidly growing middle class of the country, who welcomed encyclopaedias designed to provide them with an adequate cultural background for polite society. Johann Theodor Jablonski's illustrated *Allgemeines Lexicon* (1721) continued in this same style, and there were similar works compiled by the Swiss theologian and philologist Jakob Christoph Iselin and Antonius Moratori (1727). Johann Heinrich Zedler's huge *Grosses vollständiges Universal-Lexicon* ("The Great Comprehensive Universal Lexicon"; 1732–50) was in the older tradition but is important for its accuracy and its biographical and bibliographical material. An attempt to produce a German type of *L'Encyclopédie* in 1778–1807 was, however, a failure. Friedrich Arnold Brockhaus recognized the real need of the German people. Reworking Renatus Gotthelf Löbel's bankrupt encyclopaedia, he produced his first *Konversations-Lexikon* (1796–1811), thereby setting the pattern for at least half of all succeeding encyclopaedias throughout the western world. Brief, well-designed articles tightly packed with facts, comprehensive coverage, and a reputation for accuracy and up-to-dateness were the ingredients for one of the most successful of encyclopaedias.

Brockhaus' *Konversations-Lexicon*

The 19th century. Having served a long apprenticeship as a reviser of Chambers' *Cyclopaedia*, Abraham Rees at last produced a completely original and finely illustrated work, *The New Cyclopaedia* (1802–20), the only serious rival to the *Britannica* in a generation that saw some dozen "new" encyclopaedias rise and fall. What might have been the greatest encyclopaedia of the century, the *Encyclopaedia Metropolitana* (1817–45), failed miserably because of the early withdrawal of its designer, Samuel Taylor Coleridge, and subsequent financial troubles; but from it came the most notable contribution to the philosophy of encyclopaedia making since Bacon—Coleridge's profound "Treatise on Method" (1817).

To the principal influences on the compilation of encyclopaedias—Bacon, Diderot, the *Britannica*, and *Brockhaus*—must be added that of the Frenchman Pierre Larousse. His completely original approach to encyclopaedia making has given the series of encyclopaedias that bear his name a unique reputation. Emphasis throughout has been on readability; style has never been sacrificed to conciseness, and the successive editors of *Larousse* have paid very close attention to the changing public taste among French readers concerning the presentation of information.

The advent of Noah Webster was fully as epoch-making as that of Brockhaus and Larousse. Webster's informative *American Dictionary of the English Language* (1828) was encyclopaedic in character, but he avoided the long entries for the more important subjects that were such a feature of *Larousse*. Webster's approach appealed to the American taste and captured a huge market that has only increased with the years.

Brockhaus soon faced opposition, for his encyclopaedia was stronger on the humanities than on scientific and technical subjects. Joseph Meyer's *Der grosse Conversations-Lexikon* (1840–52) rectified this imbalance and was the first of a highly successful series that competed vigorously with *Brockhaus* for 100 years. In addition, Herder's *Conversations-Lexikon* (1853–57) and its subsequent editions provided the Catholic counterbalance in a country where Protestants and Catholics were almost equal in numbers.

The market for encyclopaedias in 19th-century Great Britain seemed inexhaustible, but many publishers lost money by putting out works that failed to capture the public's fancy. An exception was *Chambers's Encyclopaedia* (1860–68), which was unconnected with Ephraim Chambers' classic. Influenced by childhood access to a copy of the *Britannica*, Robert Chambers and his brother William compiled an original work, *Chambers's Encyclopaedia*, that took the *Konversationslexikon* form and thus found a new market that has continued to the present day.

Chambers's Encyclopaedia in England

In the first half of the 19th century there was increasing activity in other countries. Poland produced the *Encyklopedia Powszechna* (1858–68), known as "Orgelbrand" after its publisher. The Hungarians had followed the Bohemian *Slovník naučný* ("Scientific Dictionary"; 1860–90) with the *Egyetemes magyar encyclopaedia* ("Universal Hungarian Encyclopaedia"; 1861–76). The Russians had produced half an encyclopaedia, V.N. Tatishchev's *Leksikon rossyskoy* ("Russian lexicon"), in 1793, and then issued A. Starchevsky's *Spravochny entsiklopedichesky slovar* ("Encyclopaedic Reference Dictionary"; 1847–55) on the *Brockhaus* model. More important was the famous *Entsiklopedichesky slovar* ("Encyclopaedic Dictionary"; 1895), which became known as "Granat" after the Granat Russian Bibliographical Institute that produced it. A later edition (1910–48) of "Granat," in 58 volumes, is not exported from the Soviet Union. Modelled on the *Britannica*, this edition contained many important articles, such as Lenin's contribution on "Marx" and on "The Russian 19th-Century Agrarian Problem." Successive ideological changes in Russian society caused many changes in the text of "Granat," and it remains one of the most inaccessible of all Russian encyclopaedias outside the Soviet Union.

Larousse did not go unchallenged. Inspired by the French politician Ferdinand-Camille Dreyfus, *La Grande Encyclopédie* (1886–1902) provided France with a superb, authoritative, and comprehensive work, well documented, and of excellent scholarship throughout; it is still in use today. In Denmark the century ended with the issue of no fewer than three new good multivolume encyclopaedias: *Allers* (1892–99), *Hagerups* (1892–1900), and *Salmonsens* (1893–1911), a situation without parallel in the history of encyclopaedias. During the course of the century practically every feature of the modern encyclopaedia had been introduced, and editorial standards had at times risen to a height that modern editors can only envy.

The 20th century. In 1890–1906 a Russian edition of *Brockhaus*, which subsequently had considerable success, was issued from the St. Petersburg office of *Brockhaus*. In contrast, S.N. Yushakov designed his *Bolshaya entsiklopedya* ("The Great Encyclopaedia"; 1900–09) on the "Meyer" model. After "Granat" the next important encyclopaedia was the 65-volume *Bolshaya sovetskaya entsiklopedya* ("The Great Soviet Encyclopaedia"; 1926–47), which was eventually discredited; the second edition (1949–58) had a Marxist-Leninist approach but was less biassed on nonpolitical subjects. It represented almost the whole of the Soviet Union's cultural resources: 8,000 scholars contributed articles, and the appended bibliographies were truly international in scope. One complete

Russian encyclopaedias

volume was devoted to the Soviet Union. The yearbooks that supplemented this encyclopaedia were very well produced and maintained the high standards of the original work.

There has also been a series of editions of the much smaller *Malaya sovetskaya entsiklopedya* ("The Little Soviet Encyclopaedia"), first issued in 1928–31.

In the U.S., the first edition of the *New International Encyclopaedia* was issued in 1902–04 and was subsequently supplemented by yearbooks. *The Encyclopedia Americana*, which traces its ancestry back to an English-language adaptation (1829–33) of the 7th edition of *Brockhaus*, took on new strength in 1902 when the editor of *Scientific American*, Frederick C. Beach, was appointed editor of the *Americana*. It has enjoyed growing success through its policy of following the continuous revision system, and yearbooks have supplemented it from 1923 onward. In 1950–51 a completely new American work, *Collier's Encyclopedia*, appeared in 20 volumes, and subsequent editions have been supplemented by yearbooks since 1960. *Collier's* is noted for its large number of illustrations and maps.

The "Espasa," the *Enciclopedia universal ilustrada europeo-americana* (1905–70), is not only a great encyclopaedia, but—like the *Enciclopedia italiana*—it eschews revision in favour of a series of sizable supplements. Physically, it is the largest current encyclopaedia in the world. One complete volume is devoted to Spain and is separately revised and reissued from time to time. A smaller encyclopaedia, the *Diccionario Salvat* (first issued in 1907–13), is revised at frequent intervals. Another major Spanish encyclopaedia, the *Enciclopedia labor* (1955–60), is remarkable for devoting each volume to a major subject field, an index volume providing the key to the total contents. This encyclopaedia is valuable for the attention it pays to every part of the Spanish-speaking world.

Italian encyclopaedias

One of the most important of all encyclopaedias, the *Enciclopedia italiana* (1929–39), is famous for its lavish production, its superb illustrations, and its lengthy, scholarly, and well-documented articles. Even its defense of Fascist ideology is not allowed to impinge on the general impartiality of the text. Supplements have been issued since World War II. The postwar *Dizionario enciclopedico italiano* (1955–63), issued by the same publishers, is a much smaller but well-illustrated work. In Germany, the three giants of the German encyclopaedia world—*Brockhaus*, "Meyer," "Herder"—continue to produce new editions, unchallenged by any new encyclopaedia of comparable scope.

In spite of the continuing popularity of *Larousse*, France has produced three other encyclopaedias of note in the 20th century. The *Encyclopédie française* (1935–66) is an outstanding collection of monographs by well-known scholars and specialists, arranged in classified form and issued in looseleaf binders, supplemented by a continuously revised index. The *Encyclopédie de la Pléiade* (1955–) is an encyclopaedic series, each work (some in more than one volume) being a self-contained treatment of a broad subject field written in narrative form. One of the most interesting new encyclopaedias is the *Encyclopaedia Universalis* (1968–), edited by Claude Grégory and owned by the French Book Club and Encyclopædia Britannica, Inc. This work, inspired by *L'Encyclopédie*, eschews the inclusion of minor items in favour of extensive and very well-illustrated articles on important subjects, and it pays special attention to modern science and technology. Of the 20 volumes planned, volumes 17–19 will constitute a "thesaurus" of fact entries and index, and volume 20 will be devoted to an exposition of the classifications used in constructing the encyclopaedia and to further elaborations.

Most countries have issued at least one good encyclopaedia, among which should be mentioned:

Bulgaria: *Kratka bŭlgarska entsiklopediia* (1963–).

Czechoslovakia: *Ottův slovník naučný* (1888–1909); *Masarykův slovník naučný* (1925–33); *Nový velký ilustrovaný slovník naučný* (1929–34); *Komenského slovník naučný* (1937–38); *Příruční slovník naučný* (1962-67).

Denmark: *Gyldendals Nye leksikon* (first issued 1931–32); *Raunkjaers konversations leksikon* (1948–57); *Gyldendals store opslagsbog* (1967–70).

Estonian S.S.R.: *Eesti nõukogude entsüklopeedia* (1968–).

Finland: *Otavan iso tietosanakirja/Encyclopaedia fennica* (1960–65); *Nuorten tieto* (1963–); *Fokus: Otavan kertovasti kuvitettu tietosanakirja* (1963–).

Greece: *Megalē hellenikē enkyklopaideia* (1926–34); *Eleutheroudakē enkyklopaidikon leksikon* (1927–31).

Hungary: *Révai nagy lexikona* (1911–35); *Új magyar lexikon* (1959–62); *Új idők lexikona* (1936–42).

The Netherlands: *Eerste nederlandse systematisch ingerichte encyclopaedie* ("ENSIE"; 1946–52); *Grote Winkler Prins encyclopedie* (1966– ; the latest in an outstanding series of "Winkler Prins" encyclopaedias starting in 1870–82); *Algemene nederlandse systematisch ingerichte encyclopedie* ("ANSIE"; 1955–); *Oosthoeks encyclopedie* (1968– ; the latest edition of a series starting in 1916–23).

Norway: *Aschehougs Konversasjons leksikon* (1970– ; the latest in a series started in 1907–13); *Norsk konversasjonsleksikon* (first published 1931–34); *Norsk allkunnebok* (1948–61).

Poland: *Wielka Encyklopedia Powszechna PWN* (1962–).

Romania: the Romanian Academy's *Dictionar enciclopedic român* (1962–66).

Sweden: *Bonniers konversations lexikon* (first published 1922–32); *Focus uppslagsbok* (1958–60).

Turkey: *Türk ansiklopedisi* (1946– ; first called *Inönü ansiklopedisi*).

Yugoslavia: *Enciklopedija Jugoslavije* (1955–); *Enciklopedija leksikografskog zavoda* (first published 1955); *Pomorska Enciklopedija* (1954–64); *Hrvatska enciklopedija* (1941–45; unfinished, A–El only).

Lithuania (published in Boston): *Lietuvių enciklopedija* (1953–69).

ENCYCLOPAEDIAS IN THE EAST

China. The contribution from the East to the history of encyclopaedias is distinctive and covers a longer period than that of the West. The Chinese have produced encyclopaedias for approximately 2,000 years, but traditionally they differ from the modern Western encyclopaedia in that they are mainly anthologies of significant literature with some elements of the dictionary. Compiled by scholars of eminence, they have been revised rather than replaced over hundreds of years. In the main, they followed a classified form of arrangement; very often their chief use was to aid candidates for the civil service. The first known Chinese encyclopaedia, the *Huang-lan* ("Emperor's Mirror"), was prepared by order of the emperor in about AD 220. No part of this work has survived. Part of the *Pien-chu* ("Stringed Pearls of Literature"), prepared around 600, is still extant. About 620 the *I-wen lei-chü* ("Anthology of Art and Literature") was prepared by Ou-yang Hsün (557–641) in 100 chapters divided into 47 sections. The *Pei-t'ang shu-ch'ao* ("Extracts for Books") of Yü Shih-nan (558–638) was more substantial and paid particular attention to details of the organization of public administration. An annotated edition, edited by K'ung Kuang-t'ao, was published in 1880.

The earliest Chinese encyclopaedia

The *Ch'u-hsüeh chi* ("Entry into Learning") was a modest work compiled about 700 by Hsü Chien (659–729) and his colleagues. A more important book was the *T'ung-tien* ("Comprehensive Statutes") compiled by Tu Yu (735–812), a writer on government and economics. Completed in about 801, it contained nine sections: economics, examinations and degrees, government, rites and ceremonies, music, the army, law, political geography, national defense. In 1273 it was supplemented by Ma Tuan-lin's enormous and highly regarded *Wen hsien t'ung k'ao* ("General Study of the Literary Remains"), which included a good bibliography. Supplements to this work were published in the 17th, 18th, and 20th centuries. Under the order of the second Sung emperor, Sung T'ai Tsung, the statesman Li Fang organized the

compilation of the vast *T'ai-p'ing yü-lan* ("Emperor's Mirror"), which included extracts from many works of literary and scientific standing that are no longer extant. In 1568–72 the *T'ai-p'ing yü-lan* was revised and reprinted from movable type; a new edition revised by Yüan Yüan appeared in 1812. The *Ts'e-fu yüan-kuei* (*c.* 1013), particularly strong in historical and biographical subjects, was almost as large as the *T'ai-p'ing yü-lan.*

The historian Cheng Ch'iao (1108–66) compiled the *T'ung chih* ("General Treatises"), an original work with a strong personal contribution; the printed edition (1747) was in 118 volumes. One of the richest and most important of all Chinese encyclopaedias, the *Yü-hai* ("Sea of Jade"), was compiled about 1267 by the renowned Sung scholar Wang Ying-lin (1223–92) and was reprinted in 240 volumes in 1738.

World's largest encyclopaedia

What was probably the largest encyclopaedia ever compiled, the *Yung-lo ta-tien* ("Great Handbook"), was issued at the beginning of the 15th century. Unfortunately, only a very small part of its 22,937 chapters has survived; these were published in 1963. A number of small encyclopaedias were issued in the 16th century, but the next important event was the publication of the small but profusely illustrated *San ts'ai t'u-hui* (1607), compiled by Wang Ch'i and his son Wang Ssu-i. In 1704–11 the Chinese literary encyclopaedia *P'ei-wen yün-fu* was compiled by order of the emperor K'ang-hsi; this was supplemented by the *Yün fu shi I* (1720). Other works ordered by the emperor include the *Pien-tzu lei-pien* (1726) and the *Tzu shih ching hua* (1727). In 1726 the huge *Ku-chin t'u-shu chi-ch'eng* ("Collection of Pictures and Writings") was published by order of the emperor. Edited by the scholar Ch'en Meng-lei, it filled over 750,000 pages and attempted to embody the whole of the Chinese cultural heritage.

At the turn of the century, a number of encyclopaedias were issued. Wang Chi's *Shih wu yüan hui*, which covered well over 2,000 topics, was compiled in 1796. Lu Feng-tso's *Hsiao chih lu* (1804) is particularly valuable for its attention to technical terms, which previous works had ignored. Ch'en Wei's *Ching Chuan II* (1804) concentrated on history and the great Chinese classics, whereas Wang Ch'eng-lieh's *Ch'i ming chi shu* (1806) is stronger in biographical material. Tai Chao-ch'un compiled the *Ssu shu wu ching lei tien chi ch'eng* (1887), a historical work for the use of civil-service candidates. Wei Sung's *I shih chi shih* (1888) had actually been compiled 65 years previously, but it paid far more attention to practical matters. The *Chiu T'ung T'ung* (1902) of Liu K'o-i was in large measure a reassembly of material in the older encyclopaedias in a more efficient classification. A more important work of the period is the largely historical and biographical *Erh shih ssu shih chiu t'ung cheng tien lei yao ho pien* (1902). The *Ch'ing ch'ao hsü wen hsien t'ung k'ao* (1905), compiled by Liu Chin Tsao, was revised and enlarged in 400 volumes in 1921. It includes contemporary material on fiscal, administrative, and industrial affairs and gives some attention to technical matters. Lu Erh-k'uei's *Tz'u-yüan* (1915), with a supplement issued in 1931, was the first really modern Chinese encyclopaedia and set the style for nearly all later works of this nature.

Japan. In the Edo, or Tokugawa, era (1603–1867) there appeared a kind of encyclopaedia that consisted of extracts of major works in Japanese and Chinese. *Koji-ruien* (51 volumes, 1879–1914) and *Nihon-hyakka-daijiten*, or the "Great Japanese Encyclopaedia" (10 volumes, 1908–19) were somewhat more akin to modern encyclopaedias but were mostly compilations of scientific works. More complete general encyclopaedias appeared in the Showa period (1926–); *Dai-hyakka* (28 volumes, 1931–35), *Kokumin-hyakka* (15 volumes, 1934–37), *Sekai-dai-hyakka* (24 volumes, 1955–68), and *Japonica* (19 volumes, 1967–72) are examples of well-compiled works. In 1972 the *Britannica International Encyclopaedia* (28 volumes when completed) began publication. Unlike most of the formerly published Japanese encyclopaedias, which featured simple short entries, this work consisted of systematic long entries, supplemented with shorter "fact index" entries.

The Arab world. The early encyclopaedias written in Arabic can be roughly divided into two classes: those designed for people who wished to be well informed and to make full use of their cultural heritage, and those for the rapidly growing number of official administrators. The latter type of encyclopaedia originated when the Arabs established their rule through so many parts of the Mediterranean region. The first true encyclopaedia was the work of Ibn Qutayba (828–889), a teacher and philologist, who dealt with his topics by quoting traditional aphorisms, historical examples, and old Arabic poems. The arrangement and contents of his *Kitāb ʿUyūn al-Akhbār* ("The Best Traditions") set the pattern for many later encyclopaedias. The ten books were arranged in the following order: power, war, nobility, character, learning and eloquence, asceticism, friendship, prayers, food, women. Ibn ʿAbd Rabbih of Córdoba improved on Ibn Qutayba's work in his *ʿIqd* ("The Jewelled Necklace") by including more contemporary items of note.

Work of Ibn Qutayba

What has often mistakenly been referred to as the first encyclopaedia, the *Mafātīḥ al-ʿUlūm* ("Key to the Sciences"), was compiled in 975–997 by the Persian scholar and statesman al-Khwārizmī, who was well aware of the content of the more important Greek writings. He divided his work into two sections: indigenous knowledge (jurisprudence, scholastic philosophy, grammar, secretarial duties, prosody and poetic art, history) and foreign knowledge (philosophy, logic, medicine, arithmetic, geometry, astronomy, music, mechanics, alchemy). The Ikhwān aṣ-Ṣafāʾ ("Sincere Brethren"), a religious or political party founded at Basra in the 10th century, published the *Rasāʾil Ikhwān aṣ-Ṣafāʾ*, a remarkable work that consisted of 52 pamphlets written by five authors, comprising all the knowledge available in their milieu. The work included (1) mathematics, geography, music, logic, and ethics; (2) the natural sciences and philosophy; (3) metaphysics; and (4) religion, astrology, and magic. A complete edition was published in 1887–89.

The Egyptian historian and civil servant an-Nuwairi (1272–1332) compiled one of the best known encyclopaedias of the Mamlūk period, the *Nihāyat al-ʿarab fī funūn al-adab* ("The Aim of the Intelligent in the Art of Letters"), a work of almost 9,000 pages. It comprised: (1) geography, astronomy, meteorology, chronology, geology; (2) man (anatomy, folklore, conduct, politics); (3) zoology; (4) botany; (5) history. A complete edition was issued in 1923. The *Masālik al-abṣār* ("Sight-Seeing Journeys") of al-ʿUmarī (1301–48) was chiefly strong on history, geography, and poetry. A third Egyptian, al-Kalkashandī (died 1418), compiled a more important and well-organized encyclopaedia, *Ṣubḥ al-aʿshā*, that covered geography, political history, natural history, zoology, mineralogy, cosmography, and time measurement. Ibshīhī (flourished 1440) compiled a very individual encyclopaedia, the *Mustatraf* ("Spiritual Discoveries"), that covered the Islāmic religion, conduct, law, spiritual qualities, work, natural history, music, food, and medicine. At the turn of the Arab fortunes, Ibshīhī had recapitulated all that was best in their culture.

15th- and 16th-century Arabic encyclopaedias

The Persian lawyer ad-Dauwānī (1427–1501) published a kind of encyclopaedia, entitled *Unmūdag al-ʿūlūm*, that consisted of documented questions and answers and technical inventions on a very wide range of subjects. As-Sīrazī (died 1542) soon issued a refutation to it, the *Maqālatar radd ʿalā unmūdag al-ʿulum al-Galā lija.* The *Magmaʾ multaqat az-zuhūr birauḍa min al-manẓūm wal mantur* (1524) of al-Ḥanafī comprised an encyclopaedic survey and description of the various branches of knowledge, with an appendix containing an alphabetical list of the names of God. In Lebanon, Buṭrus al-Bustānī and his sons compiled the *Dāʾirat al-maʿārif* (1876–1900). A second edition (1923–25) was prepared by Muḥammad Farid Wajdi, and a third edition was begun by Fu'ād Afrām al-Bustānī in 1956. In 1955 Albert Rihani issued the one-volume *al-Mawsūʿat al-ʾarabīyah.*

Other areas. Other important encyclopaedias from the East include:

Burmese: the Burma Translation Society's *Myanma swezon kyan/Encyclopaedia Burmanica* (1954–).

Hebrew: *ha-Entziqlopedia ha-'Ivrit* (*Encyclopaedia Hebraica* 1949–); *Entziqlopedia tevel u-melo'a* (1962–); *Entziqlopedia kelalit Massadah* (1958–60).
Indonesian: *Ensiklopedia Indonesia* (1954).
Sinhalese: *Sinhalese Encyclopedia* (1963–).

AIMS AND EFFECTS OF ENCYCLOPAEDIAS

For much of their existence encyclopaedias have tried to be all things to all men. The effort has been magnificent, the results uncertain. No one can know what specific influence an encyclopaedia may have had on the thoughts and actions of any particular individual. One can be sure, however, that no one can look at an encyclopaedia without reacting in some way to its contents. The most ordinary effect is to remind the reader how little he knows of the world around him. This may produce in him a feeling that he is content to leave the situation as it is; it is far more likely to stimulate him at least to make some temporary effort to improve his understanding, and out of this may develop further studies that may interest him for a lifetime.

While trying to be a source for all that the reader wants, encyclopaedias have nevertheless omitted some things throughout their history. For a long time biographical information concerning living people was not thought suitable for inclusion. Humour and satire have always been found out of place; they date a work or become incomprehensible after a time. Up until the 19th century, "do-it-yourself" items were hardly ever more than a sporadic feature of any encyclopaedia, but nowadays many encyclopaedias cover a bewildering mass of facts and explanations. There are, however, still numerous aspects of ordinary life that are not considered part of encyclopaedic information. For example, it is extremely doubtful that an encyclopaedia will describe the kinds of clothing that are advisable for a visit to a foreign land, the precautions to be taken, the local customs to observe. Rightly or wrongly, this type of information is regarded as the preserve of the travel guide.

Even in this age, the editorial policies of most encyclopaedias have drawn a line between domestic needs—such as food recipes, touring routes, postal charges—and public concerns—such as population statistics, historical facts, explanations of scientific phenomena and technical inventions.

BIBLIOGRAPHY. There are two short and very readable introductions to the subject: JAMES M. WELLS, *The Circle of Knowledge: Encyclopaedias Past and Present* (1968), a well-illustrated guide issued in connection with a Newberry Library exhibition; and S.H. STEINBERG, "Encyclopaedias," *Signature*, New Series, no. 12, pp. 3–22 (1951), which is a brilliant conspectus of the whole field of encyclopaedia history. R.L. COLLISON, *Encyclopaedias: Their History Throughout the Ages*, 2nd ed. (1966), lists and describes in one chronological sequence encyclopaedias from both East and West, and pays particular attention to *L'Encyclopédie*, *Brockhaus*, the *Britannica*, the *Metropolitana*, and *Larousse*. It also includes a reprint (pp. 238–295) of SAMUEL TAYLOR COLERIDGE'S philosophical essay on the design of encyclopaedias, the "Treatise on Method." FRITZ SAXL, "Illustrated Mediaeval Encyclopaedias," in his *Lectures*, vol. 1, pp. 228–254, and vol. 2, plates 155–174 (1957), is an important and original contribution to the subject, the 20 illustrations being especially interesting. The *Journal of World History* devoted one complete issue (vol. 9, no. 3, 1966) to an international symposium on encyclopaedias, special attention being paid to St. Isidore, Hugh of Saint-Victor, Raoul Ardent, Vincent of Beauvais, Sahagún, *L'Encyclopédie*, the *Metropolitana*, the *Britannica*, *L'Encyclopédie française*, and Arabic and Chinese encyclopaedias of the classical period. A.J. WALFORD (ed.), *Guide to Reference Material*, 2nd ed., vol. 3, pp. 83–107 (1970); and CONSTANCE M. WINCHELL, *Guide to Reference Books*, 8th ed., pp. 81–91 (1967), and their supplements, both provide scholarly evaluations of the principal current English- and foreign-language encyclopaedias. GERT A. ZISCHKA, *Index Lexicorum* (1959), is important both for its excellent summary of the history of the encyclopaedia and for its extensive bibliography of past and present encyclopaedias. "The Uses of Encyclopaedias: Past, Present, and Future," in the *American Behavioral Scientist*, 6:3–40 (1962), is a stimulating symposium with contributions by Livio C. Stecchini, Jacques Barzun, Harry S. Ashmore, W.T. Couch, Charles Van Doren, Francis X. Sutton, David L. Sills, Carl F. Stover, and Alfred de Grazia. HERMAN KOGAN, *The Great EB* (1958), is a well-written and fascinating account of the *Britannica* and its history, but it is also valuable for the light it throws on the more practical problems and techniques of the encyclopaedia world in general. S. PADRAIG WALSH, *Anglo-American General Encyclopedias* (1968), is a dictionary of English-language encyclopaedias issued during the period 1703–1967, and has helpful evaluative notes. In 1969 the REFERENCE AND SUBSCRIPTION BOOKS REVIEW COMMITTEE OF THE AMERICAN LIBRARY ASSOCIATION issued "Purchasing a General Encyclopaedia," a pamphlet suggesting 12 criteria for evaluating the quality and usefulness of any encyclopaedia, and containing the committee's recommendations concerning a number of named encyclopaedias. In each encyclopaedia the entry under the word "Encyclopaedia" or "Encyclopedia" will usually (but not invariably) provide information concerning its own history, and often gives very useful information on the history of encyclopaedias in general. Additional details may often be found in the encyclopaedia's general introduction, which is usually printed in the first volume.

(R.L.C.)

Endocrine System, Human

The human endocrine system is made up of a large number of glandular tissues that have in common the fact that they secrete directly into the blood stream "chemical messengers" that regulate and integrate body functions. These chemical messengers, or hormones, usually but not invariably act on tissues some distance away from the cells that secrete them. One function of the endocrine system is the regulation of reproduction and closely associated phenomena. Other functions resemble those of the nervous system in that they help individuals cope with changes in their environment. Both hormonal and neural messages are specific. Nerve impulses pass along their axons (see NERVE IMPULSE). Hormones, on the other hand, enter the general circulation and are spread to all or almost all of the tissues in the body. Only certain tissues, however, are "programmed" to respond to a given hormone. Endocrine messages thus owe their specificity to the ability of those particular tissues to "understand" them.

This article is divided into the following major sections:

The nature of the human endocrine system and its hormones
The glands and tissues comprising the system
- The testes
- The ovaries
- The thyroid gland
- The adrenal glands
- The pituitary gland
- The structure and endocrine functions of the hypothalamus
- The pancreatic islets
- The parathyroid glands
- The gastrointestinal mucosa
- The thymus
- The pineal body
- The kidneys
- Other possible hormones and endocrine organs

The endocrine systems of animals other than man are dealt with in the article ENDOCRINE SYSTEMS. For a detailed description of the secretions of those systems, see the article HORMONE.

The nature of the human endocrine system and its hormones

Many known tissues secrete hormones in humans; probably still others are as yet undiscovered. The principal endocrine glands and tissues are shown in Figure 1, and the known or suspected endocrine tissues in humans, their hormones, and the chemical nature of these secretions are summarized in Table 1. Some hormones are tropic hormones that act on other endocrine cells. Four of the anterior pituitary hormones, for example, are tropic hormones, and through them the pituitary regulates the secretions of the adrenal cortex, the thyroid, and the sex glands.

Many hormones are proteins and polypeptides—*i.e.*, complex chains of amino acids. The structure of most if not all of these hormones varies from one animal species to another. In mammals the differences are usually so

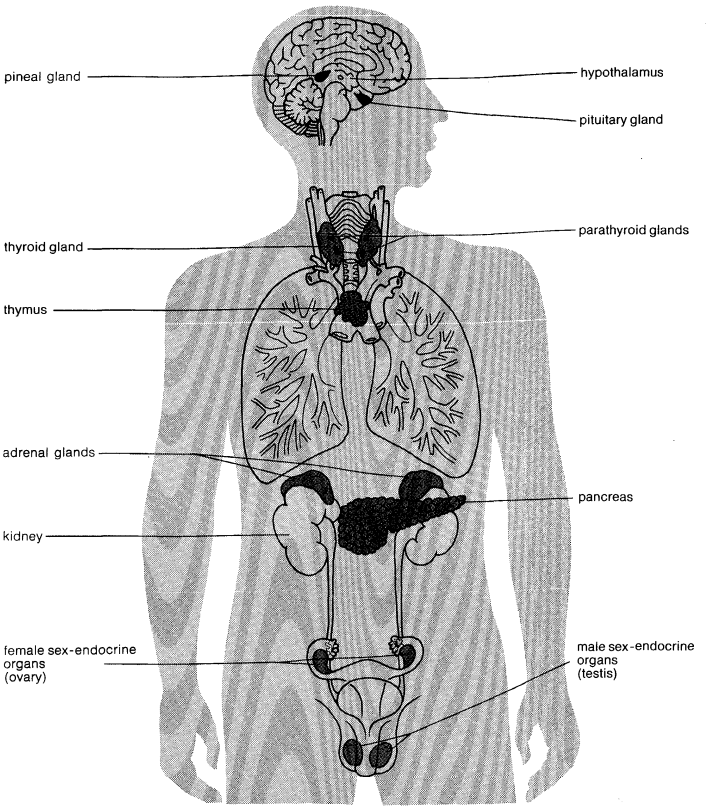


Figure 1: The human endocrine glands.

slight that the protein hormones of one species are generally, but not always, active in others. The structural differences are detected by the antibody-forming mechanism, however, and antihormone antibodies develop over a period of time. In human diabetics treated for prolonged periods with insulins of animal origin, for example, this can become a serious problem.

Neural control of the endocrine system

A large segment of the endocrine system is under nervous system control. Changes in the external or internal environment initiate changes in the sense organs that monitor these environments. Impulses set up in the nerve fibres from these receptors travel to areas of the brain where they make connections that activate effector mechanisms. The effectors available to the nervous system include somatic, smooth, and cardiac muscles, and glands.

The somatic muscles are the voluntary muscles responsible for moving, speaking and breathing. The smooth and cardiac muscles are responsible for visceral functions (*e.g.*, the peristaltic movements of the digestive tract are brought about by smooth muscle; only the heart contains cardiac muscle). Some of the gland cells are endocrine; *i.e.*, they secrete hormones directly into the bloodstream. The endocrine glands that are at least in part under neural control include the adrenal medulla, the juxtaglomerular cells of the kidneys, the beta cells in the islets of the pancreas, and the pituitary gland. These glands will all be described in the paragraphs that follow. Because the pituitary secretes at least ten hormones, of which four are hormones tropic to other endocrine glands, its regulatory effects are extremely widespread.

The nature of neural control varies from gland to gland. In the case of the adrenal medulla, the secretory cells are innervated by nerve cells. In the case of the juxtaglomerular cells, the regulatory effect appears to be mediated by circulating epinephrine (adrenalin), the hormone secreted by the adrenal medulla, as well as by a direct innervation of the secretory cells. The regulation of pituitary secretion is largely a function of the hypothalamus, the portion of the brain that lies immediately above the pituitary gland. In the anterior pituitary, the gland cells are regulated by hormones called "releasing factors" and "inhibiting factors" secreted by nerve cells in the hypothalamus into the portal hypophysial vessels, a collection of special blood channels that run directly from the hypothalamus to the anterior lobe of the pituitary. The posterior pituitary hormones are secreted directly into the general circulation from the endings of neurons whose cell bodies are located in the hypothalamus. Intermediate lobe hormones apparently are secreted from innervated gland cells and possibly are regulated by humoral agents from the hypothalamus as well.

Secretion of hormones by such neurons is not unique. Most, possibly all, neurons secrete chemical substances at their endings. In some cases, these substances act locally to produce excitation in the nerve or muscle cell on which the particular neuron ends; in other cases, the secreted material enters the blood or lymph circulation.

Neuroendocrine control of endocrine function is important in that it regulates the secretions of a large part of the endocrine system in accordance with the needs of the individual.

Feedback control

Another mechanism that regulates endocrine secretion is known as feedback control. Negative feedback is important. The secretion of parathyroid hormone is an example of negative feedback: parathyroid hormone increases blood-plasma calcium levels, and increased plasma calcium acts directly on the parathyroid gland cells to inhibit parathyroid secretion. If an extraneous factor increases plasma calcium, the secretion of parathyroid hormone is decreased and the rise in calcium is counteracted. A similar negative feedback plays a regulatory role in the interaction between the pituitary gland and some of the endocrine organs on which its hormones exert tropic effects. Adrenocorticotropic hormone (ACTH) from the pituitary, for example, increases the secretion of cortisol, a hormone of the adrenal cortex; and cortisol feeds back in a negative fashion to inhibit ACTH secretion.

Positive feedback also occurs in a few special circumstances in the endocrine system when a change in the substance regulated feeds back to change endocrine functions in a way that accentuates, rather than counteracts,

Table 1: The Human Endocrine System

gland or tissue	hormone	chemical nature
Testis	testosterone	steroid
Ovary	17β-estradiol	steroid
	progesterone	steroid
	relaxin	polypeptide
Thyroid gland	thyroxine	amino acid
	triiodothyronine	amino acid
	calcitonin	polypeptide
Adrenal gland		
Medulla	epinephrine	amine
	norepinephrine	amine
Cortex	cortisol	steroid
	corticosterone	steroid
	aldosterone	steroid
	androgens	steroid
	estrogens	steroid
Pituitary gland		
Anterior lobe	adrenocorticotropin (ACTH, corticotropin)	polypeptide (mw = 4,567)*
	growth hormone (GH, STH, somatotropin)	protein (mw = 21,734)*
	thyrotropin (TSH, thyroid-stimulating hormone)	glycoprotein (mw ≈ 25,000)*
	follicle-stimulating hormone (FSH)	glycoprotein (mw ≈ 30,000)*
	luteinizing hormone (LH, interstitial cell stimulating hormone ICSH)	glycoprotein (mw = 26,000)*
	prolactin (luteotropic hormone, LTH, lactogenic hormone, manmatropin)	protein (mw = ?)
Posterior lobe	oxytocin	polypeptide (mw = 1,007)*
	vasopressin (ADH, antidiuretic hormone)	polypeptide (mw = 1,228)*
Intermediate lobe tissue	α-melanocyte-stimulating† hormone (α-MSH)	polypeptide (mw = 1,823)* (not secreted?)
	β-melanocyte-stimulating† hormone (β-MSH)	polypeptide (mw = 2,734)*
Hypothalamus	corticotropin releasing factor (CRF)	polypeptide?
	growth hormone releasing factor (GRF)	polypeptide?
	thyrotropin releasing factor (TRF)	polypeptide
	follicle-stimulating hormone releasing factor (FSHRF)	polypeptide?
	luteinizing hormone releasing factor (LHRF)	polypeptide
	prolactin inhibiting factor (PIF)	polypeptide?
Pancreatic islets	glucagon	polypeptide
	insulin	polypeptide
	gastrin	polypeptide
Parathyroid gland	parathyroid hormone	polypeptide
Gastrointestinal mucosa		
Stomach	gastrin I	polypeptide
	gastrin II	polypeptide
Duodenum	cholecystokinin-pancreozymin (CCK)	polypeptide
	secretin	polypeptide
	intestinal glucagon (GLI)	polypeptide?
Small intestine	villikinin	—
	enterocrinin	—
Thymus	thymosin	polypeptide
Pineal	melatonin	amine (not secreted?)
Kidneys	renin	protein
	renal erythropoietic factor (REF)	protein?
Placenta	human chorionic gonadotropin (HCG)	glycoprotein
	human chorionic somatomammotropin (HCS)	protein
	renin	protein
	estrogens	steroids
	androgens	steroids
	progesterone	steroid

*Molecular weight (mw) in daltons. †Intermediate lobe hormones referred to collectively as melanotropin or intermedin.

the change. It is apparent that such control must be short-lived or self-limiting.

An example of positive feedback that includes a neuroendocrine link is the regulation of the posterior pituitary hormone oxytocin secretion during the delivery of a baby. This hormone acts on the uterus to increase its contractions. As the baby's head begins to dilate the neck of the uterus and to descend into the birth canal, nerve impulses from these regions pass up the spinal cord to the brain, where they stimulate oxytocin secretion from the posterior pituitary. The oxytocin in turn increases the force of uterine contractions, pushing the baby's head further along and causing greater sensory stimulation with increased oxytocin secretion. The resulting crescendo of oxytocin secretion is not primarily responsible for delivery but does aid parturition. The process ceases, of course, with the expulsion of the baby from the mother's body. Positive feedback is probably also involved in the secretion of the hormones responsible for ovulation.

The negative feedback component in the control of the secretion of some hormones is indirect. This is particularly true of hormones that are under hypothalamic control. Hormones secreted by the ovaries and testes, for example, exert a negative feedback effect primarily on the hypothalamus, with a secondary decrease in the secretion of the pituitary hormones that regulate ovarian and testicular hormone secretion. In the case of ACTH and thyroid stimulating, or thyrotropic, hormone (TSH), the target organ hormones appear to feed back on both the hypothalamus and the anterior pituitary gland. In addition, evidence has been found that the pituitary tropic hormones may feed back on the hypothalamus to either inhibit or stimulate the secretion of the corresponding hypothalamic releasing or inhibiting factors. This type of feedback has been dubbed "short-loop feedback." It has also been suggested that there is also a "minifeedback" in which releasing and inhibiting factors actually exert an inhibitory effect on their own secretion.

Protein binding of hormones

Another feature of general interest in the endocrine system is the binding or linking of a number of hormones to proteins in the circulation. Protein-bound hormones, generally believed to be physiologically inactive, exist in equilibrium with the free, active form of the hormone. Hormones that are protein bound include those of the thyroid and steroid secretions of the adrenal cortex, testes, and ovaries. The degree of binding varies. In the case of the adrenocortical hormones, there is relatively little binding of aldosterone, but bound cortisol accounts for about 96 percent of the total circulating cortisol. Practically all (99 percent) of the circulating thyroxine is bound. The function of the large circulating pools of bound, inactive hormones is uncertain, but it seems likely that they provide reservoirs of readily available hormones, as an extra safeguard against a dangerous drop in the amount of circulating free hormone.

Because there are so many hormones and their effects are extremely varied, it might be expected that the molecular mechanisms by which they exert their actions would be equally varied. Research, however, has demonstrated that there are two mechanisms of action common to a large number of hormones. One is by way of the cellular apparatus that regulates protein synthesis (genetic code), and the other via the intracellular formation of the nucleotide cyclic adenosine-3,5-monophosphate (cyclic AMP).

Evidence that some hormones act through the mechanism of the genetic code first came from experiments on insects. Deoxyribonucleic acid (DNA) in the nucleus of all cells contains the "instructions" for the formation of all the proteins that the cell can make. Parts of the genetic message are copied by messenger ribonucleic acid (mRNA), which then migrates to the ribosomes in the cytoplasm, where it supervises the formation of the proteins from amino acids (see also GENETICS; PROTEIN). The protein synthetic mechanism is the key to the function of the cell, since the synthesized proteins include the enzymes necessary to catalyze all of the cell's chemical reactions. All steroid hormones and a number of protein hormones appear to act in ways that lead to increased formation of DNA-dependent mRNA and consequently to synthesis of key enzymes. Thus, the hormones cause the expression of portions of the genetic message that are coded in the cell nucleus but that, in the absence of the hormones, are retained there in a repressed state.

Portions of the genetic message are transcribed in response to a given hormone only in certain cells, those said to be sensitive to the hormone. This specificity appears to be imparted by the presence in the sensitive, but not in the other, cells of a specific receptor protein that binds the hormone.

Cyclic AMP

An even more interesting feature of the action of hormones, and particularly of protein and polypeptide hormones, on cells is the generation of cyclic AMP, which is capable of initiating a variety of changes that mimic the actions of particular hormones. Cyclic AMP is formed from adenosine triphosphate (ATP) in a reaction catalyzed by the enzyme adenyl cyclase (see also METABOLISM) and is inactivated by a reaction catalyzed by the enzyme phosphodiesterase. Adenyl cyclase is found in the membranes of cells and consequently can be activated by chemicals that contact the outside of the cells without penetrating them. This had led to the hypothesis that some hormones act on their target cells by activating adenyl cyclase, and the intracellular cyclic AMP thus generated acts as a "second messenger" to bring about the hormone's action. The hypothesis is an attractive one because it explains how high molecular weight protein hormones can act without entering cells; it is difficult to visualize how such large molecules could enter cells without disrupting their membranes.

Since many hormones appear to act via cyclic AMP, but each of the hormones has its own characteristic actions, a given hormone must activate adenyl cyclase only in the membranes of its target cells. The basis of such specificity is not known, but it could be related to the presence of hormone-specific receptors in the cell membranes.

In certain instances, both the cyclic AMP second messenger system and DNA-dependent mRNA synthesis appear to be involved in the action of a single hormone. There is now evidence, however, that in some cases intracellular cyclic AMP initiates the expression of the portion of the genetic message that is responsible for the action of the hormone.

The second messenger system is frequently involved when the target tissue of a particular hormone is another endocrine gland. Cyclic AMP, thus, appears to be involved in the secretion of the kidney hormone renin in response to epinephrine and norepinephrine, the adrenocortical hormones in response to adrenocorticotropic hormone, the hormones of the corpus luteum of the ovary in response to luteinizing hormone (LH), the thyroid hormones in response to thyroid-stimulating hormone, and some of the hormones of the anterior pituitary in response to the hypothalamic releasing factors. The role of cyclic AMP in the secretory process is the subject of much current research.

The secretory process: emeiocytosis

There appear to be additional features common to the secretory process in various endocrine cells. Protein and polypeptide hormones, like other proteins and polypeptides, are formed in the endoplasmic reticulum (see CELL AND CELL DIVISION). In the cellular Golgi apparatus, the hormones are condensed into granules, and each granule is surrounded by a membrane. The granular particles thus formed are secreted by a process of emeiocytosis (literally, "cell vomiting"). The hormone granules move to the edge of the secretory cell, their membranes fuse with the cell membrane, and the area of fushion then degenerates, leaving the hormone granules outside the cell and the cell membrane intact. The process is the reverse of pinocytosis, in which substances are engulfed by protoplasmic extension of the cell and taken into its interior. A typical example of emeiocytosis is the secretion of the polypeptide hormone insulin. Epinephrine and norepinephrine are apparently bound to protein and stored in similar membrane-lined granules, then secreted by emeiocytosis.

Steroid-secreting cells have a prominent smooth endoplasmic reticulum that seems to play an important part in the formation of their secretions. Emeiocytosis is not involved in this secretory process. Steroids, however, are soluble in membrane constituents and probably can cross cell membranes with relative ease by dissolving in them.

Once released, the hormones must generally pass through the basement membranes of the particular endocrine cell and of the endothelial cells that form the wall of the blood capillaries. In most endocrine glands the endothelium itself is thinned out so that there are gaps, or fenestrae, between cells closed only by a thin, homogeneous membrane. This fenestrated endothelium is also found in the intestine and kidneys, where large amounts of substances are transported across the capillary walls. For some reason, however, such endothelium is absent in the testes.

In the following sections, each of the human endocrine glands is described. Each gland's location, structure, blood supply and innervation are considered; the nature, action and metabolism of its secretions are summarized; and the regulation of its secretion is reviewed.

The glands and tissues comprising the system

THE TESTES

The testes are the male gonads (see REPRODUCTIVE SYSTEM, HUMAN). They have two functions: they produce spermatozoa, the male germ cells, and they secrete testosterone, the hormone primarily responsible for male sex characteristics. The first function is the gametogenic function, and the second is the endocrine function. Testosterone and other hormones that produce masculinization are referred to collectively as androgens.

The testes are located in the scrotum, the wrinkled bag of skin at the base of the penis. In this location, the testes are kept cooler than the tissues in the interior of the body. The process of forming sperm cells proceeds best at a temperature several degrees below body temperature; if the testes were in the abdomen, sperm formation would reduce to the point of sterility.

The testes develop during embryologic life from structures in the abdomen, and when they later migrate to the scrotum they take their blood vessels with them. The internal spermatic arteries that bring blood to the testes are branches of the aorta. The spermatic veins empty into the renal vein on the left side of the body and the inferior vena cava on the right. The nerves are branches of the sympathetic and pelvic autonomic plexuses.

The spermatozoa are produced in the testes in a series of long, folded and coiled tubes called the seminiferous tubules. The tubules drain into a network of ducts in the epididymis, a large structure that lies along the side of the testes. The ducts in the epididymis coalesce to form the vas deferens. This latter duct passes into the abdomen and conducts sperm to the urethra, the urinary passage from the bladder to the tip of the penis.

Testosterone. Between the seminiferous tubules in the testes are nests of cells that secrete testosterone. These cells, which contain lipid granules, are called the interstitial cells of Leydig. Testosterone is a steroid that contains 19 carbon atoms. Steroid hormones are derived from the cyclopentanoperhydrophenanthrene nucleus. They are formed in the body from cholesterol. The male sex hormones secreted by the testes and adrenals are

Leydig cells

From Israel S. Kleiner and James M. Orten, *Biochemistry*, 7th ed. (1966); The C.V. Mosby Co., St. Louis

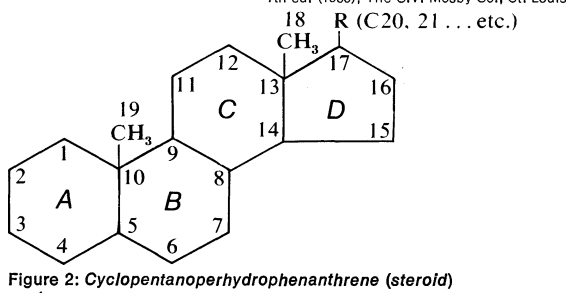


Figure 2: *Cyclopentanoperhydrophenanthrene (steroid) nucleus.*
The rings are lettered (*A*, *B*, *C*, *D*), and the carbon positions numbered. R denotes an aliphatic (nonringed hydrocarbon) side chain. Hormonal activity is altered to a large extent by additions or modifications at the 3, 11, and 17 positions (see text).

mostly carbon-19 steroids in that they contain 19 carbon atoms, while the adrenocortical steroids are mostly carbon-21 steroids that have a carbon-2 side chain on the *D* ring. Steroid formation is similar in all steroid-producing glands, and the pathways are comparable from gland to gland, the end product being determined by the particular enzymes in a given gland (see also STEROIDS).

The actions of testosterone. The actions of testosterone and of other androgens are primarily responsible for the characteristics associated with maleness. They stimulate the growth of the beard and increase the length of the larynx, producing deepening of the voice. They are protein anabolic; *i.e.*, they promote the formation of more protein in the body. They are thus responsible for the formation of the broad-shouldered, well-muscled body of the male. They also have a number of other effects, including apparent effects upon behaviour. The behavioral effects are difficult to quantitate, but testosterone does appear to promote aggressive, boisterous behaviour and to stimulate interest in sexual activity.

About 70 percent of the circulating testosterone is bound to a gonadal steroid-binding protein in the plasma, while the remaining testosterone in the circulation is in the free form. Its half-life in the circulation (time required for one-half a given amount to disappear from the circulation) is 90 minutes. The hormone is converted to various derivatives, most of which have an oxygen atom, usually contained in a keto group, attached to the *D* ring of the steroid molecule in position number 17. These compounds are therefore called 17-ketosteroids and are measured in the urine as an index of the endocrine function of the testes; 17-ketosteroids, however, are also formed from adrenal cortical hormones.

Control of testicular function

Both the gametogenic and the endocrine functions of the testes are under the control of the anterior lobe of the pituitary gland. The anterior pituitary hormone that regulates spermatogenesis is the follicle-stimulating hormone (FSH), and the anterior pituitary hormone that regulates testosterone secretion is LH. The secretion of FSH and LH is in turn under hypothalamic control, via follicle-stimulating and luteinizing hormones releasing factors, releasing factors secreted by the hypothalamus.

In the adult male, testosterone appears to be secreted at a steady rate and spermatogenesis is a steady, nonfluctuating process. Testicular function depends, therefore, on the maintenance of steady secretion of luteinizing hormone and follicle-stimulating hormone. Negative feedback control provides such stability. Luteinizing hormone secretion is under negative feedback control via testosterone. Luteinizing hormone stimulates testosterone secretion, and the circulating testosterone in turn inhibits LH secretion. The site at which the feedback occurs is probably the hypothalamus rather than the pituitary, and the testosterone appears to act by inhibiting the secretion of luteinizing hormone releasing factor, the hypothalamic factor which regulates the secretion of LH. Follicle-stimulating hormone stimulates spermatogenesis, and the secretion of it also appears to be under feedback control, since the secretion of this hormone increases if the testes are removed. The nature of the feedback inhibitor, however, has not yet been determined. It is thought probable that it is not testosterone, because in certain diseases FSH secretion is increased while testosterone secretion is normal. There is evidence that the testes produce a substance of unknown chemical nature called "inhibin," that suppresses follicle-stimulating hormone secretion.

THE OVARIES

The ovaries are the female gonads (see REPRODUCTIVE SYSTEM, HUMAN). They produce the female germ cells, or ova, and they secrete three hormones, estradiol, progesterone, and relaxin. In the adult female, each ovary is

Ovarian morphology

an irregularly ovoid almond-sized structure located in each side of the abdominal cavity at the edge of the pelvis. The ovaries receive their arterial blood supply from the ovarian arteries, branches of the aorta which are comparable to the internal spermatic arteries in the male. They join with the uterine artery. The ovarian veins, like the internal spermatic veins, drain into the inferior vena cava on the right and the left renal vein on the left. The nerves come from surrounding autonomic nervous system plexuses. The ovaries are located close to the fluted ends of the fallopian, or uterine, tubes, hollow conduits which pass on either side from the ovary to the uterus. The uterus, which in the nonpregnant woman resembles in size and shape an inverted pear, opens into the vagina through its opening, the cervix. The vagina, in turn, opens to the exterior.

The menstrual cycle. The theme of reproductive function in the mature female is the cyclicity of the menstrual cycle. This is in marked contrast to the male, in which the pattern is one of steady hormone secretion and germ cell production. In the female, cyclic changes in the ovary produce cyclic changes in the uterus, which culminate approximately every 28 days in menstruation (see also MENSTRUATION). The ovarian changes are, in turn, the product of sequential, orderly secretion of a number of anterior pituitary hormones under the control of a "biological clock" in the brain.

In the ovary of a young woman who has reached sexual maturity there are approximately 500,000 structures called primordial follicles. Each of these contains a primitive germ cell or ovum that had partially developed, then became dormant at the time of birth. At about the time each menstrual cycle ends, one of these follicles begins to enlarge. The hormone primarily responsible for the enlargement is follicle-stimulating hormone from the anterior lobe of the pituitary. If a woman receives injections of FSH, many follicles are usually stimulated, and women treated in this way frequently give birth to more than one baby. Indeed sextuplets, septuplets, and even nontuplets have been reported. The reason why follicle-stimulating hormone from the woman's own pituitary usually stimulates only one follicle, while exogenous FSH stimulates many, is unknown. The final maturation of the follicles requires both follicle-stimulating hormone and luteinizing hormone from the pituitary.

On approximately the 14th day of the menstrual cycle there is a sudden burst of luteinizing hormone and follicle-stimulating hormone secretion, and in response to the LH the mature (graafian) follicle ruptures. This rupture is the process of ovulation. The ovum is extruded from the follicle into the abdominal cavity and is usually promptly caught up in the folds around the opening of the adjacent fallopian tube. It is transported toward the uterus, largely by the action of the hairlike cilia that project from the cells lining the tube. If fertilization does not occur, the ovum passes through the uterus and out the vagina to the exterior.

The ruptured follicle fills promptly with blood, which clots to form what is called a corpus hemorrhagicum. The cells lining the corpus hemorrhagicum are stimulated, and yellowish, lipid-rich luteal cells rapidly grow into the clot. The ball of luteal cells thus formed is called the corpus luteum. In humans, the corpus luteum is

The corpus luteum

maintained by luteinizing hormone from the pituitary gland for 10–12 days. It then begins to degenerate, its function regresses, and menstruation begins. The whole process from initial stimulation of the follicle to advance regression of the corpus luteum normally takes approximately 28 days.

During the first part of the menstrual cycle, the inner lining of the uterus, known as the endometrium, grows thicker. In the endometrium the uterine glands, which are exocrine (*i.e.*, do not secrete directly into the bloodstream), become longer but remain relatively straight and secrete little. The endometrium during this stage is said to be in the proliferative phase of the cycle.

After ovulation, the uterine glands become coiled and begin to secrete mucus. The endometrium is said to be in the progestational stage of the cycle. Its glycogen content increases, and it becomes more succulent and edematous. If fertilization and implantation of the fertilized ovum in the uterus do not occur, the endometrium begins to break down as the corpus luteum regresses. Bleeding occurs, and the endometrium is sloughed except for a thin layer near the uterine muscle. The blood and sloughed endometrial tissue constitute the menstrual flow. Menstruation stops after an average of five days because the spiral arteries that supply the endometrium constrict and cut off the blood flow.

Ovarian hormones. The ovaries secrete three hormones: estradiol, progesterone, and relaxin.

Estradiol. Estradiol is an estrogen (estrogen is a generic term for the feminizing, or female, hormones) secreted

by the inner lining, or theca interna, of the follicle, and by the luteal cells. It is a steroid that is in equilibrium in the body with two closely related estrogens, estrone and estriol, and is synthesized from cholesterol in the body by way of testosterone. Most synthetic estrogens have similar structures, although some, such as stilbestrol, do not. Estrogens are responsible for "femininity" and for the growth of the endometrium during the proliferative phase of the menstrual cycle; along with other hormones, they also promote breast development, affect follicle-stimulating and luteinizing hormone secretion in a feedback fashion; and, in many laboratory mammals, act on the brain to stimulate sexual behaviour. Secreted estrogens are about 30 percent bound to protein in the circulation. Estradiol, estrone, and estriol are all excreted in the urine.

Progesterone. Progesterone is formed by the luteal cells of the corpus luteum. A small amount is also secreted by the adrenal gland in both men and women, and it is an important intermediate in the biosynthesis of other steroids. Further details of its biosynthesis are considered in the section on the adrenal cortex. Circulating progesterone is largely bound to albumin and to a globulin. It has a half-life in the circulation of only five minutes and is metabolized, principally in the liver, to pregnanediol, which in turn is conjugated to glucuronic acid and excreted.

Progesterone actions

The main action of progesterone is on the uterus, where, in concert with estrogens, it causes the previously discussed changes in the endometrium during the latter half of the menstrual cycle. Progesterone stimulates breast enlargement by promoting the development of the milk sacs. It is also thermogenic (*i.e.*, raises body temperature) and is responsible for the temperature elevation that occurs immediately after ovulation, a clinical sign that is obviously of value in family planning.

Progesterone also inhibits luteinizing hormone secretion and, when administered throughout the cycle, can prevent the increase in LH secretion that normally produces ovulation. Many derivatives of progesterone have been used for this purpose, and several contraceptive pills are a mixture of a synthetic progestational agent and a small amount of estrogen.

Relaxin. Relaxin is a polypeptide or a group of polypeptides apparently secreted by the corpus luteum. Relaxin softens the pelvic ligaments, permitting the birth canal to enlarge. It is presumably secreted at the time of delivery, but the mechanisms of its formation, secretion, and metabolism are as yet unresolved.

The cyclic pattern of reproductive functioning. The secretion of estrogen and progesterone and their interaction with the pituitary and the hypothalamus to produce the cyclic pattern of reproductive functioning in women is complex.

In the absence of fertilization. Estrogen secretion increases sharply just before ovulation, and there is a second rise in estrogen secretion in the middle of the progestational part of the cycle. Progesterone secretion is low throughout the first half of the cycle and increases only after ovulation. The secretion of estrogen from the developing follicle and the corpus luteum is stimulated by follicle-stimulating hormone in the presence of luteinizing hormone, and estrogen feeds back to inhibit FSH secretion. Progesterone secretion by the corpus luteum is dependent upon LH from the anterior lobe of the pituitary gland, and progesterone inhibits LH secretion. A sharp rise in estrogen levels can also stimulate LH secretion. The FSH level in the blood is slightly elevated early in the menstrual cycle and falls off while the LH level stays constant until midcycle. About the 14th day of the cycle, there is a sharp increase in the circulating levels of LH and FSH, the former being greater than the latter. Thereafter the levels of both hormones fall, reaching their lowest points during the cycle. The rise in LH secretion produces ovulation, while the significance of the rise in FSH secretion is still uncertain. (A more detailed description of the cycle is given in the article MENSTRUATION.)

The secretion of follicle-stimulating hormone and luteinizing hormone is controlled by the hypothalamus through follicle-stimulating hormone releasing factor and luteinizing hormone releasing factor. There is considerable experimental evidence from female animals for the existence of two brain centres to regulate gonadotropic hormone secretion. One centre in the base of the hypothalamus maintains basal levels of FSH and LH secretion, while the other in the superior and anterior hypothalamus is responsible for the bursts of luteinizing hormone secretion that produce ovulation.

Hypothalamic control

It is possible to advance a hypothesis of the way steroids and gonadotropic hormones interact to produce the menstrual cycle: the slightly elevated follicle-stimulating and luteinizing hormone levels early in the cycle lead to follicular maturation, and as the follicle matures there is a sharp increase in estrogen secretion. This estrogen peak in the relative absence of progesterone stimulates the secretion of LH and presumably FSH by positive feedback, with the production of ovulation. The corpus luteum then secretes progesterone and estrogen, and these hormones acting together depress by negative feedback the levels of FSH and LH secretion. When the corpus luteum regresses, estrogen and progesterone secretion declines, and the rise in FSH and LH secretion initiates a new cycle. The hypothesis is, however, admittedly tenuous.

The preceding paragraphs have detailed the endocrine events occurring in the menstrual cycle in the absence of pregnancy. If the ovulated ovum becomes fertilized, the events are different (see PREGNANCY; EMBRYOLOGY, HUMAN), primarily because of hormones secreted by the placenta.

Fertilization and pregnancy. Fertilization usually occurs in the midportion of the fallopian tube. The fertilized ovum, which begins to divide at once, enters the uterus and implants in the endometrium, usually on the posterior superior wall of the uterus. The developing embryo, now called a blastocyst, becomes surrounded by the chorion, a tissue made up largely of syncytiotrophoblasts. The syncytiotrophoblast is comprised of masses of multinucleate protoplasm without clear cell boundaries and is capable of eroding the endometrial tissue to make room for the developing fetus. The outside layer of cells of the blastocyst becomes the trophoblast. Later in pregnancy, the trophoblast becomes organized into the mature placenta. By this time, the syncytiotrophoblast is located on the surface of the villi and is underlain by the cytotrophoblast, a layer of trophoblast cells with distinct boundaries that divide and add to the mass of syncytiotrophoblast by fusing with it (see also EMBRYOLOGY, HUMAN).

The secretions of the placenta. Early in pregnancy, the placenta secretes large amounts of a gonadotropic hormone called human chorionic gonadotropin (HCG). The action of this hormone maintains the corpus luteum, which fails to regress and becomes known as the corpus luteum of pregnancy. No menstruation occurs because progesterone and estrogen secretion from the corpus luteum remains high, and no new follicles develop because these high levels of estrogen and progesterone inhibit the secretion of follicle-stimulating and luteinizing hormone from the pregnant woman's own pituitary.

Human chorionic gonadotropin

Human chorionic gonadotropin is secreted in greatest quantities in early pregnancy, but appreciable amounts continue to be secreted until delivery. The placenta also secretes progesterone and estrogen in increasing amounts throughout pregnancy. Ovariectomy before the third month of pregnancy causes abortion, whereas it has no effect later because the placental hormones maintain the pregnancy. In addition, the placenta secretes renin (a kidney hormone, discussed later) and human chorionic somatomammotropin (HCS), a protein hormone with lactogenic (milk-stimulating) and growth-promoting activity. Thus the placenta is a remarkable endocrine organ in its own right. The performance of the placenta is all the more amazing in view of the fact that all of the placental hormones seem to be secreted by the amorphous syncytiotrophoblast.

Lactation. After delivery of the baby, lactation begins. Breast development, both at puberty and during preg-

nancy, appears to depend on the balanced action of several hormones, but estrogens and progesterone play the principal roles. Prolactin (luteotropic hormone of the anterior pituitary gland) causes the developed breast to secrete milk, and the secretion is apparently increased following delivery, especially after suckling. The mechanism responsible for holding lactation in check during pregnancy and for initiating it after the baby is born, however, is poorly understood.

THE THYROID GLAND

The thyroid gland plays a key role in the regulation of metabolism (see also METABOLISM). It secretes two similar hormones, thyroxine and triiodothyronine. These hormones increase oxygen consumption by most of the cells in the body. In man and other mammals, the thyroid also secretes calcitonin, a plasma-calcium-lowering hormone.

Thyroid morphology

The thyroid gland, located just below the larynx at the base of the neck, is made up of two lobes of tissue connected by a narrow isthmus. It is derived embryologically from the floor of the pharynx, and a remnant of tissue, the thyroglossal duct, sometimes extends from the thyroid to the base of the tongue in the adult. The gland is supplied by two main pairs of arteries, one pair from the external carotid arteries in the neck to the superior poles of the gland and one from the subclavian arteries in the chest to the inferior poles. The thyroid veins drain into the neighbouring large veins of the neck and chest. Like all other endocrine glands, the thyroid is a vascular tissue and has a copious blood supply. Sympathetic nerve fibres pass from the cervical sympathetic ganglia to the gland, and parasympathetic fibres enter it from the vagus nerve. Afferent nerves that relay sensory information from the thyroid to the central nervous system enter the laryngeal nerves. It is probable, however, that the only function of thyroid innervation is regulation of its blood flow.

Thyroid tissue is organized into follicles in which balls of thyroid cells surround a central cavity. The follicles contain a gelatinous pink material, the colloid, in which thyroid hormones are stored bound to a protein called thyroglobulin. The structure of these cells varies with their state of activity. When the gland is inactive for a long time, the cells become flattened; when it is very active, they become square or even rectangular (columnar). Under the electron microscope, these cells show structural details which help to explain their functions. They are underlain by a basement membrane but are in close contact with the capillary blood vessels. The capillaries, like those of other endocrine glands and some other tissues, have gaps or fenestrae between the endothelial cells that make up their walls. At these fenestrae, there is only a thin membrane between the interstitial fluid and the blood, and the fenestrae are presumably the sites of easy transfer of substances into and out of the capillaries. The inner borders of the cells have multiple fine extensions, called microvilli, which project into the colloid.

The hormones of the thyroid. The hormones secreted by the thyroid gland include thyroxine, an amino acid containing four iodine atoms, and its triiodinated derivative, triiodothyronine. The raw materials necessary to form these hormones are iodine and tyrosine, an amino acid normally found in the diet. Their synthesis occurs at the border between the cells and the colloid by successive iodination and condensation of tyrosine residues attached to the thyroglobulin in the colloid. The iodine is actively pumped out of the body fluids into the thyroid cells, and it and the thyroglobulin are secreted into the colloid by the thyroid cells. In addition the thyroid cells remove the hormone from the thyroglobulin and secrete it into the blood stream.

Calcitonin

The thyroid also contains parafollicular or clear cells, which secrete the calcium-lowering hormone calcitonin. Most of these cells are found, as their name implies, in the spaces between the follicles. The parafollicular cells in mammals are homologous to the cells of the ultimobranchial bodies, which are separate endocrine organs in nonmammalian vertebrates. The function of calcitonin is discussed in detail below in the section on the parathyroid glands.

Iodine transport and hormone synthesis. In the body, inorganic iodine circulates in the form of I^-, the iodide ion. In the thyroid, iodine is actively transported from an area of low concentration (the extracellular fluid) to one of high concentration (the interior of the thyroid cell). It is also transported against an electrical gradient. Since the interior of the cell is 50–90 millivolts negative to the exterior, and the iodide ion is negatively charged, the negative ion is moving toward, instead of away from, negativity. The energy for this active transport comes from the chemical adenosine triphosphate (ATP) and related energy-rich phosphate compounds. The transport mechanism is also called the "iodide pump" and the "iodide-concentrating" or "trapping" mechanism. Once the iodide is in the cells, it normally is promptly converted to iodine and bound to the tyrosine residues on the thyroglobulin in the colloid. If this binding is prevented by a drug such as thiouracil, the avidity of the transport mechanism can be seen. The normal ratio of thyroid to plasma iodide is approximately 25:1, but in the presence of thiouracil it may reach 250:1.

Because iodide is concentrated and bound in the thyroid gland, radioactive iodine can be conveniently used for the diagnosis and treatment of certain thyroid diseases. Radioactive iodide is handled by the gland in the same way as other iodides. The magnitude of iodide uptake and binding, therefore, can be quantitated in patients suspected of having thyroid disease by administering an isotope of iodine, then monitoring thyroid radioactivity with a gamma radiation detector placed over the thyroid gland. An uptake curve is plotted and compared with a typical curve for iodine uptake in a normal human. The minute amounts of radioactivity used for diagnostic purposes have no known significant effect on thyroid function, but when large doses of radioactive iodine are administered, radioactivity concentrations in the gland may become sufficient to damage and even kill the thyroid cells. Thus radioactive iodine is of value in the treatment of some cases of hyperthyroidism and thyroid cancer.

As previously mentioned, the thyroid hormones are formed in the colloid by the progressive and orderly addition of iodine atoms to tyrosine residues on the molecules of thyroglobulin in the colloid. This reaction is followed by condensation of two of the iodinated residues to form thyroxine or triiodothyronine. Iodinated products other than thyroxine and triiodothyronine are also formed, but they are de-iodinated by enzymes in the thyroid tissue and their iodine re-used.

In addition to transporting the raw materials for thyroid hormone synthesis to the colloid, the thyroid cells remove triiodothyronine and thyroxine from the colloid and secrete it into the bloodstream. The removal process involves ingestion of bits of colloid by pinocytosis. Inside the cells, the thyroxine and triiodothyronine are detached from the thyroglobulin, presumably through the action of enzymes that are stored within the cells in structures called lysosomes. The hormones then move into the capillaries, apparently by simple diffusion.

Circulating thyroxine bound to protein

In the circulation, thyroxine is bound to plasma proteins. Triiodothyronine is also bound but to a lesser extent. Free thyroxine and triiodothyronine are physiologically active while the bound forms are not. Most of the thyroxine is bound to a circulating globulin called thyroxine-binding globulin, or TBG. Some is also bound to another protein, thyroxine-binding prealbumin, TBPA. In all, more than 99 percent of the thyroxine in the circulation is bound. The total bound and free thyroxine, however, are in equilibrium, and as free thyroxine is removed from the circulation by processes such as entry into cells, it is promptly replaced by thyroxine from the bound pool. Thus the large reservoir of bound thyroxine apparently serves to stabilize the free thyroxine level.

The thyroid hormones are metabolized by removal of iodine (de-iodination) and by combining with glucuronic acid or sulfate (conjugation). The former process occurs in all tissues, while the latter occurs principally in the liver and the kidneys. The half-life of thyroxine is comparatively long (six to seven days); the half-life of triiodothyronine is much shorter

Effects of thyroid hormones. The thyroid hormones have a number of different effects, but their primary action in adults is a stimulation of the oxygen consumption of cells. The stimulation occurs in most organs in the body, the only prominent exceptions being the brain, spleen, and testes. Secondary to the stimulation of metabolism, there is increased appetite, weight loss, often a slight rise in body temperature, and various other effects. The hormones also lower circulating cholesterol levels, an effect for which they have been used therapeutically. They are necessary for normal growth and development of children, and their deficiency is associated with dwarfism and mental retardation. The hormones appear to exert their effects primarily via DNA and RNA. Details of the actions of these hormones are considered in the articles HORMONE; ENDOCRINE SYSTEM DISEASES AND DISORDERS. The effect of a single dose of thyroxine on oxygen consumption is slow in onset and prolonged, lasting many days. The effect of triiodothyromine is much more rapid and pronounced, possibly because this hormone is less firmly bound to plasma protein than thyroxine and has a much shorter half-life.

Control of thyroid secretion

The secretion of thyroxine and triiodothyronine is under the control of thyroid-stimulating hormone from the anterior lobe of the pituitary gland. In the absence of pituitary, the thyroid continues to function, but at a low level. Conversely, injection of TSH causes a pronounced stimulation of thyroid function and increased secretion of thyroxine and triiodothyronine. Its exact site of action remains uncertain; it seems to stimulate almost all aspects of thyroid function, including iodide transport, oxidation to iodine, and iodine binding to tyrosine, as well as ingestion of colloid by the cells. The key to all these actions, however, appears to be stimulation by TSH of the formation of cyclic adenosine monophosphate in the thyroid. How this reaction leads to the others is unknown.

Thyroid-stimulating hormone secretion is controlled in part by the hypothalamus by way of the thyrotropin releasing factor secreted into the portal hypophysial vessels. Free thyroxine and triiodothyronine feed back to inhibit TSH secretion. The main site at which the thyroid hormones exert their inhibitory effect is the anterior pituitary itself. There may be an additional negative feedback of thyroid hormones at the hypothalamic level as well. It is still convenient, and probably accurate, to visualize the regulation of TSH secretion in terms of a closed-loop negative feedback system that maintains a stable day-to-day level of thyroid function and an open loop mechanism, via the nervous system and thyrotropin releasing factor, which can override the feedback control. One condition which probably works via this override system is environmental temperature; exposure to a cold environment promptly increases TSH secretion, at least in infants, whereas exposure to warmth decreases it. Psychological stimuli may also effect TSH secretion via the hypothalamus, but current evidence indicates that such stimuli do not exert major effects on thyroid function.

THE ADRENAL GLANDS

Man has two adrenal, or suprarenal, glands, one at the upper end of each kidney. Each gland is made up of an inner medulla, which secretes the hormones epinephrine and norepinephrine, and an outer cortex, which secretes steroid hormones. The cortex, in turn, is made up of an outer zona glomerulosa, which secretes the salt-regulating steroid aldosterone; and an inner zona fasciculata and zona reticularis, which secrete the other adrenal cortical hormones.

Structure. The cortical cells contain considerable quantities of lipids. The cells of the zona glomerulosa are arranged in balls, while the zona fasciculata is made up of straight cords of cells, and the zona reticularis is made up of cells arranged in an irregular net. Under the electron microscope, the adrenocortical cells are seen to contain the structures typical of steroid-secreting cells. There is a prominent smooth endoplasmic reticulum.

The medulla contains granulated secretory cells arranged in short cords or clumps. These cells, like certain other cells in the body, stain intensely with potassium bichromate and hence are called chromaffin cells. They contain dense-core granules surrounded by membranes. These granules contain epinephrine or norepinephrine bound to ATP and a protein. Morphologically, they resemble protein-secreting cells. There appear to be separate norepinephrine-secreting and epinephrine-secreting cells, the norepinephrine-secreting cells containing granules that are more dense than those of the epinephrine-secreting cells.

Blood supply of the adrenals

The adrenal glands are supplied by a large number of arterial twigs from the inferior phrenic arteries, the aorta, and the renal arteries, the major arteries in the vicinity. For the most part, these arteries form a plexus, or network, on the surface of the gland, and from it, broad, thin-walled sinusoids drain blood through the three zones of the cortex to the medulla. The capillaries of the medulla drain to a central vein. This vein drains into the vena cava on the right and the left renal vein on the left. Some arteries from the capsular plexus pass directly to the medulla. The adrenal medulla thus has a dual blood supply; some blood reaches it directly via the arteries from the capsule, but much of the blood that reaches it has first passed through the cortex. This arrangement may have functional significance; the activity of the enzyme responsible for the formation of epinephrine is increased by high concentrations of adrenal cortical hormones (see below), and it has been argued that formation of epinephrine depends on the medulla being perfused with blood from the cortex.

Innervation of the adrenal medulla

The adrenal medulla is formed from primitive neural tissue in the embryo and is densely innervated. The adrenal medullary cells, homologous to the nerve cells of the sympathetic chain, are also innervated by nerve cells that have their cell bodies in the central nervous system (preganglionic neurons). In the sympathetic ganglia, these neurons innervate postganglionic neurons that in turn innervate the viscera, while in the adrenal medulla they innervate cells that secrete directly into the bloodstream. The preganglionic neurons in both instances are cholinergic; *i.e.*, the chemical mediator liberated by the nerve endings is acetylcholine. The nerve fibres that innervate the adrenal medulla have their cell bodies primarily in the lateral gray column of the mid- and lower thoracic segments of the spinal cord and reach the adrenal via splanchnic nerves and the celiac and renal sympathetic plexuses.

Adrenal cortical hormones. The adrenal cortical hormones are steroids; *i.e.*, compounds that, as noted in the section on the testes, contain the cyclopentanoperhydrophenanthrene nucleus. They are conveniently divided on a functional basis into: (1) mineralocorticoids, hormones that have their major action on the excretion of sodium and potassium; (2) glucocorticoids, hormones that affect the metabolism of glucose, other carbohydrates, and protein; and (3) sex hormones, including both androgens and estrogens, that have the same functions as the gonadal steroids discussed previously. The mineralocorticoids and glucocorticoids are carbon-21 steroids (with a side chain on carbon atom 21), while the sex hormones are mostly carbon-19 steroids.

The adrenocortical hormones are synthesized from cholesterol, which in turn is formed from acetate. The cholesterol is converted to pregnenolone. The adrenocortical hormones are then formed by rearrangement of the A ring of the steroid nucleus and the addition of hydroxyl (−OH) groups in various positions. Each step is catalyzed by a particular enzyme or enzyme complex. By way of the so-called mineralocorticoid pathway, pregnenolone is converted to corticosterone and then to aldosterone. Pregnenolone and progesterone are also converted to their 17-hydroxy derivatives by the addition of an −OH group in the 17 position (17-α-hydroxylation). Via the 17-hydroxycorticoid pathway, they are then converted to cortisol, or, alternatively, the side chain in position 17 is split off and the androgens dehydroepiandrosterone and Δ-4-androstene-3,17-dione are formed. Appreciable quantities of these androgens are secreted, but

they are also converted to testosterone, and some of the testosterone appears to be converted to estradiol.

The enzymes involved in steroid biosynthesis are of considerable interest because different disease states caused by isolated congenital deficiency of each of the enzymes have now been described. In addition, drugs have been discovered that can inhibit individual enzymes in the biosynthetic scheme.

Mineralocorticoids. Hormones with mineralocorticoid activity increase the reabsorption of sodium by the kidney and decrease the net reabsorption of potassium so that sodium is retained in the body and potassium is lost from it. Mineralocorticoids also affect sodium and potassium metabolism through actions on other organs, but their principal effect is on the kidney.

Aldosterone is the most active mineralocorticoid secreted by the adrenal cortex. Corticosterone and deoxycorticosterone have appreciable mineralocorticoid activity, but when secreted in normal amounts, they contribute relatively little to the regulation of sodium and potassium balance. Thus aldosterone, even though its normal secretion rate is only 50–200 micrograms (millionths of a gram) per day, is the principal mineralocorticoid.

About 60 percent of the secreted aldosterone is bound to circulating proteins, mostly to albumin. It has a biological half-life of approximately 35 minutes and is metabolized in the liver and kidneys, principally by conjugation (in the 3 or 18 position) to glucuronic acid. The glucuronic acid derivatives are then excreted in the urine.

Secretion of aldosterone

The secretion of aldosterone is affected by adrenocorticotropin from the anterior lobe of the pituitary gland but is also regulated by angiotensin II, a peptide formed in the bloodstream, and by the plasma levels of sodium and potassium. Angiotensin II formation is in turn regulated by the rate at which renin is secreted by the kidneys (described later). ACTH has a marked effect on glucocorticoid secretion, but relatively large amounts are normally required to increase aldosterone secretion. On the other hand, small amounts of angiotensin increase aldosterone secretion. Larger amounts produce significant increases in glucocorticoid output as well, but even extremely large doses do not increase glucocorticoid output as much as ACTH does. A rise in plasma potassium or a fall in plasma sodium produces an increase in aldosterone secretion through a direct effect on the adrenal cortex.

Table 2: Major Stimuli That Alter Aldosterone Secretion*

	ACTH secretion	renin secretion	plasma potassium	plasma sodium
Surgery, physical trauma	+	NC or +†	NC	NC
Anxiety	+	NC or +†	NC	NC
Hemorrhage	+	+	NC	NC
Low sodium intake	NC	+	NC or +‡	NC or −‡
Standing	NC	+	NC	NC
Secondary hyperaldosteronism (some cases of heart failure, liver cirrhosis, nephrosis)	NC	+	NC	NC
High potassium intake	NC	NC or −	+	NC

*+ = increase; − = decrease; NC = no change. †Due to increased sympathetic activity. ‡Plasma sodium and potassium change only when sodium restriction is prolonged.

The major stimuli known to increase aldosterone secretion are summarized in Table 2. Stimuli, such as physical or mental stress, that increase aldosterone secretion by way of ACTH also increase glucocorticoid secretion. The increase in aldosterone secretion produced by ACTH is temporary; aldosterone output subsequently declines even if ACTH secretion remains elevated. Other factors stimulate aldosterone secretion selectively without increasing ACTH output. The increases produced by these stimuli are mediated, for the most part, via the renin-angiotensin system, although the increase produced by a high potassium diet appears to be direct. Stimuli that probably act by way of the renin-angiotensin system include a low sodium intake, bleeding, standing, and various disease states (such as nephrosis and cirrhosis) that increase aldosterone secretion. The way in which a low-sodium diet stimulates renin secretion is unknown. Other stimuli may act by decreasing blood pressure in the kidney. The aldosterone in turn causes sodium retention. Water is retained with the sodium, extracellular fluid volume is expanded, and blood pressure rises, shutting off the stimulus to aldosterone secretion. This feedback control mechanism by way of extracellular fluid volume is probably significant in the regulation of aldosterone secretion.

The role of plasma sodium and potassium, on the other hand, is less important, at least in the day-by-day regulation of aldosterone secretion. In experimental animals, relatively large changes in plasma sodium and potassium are required to produce a rapid change in aldosterone secretion. Although the increase in aldosterone brought about by a high potassium diet is probably due to a direct action of potassium on the adrenal, the plasma sodium and potassium mechanism appears to be primarily a homeostatic "backup" not normally involved in a major way in the regulation of aldosterone secretion.

There is some evidence that as yet uncharacterized factors, in addition to ACTH, the renin-angiotensin system, and plasma sodium and potassium, also affect aldosterone secretion.

Glucocorticoids. Glucocorticoids influence glucose metabolism principally by (1) exerting an anti-insulin effect, and (2) increasing the formation of glucose from amino acids and other noncarbohydrate precursors in the liver, a process known as gluconeogenesis. They also cause extensive protein breakdown, however, and have widespread additional effects in the body. In large amounts, but probably not in quantities normally secreted, they can counteract allergic and other types of inflammatory reactions. It is on this basis they have been used in the treatment of asthma, rheumatoid arthritis, and various skin diseases.

Cortisol

The principal glucocorticoid in man is cortisol. Corticosterone also has appreciable glucocorticoid activity, but normally there is less of it in the circulation. The other secreted hormones account for only a small portion of the glucocorticoid activity in the body fluids. Cortisol and corticosterone are secreted in much larger quantities than aldosterone. Under resting conditions in normal adult humans, the 24-hour cortisol output is about 20 milligrams (thousandths of a gram) and the corticosterone output about three milligrams. Cortisol is bound in the circulation to a plasma protein called transcortin, or corticosteroid-binding globulin, which is synthesized in the liver. A small amount is also bound to plasma albumin. The normal morning value for circulating cortisol is approximately 12 micrograms per 100 millilitres (thousandths of a litre) of plasma. Approximately 96 percent of this is bound and only 4 percent free. Corticosterone is similarly bound, but not to as great a degree; the normal morning value for corticosterone is 0.6 microgram per 100 millilitres of plasma, and only 90 percent of this is bound. Bound hormones are not as available for metabolism, so the half-life of cortisol in the circulation is 60–90 minutes, while the half-life of corticosterone is approximately 50 minutes.

The amount of transcortin in the circulation obviously affects the equilibrium between free and bound cortisol and corticosterone. If the amount of circulating transcortin increases, more glucocorticoids are bound. The bound glucocorticoids, unlike the free, appear to be physiologically inactive. The decrease in free glucocorticoids, however, stimulates adrenocorticotropin secretion, causing more glucocorticoids to be secreted, and a new equilibrium is reached in which the free glucocorticoid level is restored to normal.

Cortisol is removed from the circulation primarily in the liver. Some of the cortisol is reduced to form tetrahydrocortisol, which is then conjugated to glucuronic acid to form tetrahydrocortisol glucuronide. Some of the cortisol is converted to 17-ketosteroids, and some is converted to cortisone. The cortisone is, in turn, reduced and conjugated. Cortisone, then, an active glucocorticoid and well-known for its extensive use in the treatment of allergic and related diseases, is not a physiologically significant secretion of the adrenal cortex but is instead

primarily a liver metabolite of cortisol. The glucuronic acid derivatives of cortisol and cortisone are readily soluble in water, and are excreted in large quantities in the urine. Smaller amounts of other metabolites and of free cortisol are also excreted. The metabolism of corticosterone is similar to that of cortisol, except that it is not converted to cortisone.

One of the body's notable reactions to injury or threatened injury is an increase in glucocorticoid secretion. This increase is of great importance; in its absence even relatively minor injuries can be fatal. The multitude of noxious and potentially noxious stimuli that increase glucocorticoid secretion are generally lumped together under the term "stress." There is in addition a regular diurnal fluctuation in resting levels of glucocorticoid secretion. Both stress-induced and diurnal variations are caused by fluctuations in the secretion of adrenocorticotropin from the pituitary.

Regulation of glucocorticoid secretion

The regulation of glucocorticoid secretion is almost exclusively under the control of ACTH from the pituitary gland, and, although certain other factors including angiotensin II can increase glucocorticoid secretion, it appears that under all but the most extreme conditions, these latter factors play no physiological role. In the absence of stress, adrenocorticotropin is secreted in irregular spurts throughout the day, but the number and size of these spurts is greater in the early morning hours while the subject is still asleep. The circulating glucocorticoid level follows the adrenocorticotropin level, with a general downward trend throughout the waking hours.

The secretion of adrenocorticotropin is inhibited by cortisol and other glucocorticoids from the adrenal, and this negative feedback control provides stability to the system. The site at which the cortisol feeds back to inhibit ACTH secretion is probably the pituitary, but there may be an additional feedback on the hypothalamus or some other part of the brain. The increases in adrenocorticotropin secretion produced by most stimuli are mediated by way of the nervous system. Neural stimuli converge on the ventral portion of the hypothalamus, where they affect the production of the corticotropin releasing factor (CRF). By way of the portal hypophysial vessels the CRF enters the anterior pituitary, where it regulates the secretion of adrenocorticotropin. The neural impulses responsible for the increases in adrenocorticotropin secretion produced by physical stress probably ascend to the hypothalamus through the reticular formation, the ancient core of the brain (see also NERVOUS SYSTEM, HUMAN). Some drugs may act directly on the hypothalamus. Psychological stimuli act via the limbic system, that part of the central nervous system in the temporal lobes of the brain concerned with the genesis and expression of emotion. The "biological clock" responsible for the diurnal variation in adrenocorticotropin secretion also appears to be located in the temporal lobes. It also appears that there are neural pathways that inhibit ACTH secretion. The physiological role of the inhibitory pathways, however, is presently unknown.

Sex hormones. The adrenal sex hormones are normally supplemental to the gonadal sex hormones and have relatively little effect on their own. The adrenal androgens, however, particularly dehydroepiandrosterone, play a role in the growth of pubic and axillary hair in women. Secreted in excess, they can produce marked masculinization. The secretion of the adrenal sex hormones, like that of the glucocorticoids, is apparently controlled primarily by adrenocorticotropin. Gonadotropins do not appear to play any role. The androgens are excreted for the most part as 17-ketosteroids.

Adrenal medullary hormones. The adrenal medullary hormones are norepinephrine and epinephrine. Norepinephrine is formed by decarboxylation and successive hydroxylation of the amino acid tyrosine. Tyrosine is taken up by adrenal medullary cells and converted to dihydoxyphenylalanine (dopa). Dopa is then decarboxylated to the corresponding amine, dopamine. Dopamine, norepinephrine, and epinephrine are all amines (organic compounds structurally related to ammonia), and because of their structure they are frequently referred to collectively as catecholamines (structurally related to phenol alcohols). Dopamine enters the dense-core granules in the cells and is hydroxylated to form norepinephrine. The reactions to this point are identical to those occurring in the norepinephrine-secreting adrenergic neurons in the brain and sympathetic nervous system. In adrenal medullary cells that secrete epinephrine, however, there is an additional reaction. This reaction, the methylation of norepinephrine to form epinephrine, is catalyzed by the enzyme phenethanolamine-N-methyl transferase (PNMT). There is evidence that PNMT formation requires a milieu in which the glucocorticoid concentration is high. Such a milieu is provided only in the adrenal medulla, since as noted above, the medulla is perfused with blood that has first passed through the adrenal cortex. This may explain why appreciable amounts of PNMT activity, and consequently of epinephrine synthesis, are found only in the adrenal medulla.

Synthesis of norepinephrine and epinephrine

The actions of epinephrine and norepinephrine are generally similar, but there are certain important exceptions to this generalization. Both amines increase the force of contraction of the heart, break down fats and thus increase the level of circulating free fatty acids, raise the metabolic rate, and increase alertness by an action on the brain. Norepinephrine, however, causes a rise in blood pressure because it causes constriction of all blood vessels—except possibly the coronary vessels supplying the heart. Epinephrine causes constriction in many vascular beds, but it dilates the blood vessels in the skeletal muscles and the liver. The net effect on the peripheral resistance against which the heart must pump is, therefore, relatively small, and there is little change in mean blood pressure. Epinephrine breaks down the glycogen in the liver (glycogenolysis) with a consequent outpouring of glucose from the liver into the circulation. Norepinephrine has little glycogenolytic activity.

Circulating epinephrine and norepinephrine have short half-lives. They are inactivated primarily by O-methylation to form the inactive compounds metanephrine and normetanephrine. The enzyme catalyzing O-methylation is catechol-O-methyltransferase (COMT). Some of the normetanephrine and metanephrine are then oxidized to 3-methoxy-4-hydroxy mandelic acid (VMA). Intracellular catecholamines, on the other hand, are subject to oxidation before secretion because of the presence in the mitochondria of the enzyme monoamine oxidase (MAO). The inactive oxidation products enter the circulation and are converted to VMA by COMT. Other metabolites are formed in small amounts. About 2 percent of the secreted catecholamines appears unaltered in the urine; about 35 percent is excreted as metanephrine and normetanephrine and 35 percent as VMA.

In humans, epinephrine makes up 80 percent of the total catecholamine secretion of the adrenal medulla. The remaining 20 percent is norepinephrine, although it is possible that traces of dopamine are also secreted. Norepinephrine also reaches the circulation from sympathetic nerve endings, so the normal resting plasma norepinephrine level is about 0.4 nanogram (1 nanogram = 10^{-9} gram) per millilitre while the epinephrine level is 0.1 nanogram per millilitre or less.

Stimuli for medullary secretion

The stimuli that increase adrenal medullary secretion are, in general, those that increase sympathetic discharge. These include most, but not necessarily all, of the stressful stimuli that increase ACTH secretion. The sympathetic discharge, because it prepares the animal to meet emergency situations, is sometimes called the "fight or flight reaction." Among the stimuli that increase adrenal medullary secretion are low blood sugar, oxygen lack, and cold. A decline in blood pressure also initiates increased secretion. Control of the gland is almost completely neural, via the preganglionic nerve fibres that innervate the adrenal medullary cells. The adrenal medulla is one of the few endocrine tissues in which the gland cells are directly innervated by secreto-motor nerve fibres. There is some evidence for selective secretion of either epinephrine or norepinephrine from the adrenal medulla under certain circumstances, but most stimuli provoke a proportionate increase in the secretion of both

hormones. The response to bleeding is probably mediated via the pressure-sensing carotid and aortic receptors in the arteries (see BLOOD CIRCULATION, HUMAN), and the receptors that initiate the responses to low blood sugar, oxygen lack, and cold are probably neural elements in the central nervous system or the peripheral tissues. It is notable that there is only indirect negative feedback control in the adrenal medullary system, and catecholamines do not appear to act directly to inhibit their own secretion.

THE PITUITARY GLAND

The pituitary gland is located deep in the head at the top of the nasal cavity in a hollow, the sella turcica, in the base of the skull. It is connected to the overlying hypothalamus by a stalk. Except where it is pierced by the pituitary stalk, the sella is covered by a fold of dural membrane, the tentorium, or diaphragma sella. The pituitary is formed during embryonic development by the union of a ball of cells that breaks off from the roof of the mouth (Rathke's pouch) with a projection from the base of the brain. This accounts for the formation of an anterior, an intermediate, and a posterior lobe. In a number of mammalian species, however, the intermediate lobe tissue is rudimentary, and in humans it is intermingled with the tissue of the anterior lobe.

The anterior pituitary and its hormones. The anterior pituitary receives few if any nerve fibres from the hypothalamus. In man and some other mammalian species, it also lacks any appreciable direct blood supply; instead, almost all the blood reaching it first perfuses the base of the brain and is transported to the pituitary by the portal hypophysial vessels. These vessels drain a primary network of capillary loops in the median eminence of the hypothalamus and adjacent upper portion of the pituitary stalk. They pass down the stalk to the anterior lobe, where they supply the gland cells. The primary network from which the portal vessels arise is formed by branches from the carotid and other neighbouring arteries. The blood supply to the anterior pituitary is thus a true portal system since it begins and ends in capillaries without going through the heart. The blood from the anterior pituitary drains into the neighbouring cavernous sinus, a venous sinus that, in turn, drains into the jugular veins and thence into the heart.

Morphology of the anterior lobe

The anterior lobe is divided into two parts: (1) the pars tuberalis, the thin layer of tissue on the pituitary stalk through which the portal vessels run; and (2) the pars distalis, the main secretory portion of the gland. The cells of the pars tuberalis are not well differentiated, but in the pars distalis are found various types of secretory cells, including some intermediate lobe cells. The anterior lobe cells generally have been divided, on the basis of their staining characteristics, into acidophils (staining with acid dyes), basophils (staining with basic dyes), and chromophobes (staining poorly with either type of dye). More refined techniques and electron microscopy, however, have shown that, at least in some species, there are six distinct cell types, one for each of the six anterior lobe hormones. This is probably also true in man, although all six cell types have not as yet been absolutely identified.

The anterior lobe of the pituitary secretes six polypeptide or protein hormones; adrenocorticotropic hormone (ACTH), growth hormone (GH), thyroid-stimulating hormone (TSH), follicle-stimulating hormone (FSH), luteinizing hormone (LH), and prolactin. The human anterior pituitary also contains the intermediate lobe hormones, α- and β-melanocyte stimulating hormone (α- and β-MSH). ACTH is a simple polypeptide. Growth hormone and prolactin are simple proteins, while TSH, FSH, and LH are glycoproteins that contain carbohydrates such as fucose. The hormones are synthesized, as are other proteins, in the rough endoplasmic reticulum and stored in membrane-lined granules in the cells. Secretion involves the previously considered process of emeiocytosis —the movement of these granules to the cell membrane; the membrane around the granule then fusing with the cell membrane, and the area of fusion breaking down, leaving the secretory granule on the outside of the cell and the membrane intact.

The tropic hormones. ACTH, TSH, FSH, and LH are tropic hormones in the sense that they act to regulate other endocrine glands, whereas growth hormone and prolactin act directly on non-endocrine tissues. The details of the functions of the tropic hormones are treated in the sections on the particular glands they regulate.

Growth hormone. GH promotes protein synthesis and plays an important role in growth. Dwarfism results from its absence. Of the variety of causes of dwarfism, many are associated with skeletal and other deformities. In congenital isolated growth hormone deficiency, the body proportions are normal except for short stature, and sexual function is preserved. Charles Stratton, the "midget" whom P.T. Barnum, the famed circus entrepreneur, exhibited under the name General Tom Thumb, was in all probability an example of isolated growth hormone deficiency.

Growth hormone is also secreted in adults; in fact, the rate of growth hormone secretion in adults is about as high as it is in children. It exerts an anti-insulin effect and also acts to mobilize free fatty acids from fat depots. Its actions are discussed in more detail in HORMONE; ENDOCRINE SYSTEM DISEASES AND DISORDERS.

Control of GH secretion

Growth hormone secretion is controlled by growth hormone releasing factor (GRF) from the hypothalamus. Resting plasma growth hormone levels are normally zero to three nanograms per millilitre. The half-life of the hormone in the circulation is 20–30 minutes, and it is metabolized, at least in part, in the liver. Growth hormone secretion is increased during sleep, but the cause and the significance of this increase remain obscure. The growth hormone-stimulating action of the pancreatic hormone glucagon may help the body combat a low blood sugar. Other GH-secreting stimuli generally fall into one of three categories: (1) stressful stimuli, (2) stimuli that increase circulating amino acid levels, and (3) conditions such as low blood sugar and fasting in which there is a deficiency, or at least a threatened deficiency, of fuel to be burned to meet the body's energy needs.

Growth hormone has no target endocrine gland, and hence no direct negative feedback control via a target organ hormone. There is evidence, however, that growth hormone feeds back to inhibit its own secretion, presumably by an action at the hypothalamic level. Growth hormone secretion is thus regulated in part by short-loop feedback. The receptors responsive to blood glucose and amino acid levels are presumably in the brain, but their exact location has not been established.

Prolactin. The only proved function of prolactin in mammals is stimulation of the synthesis of milk in the breast. Its function in males, despite a great deal of speculation, is unknown. The action of prolactin on the breast requires the presence of other hormones, including estrogens, progesterone, and adrenocortical hormones. Prolactin secretion is increased after parturition when milk is being formed for the newborn infant. In abnormal states, however, prolactin can also produce milk secretion in virgin women and even in men. In laboratory mammals, prolactin secretion is increased by suckling. There is a sharp increase in circulating prolactin along with the increase in circulating LH and FSH at the time of ovulation, but during the rest of the sexual cycle and during pregnancy, the levels are low. There is no reason to believe that the pattern of prolactin secretion in humans is different, but proof that it is similar must await the development of a method for measuring prolactin in man.

One of the reasons that methods for measurement of prolactin have not been successful is the difficulty in separating human prolactin from human growth hormone. The similarity between the two hormones is so great that it has been postulated that they are identical. The growth hormone-like and prolactin-like activities released by human pituitary cells in tissue culture, however, are not always parallel, and female dwarfs with congenital growth hormone deficiency lactate normally

and have prolactin-like activity in their blood when they nurse their babies. Thus, it appears that the two hormones are separate, although human prolactin has not as yet been isolated or characterized.

Control of prolactin secretion

Prolactin secretion is under inhibitory control by the hypothalamus, and severance of the portal vessels by cutting the pituitary stalk or transplanting the pituitary leads to increased prolactin secretion. Section of the pituitary stalk can bring about lactation in nonpregnant women. A prolactin inhibiting factor (PIF) has been extracted from the hypothalamus, and there is some evidence that there may be a prolactin stimulating factor as well. There is also evidence that prolactin feeds back in a short-loop fashion to inhibit its own secretion.

Intermediate lobe and its hormones. Two intermediate lobe hormones, α-MSH (α-melanocyte-stimulating hormone) and β-MSH, are found in the pituitaries of mammals (see HORMONE). The first is a polypeptide containing 13 amino acid residues that has the same structure in a variety of mammals and probably in man. The structure of β-MSH varies from species to species; in humans, it is a polypeptide containing 22 amino acid residues. β-MSH is found in human blood, but α-MSH has not been detected in any of the body fluids. Little is known about the metabolism and excretion of these compounds.

Functions of the intermediate lobe hormones

Intermediate lobe hormones stimulate melanophores, cells that contain motile granules in the skin of reptiles, fish, and amphibia. When the granules are aggregated in the centre of these cells, they are relatively clear, and the skin of the animal assumes a light colour; when the granules spread throughout the cells, the skin becomes dark. In many animals, the secretion of intermediate lobe hormone plays an important role in protective changes in skin coloration. There are no melanophores, however, in the skin of mammals. Melanocyte-stimulating hormones can produce some darkening of the human skin by increasing the formation of dark pigments, but this does not occur under normal conditions. Many other functions have been hypothesized for α- and β-MSH in humans, but there is insufficient evidence to consider any of them as established. It is known that the β-MSH level in the circulation tends to parallel that of ACTH, being high when cortisol secretion is low and low following administration of adrenocortical hormones. ACTH has some MSH activity because it contains a sequence of amino acids that is also found in the latter. In certain diseases, high levels of β-MSH in the blood are associated with increased pigmentation of the skin. These are also diseases in which the rate of ACTH secretion is high, but the ACTH is not present in sufficient quantities to produce the pigmentation by itself.

In some animal species, the intermediate lobe appears to be under direct neural control via nerve fibres from the hypothalamus that inhibit its secretion. The existence in the hypothalamus of two factors that influence MSH secretion has been claimed—one an MSH-inhibiting factor (MIF) and the other an MSH-releasing factor (MRF), but the role that these factors play, if any, in the regulation of MSH secretion is uncertain.

The posterior pituitary and its hormones. The posterior lobe of the pituitary gland secretes two hormones: vasopressin and oxytocin. Each is a small polypeptide.

The posterior pituitary is embryologically and functionally a part of the brain (many of the neurological structures considered here are described in the article NERVOUS SYSTEM, HUMAN). It is made up primarily of nerve fibres that end around blood vessels. These nerve fibres are the axons of nerve cells that have their cell bodies in two prominent nuclei of the hypothalamus, the supraoptic and paraventricular nuclei. The blood vessels that supply the posterior lobe come from nearby arteries, and the venous drainage, like that of the anterior lobe, is into the nearby cavernous sinus.

The posterior lobe hormones are formed in the neurons of the supraoptic and paraventricular nuclei and released from the endings of their axons into the capillaries of the posterior lobe. These capillaries have fenestrated endothelial linings. Release is triggered by nerve impulses that travel along these neurons to their endings. The axons from the supraoptic nuclei, whose principal secretion is vasopressin, end mainly in the posterior lobe; the axons from the paraventricular nuclei, which secrete mainly oxytocin, have an appreciable number of endings in the stalk and median eminence as well. Under the electron microscope, the endings of the neurons are found to contain many granules that are 1000–1500 angstrom (1 angstrom = 10^{-10} metre) units in diameter. These granules are probably collections of oxytocin or vasopressin bound to several small proteins called neurophysins. It is still uncertain how the hormones pass from the nerve endings into the blood vessels, although it may be by emeiocytosis or by dissolution of the granules.

Herring bodies

The supraoptic and paraventricular neurons contain globules of a material that takes on certain specific stains (*e.g.*, the Gomori stain). In mammals, they are the only neurons that stain in this fashion. The stained globules, called Herring bodies, are apparently aggregations of carrier-bound posterior pituitary hormones, and the material is known as "neurosecretory." If the axons of the supraoptic and paraventricular nuclei are cut, neurosecretory material disappears beyond the cut and accumulates just proximal to it. Thus it appears that oxytocin and vasopressin are manufactured in the cell bodies of the neurons and transported along the axons to their endings. Most, if not all, neurons secrete chemical agents at their axon terminals, and there is evidence that many of these secreted chemical agents are synthesized in part in the cell bodies of the neurons and carried along their axons to the endings.

Vasopressin. Vasopressin is the antidiuretic hormone. It acts primarily on the walls of the terminal portions of the kidney tubules, causing an increase in water reabsorption. This makes the urine more concentrated and consequently dilutes the body fluids. Vasopressin plays an important role in maintaining the body's water and electrolyte balance. The hormone acts by way of cyclic adenosine monophosphate. It has a half-life in the circulation of only three minutes and is metabolized in the liver and kidneys.

Oxytocin. Oxytocin acts on muscle-like cells (myoepithelial cells) in the walls of the ducts of the breast. Contraction of these cells squeezes milk out of the ducts and consequently out of the nipples to the exterior. Oxytocin also induces contractions of the uterine muscle, but this action is strong only in the estrogen-dominated uterus; *i.e.*, when the estrogen blood-level is high. Oxytocin aids in parturition and may also assist the movement of sperm up the female genital tract after sexual intercourse. The half-life of oxytocin in the circulation is approximately the same as that of vasopressin, and it is metabolized for the most part in the kidneys.

The regulation of vasopressin and oxytocin secretion is considered in the paragraphs that follow.

THE STRUCTURE AND ENDOCRINE FUNCTIONS OF THE HYPOTHALAMUS

The hypothalamus is not only a major regulator of endocrine secretion, it is also an endocrine gland in its own right, as its nerve cells secrete blood-borne factors that regulate anterior pituitary activity, water balance, and milk secretion. It forms the anterior end of the stem of the brain. It is bounded on each side by the temporal lobes of the cerebral cortex, above by the thalamus, and behind by the midbrain. The bottom portion of the hypothalamus is called the tuber cinereum, and its midportion, from which the portal hypophysial vessels originate, is the median eminence of the tuber cinereum.

Morphology of the hypothalamus

The hypothalamus contains many nuclei that regulate a variety of endocrine and vegetative, or involuntary, functions. The supraoptic and paraventricular nuclei, the large-celled nuclei that regulate posterior pituitary secretion, are mentioned above. The axons of the cells in these nuclei pass to the median eminence and posterior pituitary, forming the hypothalamo-pituitary tract. In the median eminence are many nerve endings closely associated with the primary plexus of the portal hypophysial vessels. It is probable that most of these endings secrete the releasing and inhibiting factors that regulate

anterior pituitary secretion. The major factors are summarized in Table 1. There may also be a growth hormone inhibiting factor (GHIF) and a prolactin stimulating factor (PSF). Thyrotropin releasing factor (TRF) has been isolated and synthesized; it is a tripeptide containing three amino acid residues. The other factors are probably polypeptides, but their exact structure has not been determined. At least some of them appear to act via cyclic AMP, but the details of their actions and metabolism are unknown.

There are few cell bodies in the median eminence; therefore, most of the neurons that secrete the factors affecting anterior pituitary function must have their cell bodies located elsewhere. The neurons concerned with gonadotropic hormone secretion appear, for the most part, to have their cell bodies in the arcuate nucleus near the ventromedial nucleus. The neurons concerned with the regulation of thyroid-stimulating hormone secretion have their cell bodies in the anterior hypothalamus. The locations of the cell bodies of the other neurons that secrete releasing and inhibiting factors are unknown.

The regulation of growth hormone and vasopressin secretion

The control of growth hormone secretion is an example of the regulation of anterior pituitary secretion mediated by the hypothalamus. The major factors that increase the secretion of this hormone are energy deficiencies, stressful stimuli, and certain amino acids. Presumably these stimuli all act on the nervous system in different ways, but the net effect of them all is stimulation of the neurons that secrete growth hormone releasing factor (GRF). This factor is liberated from the endings of these neurons into the portal vessels and passes by this route to the anterior pituitary where it stimulates release of growth hormone.

The control of the secretion of vasopressin is an example of the type of neuroendocrine reflex that controls posterior pituitary secretion. Vasopressin secretion is increased when the osmolality, or osmotic pressure, of the blood plasma increases; *i.e.*, when plasma becomes more concentrated. It is also increased by a decrease in the circulating blood volume, by pain, and by certain drugs.

Osmotic control is mediated through cells in the hypothalamus that are sensitive to changes in the osmolality of the body fluids. When the pressure increases, these cells bring about an increase in the rate of discharge of the supraoptic neurons and the amount of vasopressin released from their endings is increased. The vasopressin acts on the kidneys, causing water retention and thus diluting the body fluids back toward their original osmolality. With the return of normal osmotic pressure the stimulus to increased vasopressin secretion is abolished. Conversely, a fall in plasma osmolality decreases the discharge rate in the supraoptic neurons; less vasopressin is secreted; less water is reabsorbed in the kidney; and water is lost in the urine until the plasma osmolality is again restored to normal. Thus the osmoreceptor-vasopressin mechanism is a feedback mechanism, which, in the absence of complicating factors, constantly operates to maintain the plasma osmolality at a normal level.

Vasopressin secretion is also increased by a decrease in blood volume and decreased by an increase in blood volume. This type of control of vasopressin is mediated through pressure-change receptors in the large arteries, veins and heart. When blood volume is increased, venous pressure rises. This initiates increased discharge in the nerves from the pressure sensors in the large veins and the atria of the heart. When the increase in blood volume is marked, arterial blood pressure also rises and there is increased discharge from the arterial pressure sensors in the carotid sinus and arch of the aorta as well. The increased discharge inhibits vasopressin secretion. Conversely, when blood volume is decreased, as it is, for instance, after severe bleeding, the rate of discharge from the pressure receptors decreases and vasopressin secretion is increased.

The supraoptic neurons are nerve cells as well as endocrine cells. It is, therefore, not surprising that they also respond to painful stimuli with increased discharge, and to certain drugs that act on the brain. The main control of vasopressin secretion normally is osmotic, but when the other mechanisms are activated they interact with the osmotic stimuli in complex but predictable patterns.

Regulation of oxytocin secretion

The regulation of oxytocin secretion is another example of neuroendocrine integration. A prime stimulus to oxytocin secretion is the infant suckling at the breast. The touch receptors, stimulated by the baby's mouth, initiate impulses which ascend the spinal cord and brainstem to the paraventricular nuclei. The paraventricular neurons become excited, and oxytocin is secreted into the bloodstream at their endings in the median eminence and posterior lobe. The oxytocin causes the mammary ducts to contract, and milk pours out into the waiting mouth of the infant.

Oxytocin also plays a role in the delivery of the baby at the end of pregnancy. The hormone apparently does not initiate labour, but once the uterus starts to contract, the pressure of the infant's head against the tissues initiates nerve impulses that ascend to the paraventricular nuclei. Oxytocin secretion is increased, and the hormone strengthens the contractions of the uterus, aiding the process of labour. There is also evidence that, at least in some species, oxytocin liberated by nerve impulses from the genital region during copulation can increase sperm transport through the nonpregnant uterus. Spermatozoa move up the genital tract more rapidly than can be accounted for by their own swimming motion, and contractions of the uterus appear to propel them along. It is uncertain, however, whether oxytocin aids sperm transport in humans.

The regulation of pituitary secretion is not the only endocrine function of the hypothalamus. It also exerts a minor effect on insulin secretion via the vagus nerves, and increases adrenal medullary secretion. Available evidence indicates that the hypothalamus mediates the increase in epinephrine and norepinephrine secretion seen during such emotional states as anger and fright.

THE PANCREATIC ISLETS

The pancreas is both an exocrine and an endocrine gland. Its exocrine secretion is a juice rich in digestive enzymes that is secreted by way of the pancreatic ducts into the duodenum. Its endocrine secretions are released into the blood stream from the islets of Langerhans, collections of secretory cells scattered throughout the pancreas. The islet cells secrete the polypeptide hormones insulin and glucagon, and they may also secrete the polypeptide hormone gastrin.

Structure. The pancreas stretches across the upper abdomen (see also DIGESTIVE SYSTEM, HUMAN). Its head is on the right side of the body closely applied to the

From W. Bloom and D.W. Fawcett, *A Textbook of Histology*, 9th ed. (1968); W.B. Saunders Company

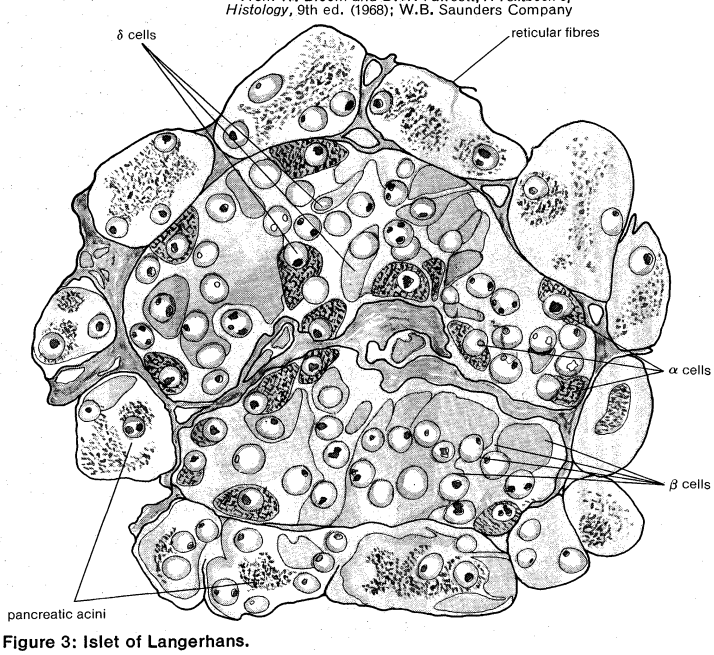


Figure 3: Islet of Langerhans.

duodenum, while its tail reaches almost to the spleen in the left upper corner of the abdominal cavity.

The pancreas receives its blood supply from several major abdominal arteries: the splenic, hepatic, and superior mesenteric. Its effluent blood drains by way of the splenic and superior mesenteric veins to the hepatic portal vein, and thence to the liver. The hepatic portal system is a system of blood vessels that arises in the intestine and other viscera and ends in the liver. Blood from the liver passes by way of the hepatic veins to the general circulation. The hormones of the pancreas (and the gastrointestinal tract) are unlike other hormones in that they are secreted into the hepatic portal system, and consequently the liver is exposed to higher concentrations of these hormones than are other tissues in the body. This fact may have physiological significance, since glucagon and insulin exert an effect on liver cells.

Unmyelinated sympathetic nerve fibres enter the pancreas, mainly from the celiac plexus. Myelinated parasympathetic branches of the vagus nerves also penetrate the gland. Most of these fibres end around the acinar cells that produce the pancreatic exocrine secretion, but some unmyelinated fibres are occasionally seen in close apposition to granulated endocrine cells in the islets. Stimulation of the right vagus nerve increases insulin secretion in animals, and norepinephrine, the mediator at sympathetic nerve endings, exerts both excitatory and inhibitory effects on insulin secretion, the details of which are considered later.

The islets are round or ovoid structures approximately 100 millionths of a metre in diameter. There are approximately 1,000,000 islets in the pancreas of an adult human, and islet tissue makes up about 1 percent of the total weight of the organ. The islets are diffusely distributed throughout the pancreas, but they are somewhat more numerous in the tail than in the head. Each islet contains granulated cells which, on the basis of their staining characteristics, are divided into alpha, beta, and delta, or D, cells. The alpha cells secrete glucagon, the beta cells secrete insulin, and the delta cells probably secrete gastrin. Approximately 70 percent of the islet cells are beta cells, while 25 percent are alpha cells and 5 percent are delta cells.

Carbohydrate metabolism

Islet cell hormones. In considering the functions of the endocrine portion of the pancreas, it is necessary to refer to a number of steps in intermediary metabolism, and especially to two steps involved in the metabolism of the principal sugar in the body, glucose. The details of intermediary metabolism are treated in the article METABOLISM but for convenience a brief summary is presented here.

The products of the digestion of dietary carbohydrates are sugars containing five or six carbon atoms, of which glucose is the most abundant. The products of protein digestion are amino acids, and those of fat digestion are fatty acids and glycerol. Glucose and other simple sugars, amino acids, and fatty acids are absorbed into the bloodstream from the intestine and metabolized in the liver and other parts of the body. Glucose enters cells, and the intracellular glucose either polymerizes to the storage form of carbohydrate, glycogen, or breaks down in orderly steps to a substance called pyruvate. In the absence of oxygen, the pyruvate is converted to lactate, while in the presence of oxygen, it enters the so-called Krebs cycle and is burned to carbon dioxide and hydrogen ions. The hydrogen ions are then oxidized to water, and coupled with this oxidation is the formation of large amounts of the energy-rich phosphate compound, adenosine triphosphate (ATP). ATP provides a ready storehouse of energy for the various energy-requiring processes in the body. Fatty acids are converted to acetylcoenzyme A, and the acetylcoenzyme A enters the Krebs cycle. Amino acids are converted to several intermediates in the metabolic pathways for glucose, and are burned in a similar fashion. It should be noted that glucose, amino acids, and fat can also be formed from intermediates in the Krebs cycle or related small molecules. Little fat is converted to glucose, but large amounts of glucose are converted to fat and there is active formation of glucose from noncarbohydrate substances, particularly from amino acids, by a process known as gluconeogenesis.

Insulin synthesis

Beta cell hormone: insulin. Insulin is synthesized and secreted by the beta cells of the islets. Like other proteins and polypeptides, it is synthesized in the endoplasmic reticulum of the cells. It then moves to the Golgi apparatus, where packets of newly synthesized insulin acquire a surrounding membrane. It is stored in membrane-lined secretory granules. The insulin in the secretory granules is in round or rhombic aggregates in which the insulin is apparently polymerized or complexed with zinc. When the insulin is secreted, the granules move to the periphery of the cell and discharge their contents by the process of emeiocytosis, leaving the cell membrane intact. Once outside the cell, the insulin apparently moves by diffusion across the basement membrane of the beta cell, across the basement membrane of the adjacent capillary endothelium, and through the fenestrated endothelium of the capillary wall into the blood system.

Human insulin is a polypeptide containing 51 amino acid residues. It is made up of two chains (the A and B chains) linked by sulfur-sulfur (disulfide) bridges. Most mammalian insulins differ from one another by only one to four amino acid residues. These differences are generally not great enough to reduce biological activity, but they are great enough to make the insulins antigenic when injected in other species; *i.e.*, they stimulate the formation of anti-insulin antibodies.

Insulin is formed as a single polypeptide chain that is then folded, permitting the disulfide bonds to be formed. The folded molecule, known as proinsulin, is then subjected to peptide chain cleavage, which splits off the amino acid chain connecting the A and B chains. The connecting peptide remains in the beta granules and is secreted along with the insulin. Neither the connecting peptide nor proinsulin have appreciable biological activity. (For a more detailed discussion of amino acid linkages and protein structure, see PROTEIN.)

The half-life of insulin in the circulation is about 30 minutes. It is degraded in almost all the tissues of the body, but most actively in the liver and kidneys.

Functions of insulin

Insulin increases the entry of glucose into muscle and other cells in the body, resulting in a decline in the glucose level in the blood. The mechanism by which this increased entry is effected is unknown, but the widespread metabolic effects of insulin can be explained largely by its fundamental action on glucose entry into cells. It seems clear, however, that insulin has a number of fundamental actions that are independent of its effect on glucose transport. These additional actions include increased movement of amino acids into cells, an increase in the difference in electrical potential across certain cell membranes, and inhibition of hormone-sensitive lipase, the enzyme in adipose tissue that catalyzes the breakdown of fats to fatty acids. The action on amino acid transport fosters protein synthesis and growth; the action on the membrane potential aids entry of potassium into cells; and the action on adipose tissue inhibits release of free fatty acids into the circulation.

The consequences of the increased entry of glucose into cells produced by insulin include not only a decline in blood sugar but a variety of other metabolic changes that are best understood by cataloging the defects that are observed in the presence of insulin deficiency. The sum of these defects is the picture known in humans as diabetes mellitus. This common, but serious, endocrine disorder is discussed in detail in the article ENDOCRINE SYSTEM DISEASES AND DISORDERS; the abnormalities associated with it are only summarized here. During insulin deficiency the blood sugar rises, and there is increased excretion of glucose in the urine. This leads to loss of water from the body and thirst and dehydration. The lack of glucose inside cells activates mechanisms that increase the breakdown of fats and proteins. Products of this breakdown are burned for energy. There is increased appetite caused by inhibition of the appetite-inhibiting cells in the brain. The accumulation of free fatty acids to increased formation of acetoacetic acid and β-hydroxybutyric acid, and their partial excretion in the lungs, giv-

ing rise to "acetone breath." These two acid ketone bodies are the cause of the ketosis seen in the disease. Furthermore, when these ketone bodies are produced in excess of the ability of the body to buffer them, the acidity of the body fluids increases to a serious degree. If untreated, the acidosis is fatal.

All these abnormalities are reversed by insulin. Providing intracellular glucose shuts off the excess fat catabolism, stops the marked production of glucose from amino acids, and permits the normal body defenses to restore acid–base balance to normal.

Large amounts of insulin depress the blood glucose to subnormal levels (hypoglycemia). Glucose is an obligatory fuel for the brain and consequently hypoglycemia is associated with disordered brain function. The symptoms start with nervousness and tremulousness and can progress to convulsions, coma, and death.

Regulation of insulin secretion

Since insulin is of pivotal importance to normal metabolism, it is not surprising that its secretion is regulated by a number of different factors. The primary regulation is by way of a direct feedback of glucose on the insulin-secreting cells. The resting, relatively low level of insulin secretion is increased by a rise in the circulating glucose and decreased by a fall. The glucose must enter the beta cells and be partially metabolized to exert its stimulatory effect. Some event during the metabolism of the glucose apparently leads to increased formation of cyclic AMP, which, in turn, stimulates insulin secretion. Insulin secretion is also increased by certain amino acids and by β-keto acids such as acetoacetic acid. Insulin decreases circulating amino acid levels by increasing their entry into muscle, and decreases β-keto acid levels because it combats ketone body formation. In addition, insulin secretion is increased by stimulation of the right vagus nerve in animals, but the physiological and pathological significance of this factor is uncertain. A variety of other factors affect insulin secretion, but most of them either affect glucose metabolism or cyclic AMP. Normal secretion also requires the presence of adequate quantities of calcium and potassium ions.

There is considerable evidence that agents that stimulate α-adrenergic receptors inhibit cyclic AMP formation, and those that stimulate β-adrenergic receptors stimulate cyclic AMP formation. Alpha-adrenergic stimulating agents inhibit insulin secretion, and β-adrenergic stimulating agents increase insulin secretion. Epinephrine has both α and β actions, so this catecholamine both inhibits and stimulates insulin secretion. The inhibitory effect usually predominates. Norepinephrine exerts a similar predominately inhibitory effect. Insulin secretion is also increased by sulfonylurea compounds such as tolbutamide, and these drugs may act by stimulating cyclic AMP formation in the islets. Tolbutamide and related oral hypoglycemic agents have seen extensive use in the treatment of diabetes, but their efficacy has been questioned. They work only in relatively mild diabetes in which pancreatic islets are still functioning, since the hypoglycemic effect is due to their ability to stimulate insulin secretion.

Another important group of agents that stimulate insulin secretion are the gastrointestinal hormones, which are secreted when food enters the stomach and small intestine. Their properties are considered below. Stimulation of insulin secretion is produced by gastrin, cholecystokinin-pancreozymin (CCK), secretin, and probably glucagon-like immunoreactive factor (GLI) found in the intestinal lining. Pancreatic glucagon also stimulates insulin secretion. It is not established which of the "gut factors" is the physiological stimulus to insulin secretion after meals, but it has been determined that glucose and amino acids in the intestine increase insulin secretion. Furthermore, the increases in insulin secretion produced by glucose or amino acids are greater when these substances are given by mouth than when they are given intravenously. The increase in insulin secretion produced by food in the gastrointestinal tract has physiological value, since it promotes a more rapid and complete utilization of the products of digestion, once they are absorbed.

The increase in insulin secretion produced by a glucose load is biphasic: an initial short-lived increase that apparently represents release of preformed hormone, followed by a slowly developing, prolonged release that appears to represent release of newly synthesized hormone as well. The first component is blocked by agents that block activation of cyclic AMP, while the second is blocked by antibiotics that inhibit protein synthesis.

In summary, insulin secretion is increased by increased amounts of glucose and amino acids in the intestine and in the blood. The hormone acts to increase the entry of glucose and amino acids into the cells, promoting their metabolism and utilization. Insulin secretion thus rises after meals and insulin has been appropriately called the "hormone of abundance."

Alpha cell hormone: glucagon. Glucagon is secreted by the alpha cells of the islets. These cells contain uniform, dense secretion granules. Hormone synthesis and secretion in these cells is presumably similar to that of insulin in the beta cells, but it has not been studied in as great detail.

The structure of porcine glucagon is known. It is a simple polypeptide containing 29 amino acid residues. The structure of glucagon in other species, including man, probably differs by varying numbers of amino acids but is basically similar.

A substance similar to glucagon is found in the walls of the stomach and small intestine. This intestinal glucagon cross-reacts with pancreatic glucagon in immunoassays and hence has been called glucagon-like immunoreactive factor (GLI), but it now appears that it is not identical to pancreatic glucagon.

Pancreatic glucagon is metabolized rapidly, with a half-life that is even shorter than the 30-minute half-life of insulin in man. Glucagon is presumably metabolized to a large extent in the liver, but the details of its metabolism and excretion are unknown.

Functions of glucagon

Glucagon acts on the liver to increase the breakdown of glycogen (glycogenolysis). The glucose thus formed enters the bloodstream and the blood glucose level rises. Glucagon also increases the formation of glucose from amino acids (gluconeogenesis). Its action on the liver is mediated by cyclic AMP. An additional action of both pancreatic glucagon and GLI is stimulation of insulin secretion.

The secretion of glucagon is increased by a decline in blood glucose and decreased by an increase in blood glucose, presumably by a direct effect on the secretory cells. In this respect, it differs from GLI; the secretion of GLI is increased by glucose in the intestinal tract, and GLI may be one of the "gut factors" that increases insulin secretion. The secretion of pancreatic glucagon is increased during fasting, and it seems probable that pancreatic glucagon is a "hormone of fasting" which helps keep the blood glucose from falling to dangerously low levels during starvation.

Glucagon secretion is also increased by amino acids. Because amino acids increase the secretion of both insulin and glucagon, it has been suggested that after a protein meal the insulin fosters the utilization of the amino acids while the glucagon prevents the hypoglycemia which the insulin would otherwise produce.

Delta cell hormone: pancreatic gastrin. Tumours made up of delta cells secrete excessively large amounts of gastrin and normal delta cells contain gastrin, the hormone that is also secreted by the gastric mucosa and increases gastric acid secretion. The delta cell tumours produce a disease in humans called the Zollinger-Ellison syndrome, which is characterized by severe, intractable ulcers of the stomach and duodenum due to the excess acid secretion. The role of pancreatic gastrin in normal physiology, however, is unknown.

THE PARATHYROID GLANDS

There are usually four parathyroid glands, two behind each lateral lobe of the thyroid gland. Their number and location, however, may vary considerably. They receive blood from local branches of the thyroid artery and drain into branches of the thyroid veins. A few nonmyelinated nerve fibres accompany the blood vessels.

Each parathyroid gland is well vascularized and disk-shaped, measuring approximately three by six by two millimetres. Each contains many small cells called chief cells, which secrete parathyroid hormone. There are also smaller numbers of large, granulated cells containing many mitochondria; these are called oxyphil cells, and their function is unknown.

The structure of bovine and porcine parathyroid hormone is known. Bovine parathyroid hormone is a simple peptide containing 84 amino acid residues. Human parathyroid hormone has a similar amino acid composition, but its exact structure is not yet known. It appears to have a relatively short half-life, but little is known about its metabolism and excretion.

Parathyroid hormone and calcium metabolism. Two hormones, parathyroid hormone and calcitonin, are primarily involved in the maintenance of a normal concentration of calcium, and specifically of ionized calcium, in the body fluids. The former hormone, which raises plasma calcium, is secreted by the parathyroid glands, while the latter, which lowers it, comes from the thyroid. The stability of the plasma calcium concentration is essential for a wide variety of body functions, including particularly the maintenance of normal nerve and muscle excitability. Long-term changes in plasma calcium bring about alterations in the structure of the bones.

Calcium enters the body through the gastrointestinal tract and leaves it in large part through the kidneys. The principal calcium storage site is the bones, which contain large quantities of calcium salts. Hormones and other calcium regulating factors, in consequence, act on the gastrointestinal tract, the kidneys and the bones. The total amount of calcium in the body of an adult human is about 1,110 grams (2.4 pounds). The normal concentration in circulating plasma is approximately 10 milligrams per 100 millilitres (5 milliequivalents per litre). There is normally an approximate reciprocal relationship between the calcium and the phosphate in the body. Whenever the product of the calcium concentration and the phosphate concentration in the body fluids exceeds a value called the solubility product, calcium phosphate precipitates. Local changes which alter the calcium or phosphate concentration may be part of the process by which bone calcifies. The details of the calcification process, however, are not known.

Role of vitamin D in calcium metabolism

In addition to the parathyroid hormone and calcitonin, vitamin D plays an important role in the regulation of calcium metabolism. This vitamin is formed in the body from dietary precursors by the action of sunlight on the skin (see also VITAMIN). It is converted to compounds which act mainly to facilitate the transport of calcium across the intestinal mucosal cells into the circulating body fluids.

A low plasma calcium level leads to hypocalcemic tetany, a condition in which there are spasms of the skeletal muscles. These are especially marked in the extremities and the larynx, and if untreated, laryngeal spasm eventually obstructs the airway and leads to death by suffocation. A high level of plasma calcium causes irritability, mental symptoms, and kidney stones as a result of calcium salt precipitation in the urinary tract.

Parathyroid hormone functions in a variety of ways to increase plasma calcium. It acts on the skeleton to increase the amount of calcium moving from the bones to the body fluids. Its action is by way of cyclic AMP, but the exact molecular mechanism responsible for the increased bone dissolution is not known. In addition, parathyroid hormone augments the action of vitamin D in facilitating transport of calcium out of the intestinal lumen into the circulating body fluids. It also acts on the kidney to increase phosphate excretion. This tends to lower plasma phosphate, and secondarily, because of the reciprocal relation between calcium and phosphate, to elevate plasma calcium.

Parathyroid secretion is regulated directly by a negative feedback effect of calcium on the parathyroid gland itself. A rise in plasma calcium inhibits, and a fall stimulates, parathyroid hormone secretion. The gland thus provides an excellent example of simple negative feedback control of the secretion of an endocrine gland. Changes in plasma magnesium appear to exert parallel effects, but the major regulation is apparently via calcium feedback.

The role of calcitonin. The cells responsible for the secretion of calcitonin have been described previously in the discussion of the thyroid gland. These parafollicular cells are found between the thyroid follicles in humans and other mammals, but in birds, fish, and other non-mammalian vertebrates they are collected into a pair of separate glands, the ultimobranchial bodies.

Human calcitonin is a polypeptide with a ring of amino acids at one end closed by the disulfide bonds of two half-cystine residues. Little is known about its metabolism or excretion. It acts on bone to counteract the effect of parathyroid hormone, possibly through an anti-cyclic AMP action. It not only lowers the plasma calcium, but it also lowers the plasma phosphate.

The regulation of calcitonin secretion appears to be mediated by a direct negative feedback of calcium on the cells secreting the hormone. In this case, a rise in calcium stimulates and a fall inhibits calcitonin secretion.

The exact physiological role of calcitonin is uncertain. Although the effects of its injection can be demonstrated in man, there is very little derangement of calcium metabolism after thyroidectomy. On the other hand, when the parathyroid glands are removed, plasma calcium drops and the individual develops hypocalcemic tetany, which, if untreated, is fatal.

THE GASTROINTESTINAL MUCOSA

The hormones secreted by cells in the mucosa, or lining, of the stomach and small intestine are of historical as well as current interest, since the whole "chemical messenger" concept stemmed from work on one of them. Six gastrointestinal hormones are listed in Table 1 (counting gastrin I and II as a single hormone). The first three of these are well established, and the first two have been synthesized. The fourth probably exists as a separate hormone, and the fifth and sixth are hypothesized to explain effects that appear to be hormonally mediated and cannot be explained by actions or interactions of the other four. These interactions are extensive, and, although gastrointestinal hormones in addition to these six may exist, it does not seem necessary at present to postulate their existence.

The cells secreting gastrin are modified epithelial cells embedded for the most part in the walls of the gastric juice-secreting pyloric glands in the distal portion (antrum) of the stomach. The identity of the cells secreting the other hormones is not known.

Hormones. *Gastrin.* Gastrin is secreted primarily from the mucosa of the antral portion of the stomach, although small amounts may also be secreted by cells in the mucosa of the small intestine (intestinal gastrin). In humans and in all other species examined gastrin is a mixture of approximately equal amounts of two components, gastrin I and gastrin II. Each contains the same 17 amino acid residues, but gastrin II has a sulfate ($-SO_3H$) group on the tyrosine residue in position 12, while gastrin I lacks this sulfate group. The physiological activities of these two gastrins are approximately equal. A tetrapeptide containing amino acid residues 14–17 of gastrin has all the activity of gastrin, although the total activity is about 10 percent of the parent compound. Other polypeptides also have gastrin activity.

Action of gastrin

The principal action of gastrin is to increase gastric secretion of acid and the proteolytic digestive enzyme, pepsin, but it also increases the movements of the stomach (gastric mobility) and has a variety of other actions. Gastrin is secreted when the antrum of the stomach is distended and when the mucosa is bathed by alcohol, protein, or the products of protein digestion. Stimulation of the vagus nerve to the stomach also stimulates gastrin secretion, and vagal stimulation acts directly to increase gastric acid secretion and motility. Thus, there is a neuroendocrine link between the vagus nerve and gastrin secretion, and the actions of the neural

and endocrine mechanism abet and supplement each other. Gastrin has been measured in the circulating blood of humans by radio-immunoassay, but very little is known about its metabolism and excretion.

Secretin. In England in 1902, two physiologists, W.M. Bayliss and E.H. Starling, first demonstrated that the excitatory effect of food in the small intestine on pancreatic secretion was due to a blood-borne factor. In subsequent experiments, they extracted a substance from the duodenal mucosa which they called secretin. They appreciated, however, that the significance of their discovery went far beyond secretin itself. They suggested that other cells in the body might secrete "chemical messengers" that entered the circulation and acted on tissues some distance away, and they coined the term hormone to categorize these substances. The list of "chemical messengers" has now grown to more than 50, and the whole science of modern endocrinology attests to the correctness of the hypothesis advanced by Bayliss and Starling on the basis of their work with secretin.

Secretin is secreted by the duodenal mucosa when the mucosa is bathed with acid or the products of protein digestion. It has a number of effects, the principal one being an action on the pancreas to cause the secretion of a dilute pancreatic juice rich in bicarbonate. The structure of porcine secretin is known. It is different from gastrin and CCK, and it resembles pancreatic glucagon; indeed, 14 amino acid residues occupy the same positions in the two hormones. Little is known as yet about the metabolism or excretion of secretin.

Cholecystokinin-pancreozymin (CCK). Experiments some years ago demonstrated that there was a hormonally mediated contraction of the gallbladder when food, particularly fat, was placed in the duodenum, and that there was, as well, a hormonally mediated increase in the secretion of an enzyme-rich pancreatic juice produced by food in the duodenum. It was initially thought that a gastrointestinal hormone called cholecystokinin was responsible for the gallbladder contraction and another hormone called pancreozymin responsible for the increased pancreatic secretion. It has since been proved that a single gastrointestinal hormone is responsible for both actions, and this hormone has been named cholecystokinin-pancreozymin (CCK). This hormone has other actions as well, including some gastrin activity.

The chemistry of CCK

The amino acid composition of porcine CCK is known, but the exact sequence of its 33 amino acids has not yet been determined. It is chemically similar to gastrin, and, like gastrin, contains a sulfated tyrosine residue near the C-terminal end of the molecule. As yet, little is known about its metabolism and excretion.

It is interesting that a polypeptide with a structure and actions that resemble both gastrin and CCK has been isolated from the skin of an Australian frog. This material, caerulein, is being used to produce gallbladder contraction in humans.

Intestinal glucagon. Another apparent gastrointestinal hormone is a factor in the intestinal mucosa which closely resembles pancreatic glucagon. This intestinal glucagon cross-reacts with glucagon in some immunoassays, but although its exact structure is unknown, it appears to differ from pancreatic glucagon. Because of its immune reactions, it has been called glucagon-like immunoreactive factor, or GLI. Its concentration in blood appears to rise after a meal. Its properties apparently resemble those of secretin and pancreatic glucagon.

Other gastrointestinal hormones. The interactions between the gastrointestinal hormones mentioned above offer an explanation for various responses that were thought to be mediated by other gastrointestinal hormones. For example, it was postulated that a hormone called enterogastrone was secreted from the intestinal mucosa and acted on the stomach to inhibit gastric secretion and emptying. It now appears that this inhibition is caused principally by the action of secretin. On the other hand, two effects presumably mediated by hormones remain unexplained, and there may be separate hormones from the intestinal mucosa that are responsible for them. One is the increased contractions of the intestinal villi when food is in the intestine, and the postulated hormone in this case is called villikinin. The other is the increased intestinal gland secretion when food is in the intestine, and the postulated hormone is called enterocrinin. Proof of the existence of these hormones awaits their isolation and chemical identification.

Cross-reactions. Although gastrin, CCK, and secretin both have a principal action, it is now clear that there are also extensive cross-reactions among these three gastrointestinal hormones and presumably with GLI as well. Gastrin, CCK, and secretin all stimulate insulin secretion; in addition, gastrin and CCK each stimulate gastric acid and pepsin secretion, gastric and intestinal motility, pancreatic secretion of enzymes and bicarbonate ions, liver secretion of bicarbonate ions, contraction of the gallbladder, and secretion of certain mucus-secreting glands in the duodenum. Secretin inhibits gastric acid secretion and gastric and intestinal motility, but otherwise its actions resemble those of gastrin and CCK. Insofar as they have been tested, the actions of pancreatic glucagon, and presumably those of intestinal glucagon as well, resemble those of secretin. In addition, the various gastrointestinal hormones synergize with, and also antagonize, one another, so that the pattern of gastrointestinal responses depends not only on single hormones but on interactions between the various hormones.

THE THYMUS

The thymus is a lymphoid organ located behind the sternum in the chest. It is made up of an outer cortex densely packed with lymphocytes and an inner medulla containing more widely scattered lymphocytes and cells arranged in clusters (Hassall's corpuscles). The arterial supply to the organ comes from branches of the mammary and thyroid arteries, and the veins drain into the innominate and thyroid veins. The few nerve fibres that enter the gland probably regulate the calibre of its blood vessels.

The thymus is large in fetal life and infancy and undergoes gradual involution with advancing age. The Hassall's corpuscles may be remnants of the branchial pouches from which the thymus arose during embryonic development. Adrenocortical hormones cause lymphoid tissue to atrophy, and since the thymus is a lymphoid organ, its size decreases markedly in individuals subjected to stress. In patients who have died after prolonged, severe infection, for example, the thymus is markedly atrophic. On the other hand, the thymus is relatively large in individuals in whom death has been sudden and not preceded by prolonged illness.

The thymus and cellular immunity

The thymus is known to be the key to the production of the lymphocytes responsible for cellular immunity. There are two types of immunity: humoral immunity, mediated by circulating antibodies, and cellular immunity, which is delayed in onset (see IMMUNITY). Cellular immune responses include the delayed hypersensitivity reactions responsible for such phenomena as rejection of skin, kidney, and heart grafts or transplants from other individuals (homografts). Early in life, the lymph nodes become populated with lymphocytes capable of mediating this type of response. It appears that the lymphocytes are affected by a thymic hormone. In animals, removal of the thymus in early life prevents the rejection of tissue grafts from other animals, and in thymectomized animals thymic grafts in tissue chambers with pores too small to permit passage of cells restore the ability to develop delayed hypersensitivity reactions. Cell-free extracts of thymic tissue accelerate the rejection of skin grafts and increase the number of circulating lymphocytes. The hormonal factor involved seems to be a polypeptide of relatively low molecular weight and has been called thymosin. The details of its formation and function remain to be determined.

THE PINEAL BODY

The pineal body properly belongs in the list of organs with suspected rather than proved endocrine functions. It is located deep between the cerebral hemispheres of

the brain at the top of the midbrain. The pineal contains neuroglia and cells which look as if they might be secretory. The gland is large in children and begins to involute before puberty. With advancing age in humans, small deposits of calcium and magnesium phosphate and carbonate (pineal sand) accumulate, making the structure visible on X-rays of the skull. The pineal receives a sympathetic nerve supply by way of sympathetic fibres that ascend through the sympathetic cervical ganglia from the chest. It receives blood from arterial vessels in the vicinity, and its effluent blood drains through the vein of Galen to dural venous sinuses in the skull.

Postulated functions of the pineal

Numerous endocrine and other functions have been attributed to the pineal, but most of the postulated functions have failed to withstand the test of time and experimental verification. Among the postulated functions in this category are a role in maintaining mental health and secretion of a factor or factors regulating adrenocortical secretion of aldosterone. Pineal tumours are sometimes associated with precocious sexual development and early puberty in boys, and this observation plus some suggestive data in animals indicates that the pineal may be involved in the regulation of gonadal function.

The postulated pineal hormone is melatonin, one of a variety of substances containing the indole ring that are found in the pineal. The pineal manufactures melatonin from 5-hydroxytryptamine (serotonin). Injected melatonin is metabolized to 5-methoxytryptamine, 5-methoxyindoleacetic acid, and 6-hydroxymelatonin, but the fate of endogenous melatonin is unknown. Melatonin lightens the skin of tadpoles, and there is some evidence that it exerts an antigonadotrophic effect in rats, slowing the onset of puberty.

The enzyme responsible for the synthesis of melatonin from 5-hydroxytryptamine is hydroxyindole-*o*-methyltransferase. In the rat, the activity of this enzyme in the pineal undergoes marked diurnal fluctuation, being high during the hours of darkness and low during the hours of light. The changes in enzyme activity have been shown to be neurally mediated and are prevented by cutting the sympathetic nerve fibres to the pineal. Light can alter reproductive cycles in the rat, and it has been suggested that the modifications are brought about, at least in part, through the action of melatonin. Melatonin, however, has not yet been detected in the blood of the rat or the human. In addition, the precocious puberty produced in humans by pineal tumours could well be due to pressure on the hypothalamus. Therefore, it cannot be stated, at least at the present time, that any endocrine function has been established for the pineal body in humans.

THE KIDNEYS

In addition to their excretory function and their role in the regulation of acid–base balance, the kidneys secrete at least two hormones. One of these, renin, acts on a protein in the blood to form a peptide which increases blood pressure and stimulates adrenocortical secretion of aldosterone. The other, the renal erythropoietic factor, stimulates the formation of a protein which increases red blood cell production in the bone marrow.

Kidney hormones. *Renin.* Renin is secreted by the juxtaglomerular cells. These granulated cells are located in the wall of each afferent arteriole, the arterial blood vessel that enters the renal glomerulus. The tubule from each glomerulus doubles back and touches the afferent arteriole at its junction with the glomerulus from which the tubule arose. At this point of contact, the tubular lining is thickened to form the structure known as the macula densa. Juxtaglomerular cells are most prominent and renin content is highest in the afferent arterioles of glomeruli in the outer part of the renal cortex, but renin is found throughout the renal cortex. There are large numbers of unmyelinated adrenergic sympathetic nerve fibres in close association with the juxtaglomerular cells, and some of the fibres appear to end on these cells. Thus, the juxtaglomerular cells appear to be innervated gland cells.

Renin is a protein with a molecular weight of approximately 50,000. It has not been obtained in pure form and its structure is unknown. Like other protein hormones in other endocrine cells, it is stored in granules in the juxtaglomerular cells. Secretion probably occurs by the process of emeiocytosis. Some renin appears to be produced by organs other than the kidney. Appreciable amounts are found in the pregnant uterus, and even in nonpregnant women and in men there are small but measurable amounts of renin in the circulation after removal of both kidneys. The physiological role of this extrarenal renin appears to be negligible.

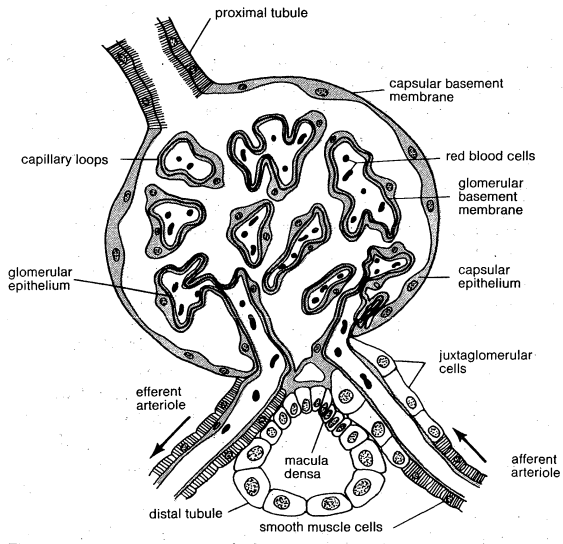


Figure 4: Diagram of a renal glomerulus showing the juxtaglomerular cells and the macula densa in relation to other structures of the kidney.
From A.W. Ham, *Histology*, 6th ed. (1969); J.B. Lippincott Co.

Renin-angiotensin system

Renin is different from other hormones in that it does not act on a tissue but on a protein in blood. This protein, called angiotensinogen, is formed in the liver, and is a globulin with molecular weight of approximately 58,000. From it renin splits off a decapeptide, angiotensin I. The angiotensin I is in turn converted to angiotensin II, a polypeptide containing eight amino acid residues, by removal of the dipeptide, histidyl-leucine, from the C-terminal end. The conversion is catalyzed by converting enzyme. There is some converting enzyme in plasma, but most of the conversion of angiotensin I to angiotensin II occurs in the lungs. The pulmonary conversion is efficient; it takes a red blood cell less than one second to traverse the pulmonary capillaries, yet during passage through the lungs, conversion of angiotensin I to angiotensin II is about 50 percent complete. It is notable that angiotensinogen can be called a "prohormone" and the liver called an endocrine gland, since angiotensinogen is converted in the blood by the sequential action of renin and converting enzyme into the physiologically active peptide angiotensin II.

Angiotensin II has several actions in the body. It acts on the vascular smooth muscle to produce constriction of small arteries and arterioles. This increases the resistance against which the heart must pump, and the blood pressure rises. Angiotensin II also stimulates the secretion of the sodium-retaining hormone aldosterone from the adrenal cortex. In addition, it may stimulate thirst by an action on the brain and may act on the brain to increase blood pressure by way of increased discharge of the sympathetic nervous system. Its action on blood pressure depends on the sodium and water balance of the individual; when the individual is salt depleted, its pressor activity is low while its aldosterone stimulating activity is maintained or even increased.

The angiotensin II in the circulation is rapidly metabolized, in part by enzymes in the blood, but also to a large degree by a poorly understood mechanism in peripheral vascular beds that appears to remove the peptides from the circulation. Its half-life in humans is less than one minute. Renin has a somewhat longer half-life, about 30–

60 minutes, and is metabolized for the most part in the liver.

Theories of renin secretion

The regulation of renin secretion has been the subject of much research and debate. Three theories have been advanced to explain how the stimuli affecting renin secretion activate the juxtaglomerular cells. A renal baroreceptor theory holds that the rate of renin secretion is inversely proportional to the degree of stretch of the juxtaglomerular cells. A so-called macula densa theory holds that the rate of renin secretion is inversely proportional to the rate at which sodium is transported from the urine to the blood across the macula densa cells. A neural theory holds that the sympathetic nervous system regulates renin via a direct secretomotor nerve supply to the juxtaglomerular cells.

The renal baroreceptor theory was advanced to explain why renin secretion is increased when there is a decline in blood pressure and decreased when the blood pressure rises. The higher the blood pressure, the more the walls of the arteries and arterioles, and hence the juxtaglomerular cells, are stretched.

The macula densa theory was advanced to explain the frequent inverse correlation between the rate of renin secretion and the rate at which sodium is excreted in the urine. According to this view, the macula densa is a sensor that monitors changes in the composition of the urine in the tubules. When the amount of sodium reaching the macula densa is low, or when sodium transport across the macula densa is inhibited by drugs such as diuretics, renin secretion increases.

It has been proved that stimulation of the sympathetic nervous system increases renin secretion, via both the renal sympathetic nerves and the circulating epinephrine from the adrenal medulla. These effects of sympathetic stimulation could be mediated by stretch of the juxtaglomerular cells or by way of induced changes in sodium excretion monitored by the macula densa. The juxtaglomerular cells appear to be innervated, and there is evidence that norepinephrine and epinephrine act directly on the juxtaglomerular cells to increase renin secretion. There is also evidence that cyclic AMP is involved in the secretory process.

These theories are not mutually exclusive. In addition, none of the theories can explain all the experimental data, and it appears that there must be more than one mechanism regulating renin secretion. The operation of renin, angiotensin I and II, and aldosterone in a negative feedback system that helps defend the volume of the body fluids has been described in the section concerning the adrenal glands. The renin-angiotensin system is also responsible for certain forms of high blood pressure in humans. The role of angiotensin in the regulation of normal blood pressure appears to be related to its effect on aldosterone secretion rather than its direct blood pressure-elevating action. Its possible role in the regulation of thirst remains to be determined.

Erythropoietin

Renal erythropoietic factor. It has been known for some time that the formation of red blood cells in the bone marrow is under hormonal control. The hormone involved is called erythropoietin. It is a glycoprotein with a molecular weight of approximately 60,000. It was thought for some time that the kidney secreted erythropoietin, but there is now reason to believe that the kidney secretes instead a substance that acts on or with a substance produced by the liver to produce erythropoietin in the circulation. The substance secreted by the kidney has been called renal erythropoietic factor (REF). Erythropoietin increases red blood cell production by stimulating the conversion of primitive precursor cells in the bone marrow into pro-erythroblasts, bone marrow cells that subsequently mature and are released into the circulation as red blood cells. The half-life of erythropoietin is approximately five hours, and it appears to be inactivated primarily in the liver.

There are apparent similarities between the renin-angiotensin system and the REF-erythropoietin system. It is clear, however, that the two systems are separate. Angiotensin II has no erythropoietic effect, and erythropoietin has no effect on blood pressure or aldosterone secretion.

The amount of erythropoietin in the circulation is increased when the number of red blood cells in the circulation is reduced, as by bleeding, and is decreased when the number of red blood cells is greater than normal; thus, erythropoietin, like so many other hormones, is under negative feedback control. Erythropoietin is also increased in persons who ascend to high altitudes or have their oxygen intake reduced with resulting oxygen lack in the tissues. The part of the kidney that secretes the REF and how its secretion is stimulated is unknown.

It is to be noted that, in humans and some other animal species, erythropoietin remains in the circulation after nephrectomy, and its circulating level can even be increased somewhat by oxygen deficiency and other stimuli. Consequently, it seems probable that the kidney is not the only source of REF. Some renin is also found in the circulation after the kidneys are removed (nephrectomy), but the amounts are physiologically insignificant.

Other kidney factors. In addition to renin and REF, the kidney may secrete a substance that lowers blood pressure. There is some reason to believe that this substance is a prostaglandin. Prostaglandins are unsaturated fatty acids that have widespread but incompletely understood actions in the body. Other investigators have argued that a protein blood pressure-lowering agent is secreted by the kidney. The nature, and even the existence, of a renal blood pressure-lowering factor is uncertain, and only future research can settle the question of whether additional hormones are secreted by the kidney.

Prostaglandins

OTHER POSSIBLE HORMONES AND ENDOCRINE ORGANS

It seems unlikely that the hormones treated above are all the chemical messengers in the human body, and it will be surprising if future research does not lead to the discovery of additional hormones. There are already several indications that as yet unisolated hormones exist. For instance, the level of circulating granulated white blood cells appears to be controlled by a humoral factor that has been tentatively named granulopoietin. The source of this substance is unknown. There is some evidence that the carotid body secretes a hormone that is involved in the control of red blood cell production but that is different from erythropoietin. There have been repeated reports of a humoral factor, sometimes called "third factor," that decreases sodium reabsorption in the proximal tubule portion of the nephrons in the kidney. This "hormone" has not been isolated nor has its source been elucidated, and many investigators have argued that it does not exist. These are but a few examples of current research pointing to the possible existence of other endocrine products and endocrine glands in experimental animals and man.

BIBLIOGRAPHY. The standard endocrinology textbooks are R.H. WILLIAMS (ed.), *Textbook of Endocrinology*, 4th ed. (1968); and C.D. TURNER and J.T. BAGNARA, *General Endocrinology*, 5th ed. (1971). Williams' text is oriented toward human disease, and the coverage of endocrine abnormalities in man is exhaustive. In Turner and Bagnara the emphasis is on comparative endocrinology with much material on birds, reptiles, and fish. A basic text, short and to the point, is C.T. SAWIN, *The Hormones* (1969). There is a concise review of the fundamentals of endocrine physiology in W.F. GANONG, *Review of Medical Physiology*, 5th ed. (1971).

There are numerous publications dealing with various specialized aspects of endocrinology. The subject of brain-endocrine interactions is covered in detail in *Neuroendocrinology*, ed. by L. MARTINI and W.F. GANONG, 2 vol. (1966–67). The same editors have published two subsequent volumes, *Frontiers in Neuroendocrinology* (1969) and *Frontiers in Neuroendocrinology* (1971). *The Pituitary Gland*, ed. by G.W. HARRIS and B.T. DONOVAN (1966), is an exhaustive three-volume reference work. Modern short volumes on various glands include: R.J. WURTMAN, J. AXELROD, and D.E. KELLY, *The Pineal* (1968); R.J. WURTMAN, *Catecholamines* (1966); S.S. LAZARUS and B.W. VOLK, *The Pancreas in Human and Experimental Diabetes* (1962); T. SYMINGTON, *Functional Pathology of the Human Adrenal Gland* (1969); and P. RHODES, *Reproductive Physiology for Medical Students* (1969).

Valuable surveys of currently active areas of endocrine research are published in the series of annual volumes called *Recent Progress in Hormone Research.*

(W.F.G.)

Endocrine System Diseases and Disorders

Endocrine glands secrete hormones that enter the bloodstream directly and influence biochemical processes within target glands or tissues. Details of the structure and function of human endocrine glands and biosynthesis of the hormones they produce are given in the article ENDOCRINE SYSTEM, HUMAN; the nature and activities of hormones—from plants up through the animal kingdom to vertebrates—are presented in HORMONE; and endocrine systems for all animals, on a comparative basis, are covered in ENDOCRINE SYSTEMS. Most of the recognized hormones are composed of amino acid chains or steroids and increase or decrease the rate of biochemical processes within the cells or tissues. Mediators between the hormone and the target cell include the chemicals adenyl cyclase, cyclic adenosine monophosphate (AMP), and other such "messengers." Products of biochemical activity within cells provide feedback signals that serve to maintain endocrine activity within the physiologic range.

Long-recognized endocrine tissues in the so-called higher centres lie within the hypothalamus, a portion of the brain located above the pituitary gland, to which it is connected by the pituitary stalk. Releasor or inhibitor hormones produced in the hypothalamus reach the anterior pituitary via the stalk and, in turn, regulate the manufacture and secretion of hormones. The latter control growth and other metabolic events (growth hormone, or GH), thyroid function (thyroid-stimulating hormone, or TSH), adrenocortical activity (adrenocorticotropin, or ACTH), gonadal maturation and production of sex hormones, ova, and spermatozoa (follicle-stimulating hormone, or FSH, and luteinizing hormone, or LH), lactation (lactogenic hormone, or prolactin), and activity of the pigment- (melanin) producing cells (melanocyte-stimulating hormone, or MSH).

Hormones of the anterior pituitary gland

This article is divided into the following sections:

I. The pituitary
  - Diseases and disorders of the anterior pituitary
  - Diseases and disorders of the posterior pituitary

II. The thyroid
  - Excesses of thyroid hormones
  - Deficiencies of thyroid hormones
  - Inflammatory disorders: thyroiditis
  - Tumours
  - Associated disorders

III. The parathyroids
  - Excesses of parathyroid hormone
  - Deficiencies of parathyroid hormone

IV. The adrenals
  - Adrenal cortex
  - Adrenal medulla

V. The gonads
  - Premature development
  - Disorders of the testes
  - Chromosomal disorders
  - Intersexes
  - Other disorders of the ovary
  - Tumours
  - The breast in endocrine disorders

VI. The pancreas: diabetes mellitus
  - Classification of diabetes mellitus
  - Occurrence of diabetes
  - Cause and treatment of diabetes

VII. Other endocrine glands
  - The pineal gland
  - The thymus

I. The pituitary

DISEASES AND DISORDERS OF THE ANTERIOR PITUITARY

Excesses of pituitary secretions. *Acromegaly.* Excesses of growth hormone secreted by a particular type of pituitary tumour—usually one made up largely of cells that take on an eosin stain (eosinophilic adenoma)—produce undue growth of the skeleton and soft tissues. If this occurs in a child, gigantism results; if in an adult, the overgrowth is confined to the bones of the jaw, hands, feet, eyebrow ridges, and the soft tissues. Such persons develop more or less characteristic physical features, with coarsening of the face; thickening—and at times infolding—of the scalp; large hands, feet, and digits; and enlargement of the heart, liver, and other internal organs.

Acromegalic individuals are likely to develop congestive heart failure, particularly when blood pressure becomes high. Muscle weakness is frequent because muscle overgrowth produced by growth hormone decreases their capacity for work. At times excessive overgrowth of bone edges involves the joints and produces pain. The bones may become thin and porous—a condition known as osteoporosis.

Signs and symptoms of acromegaly

Expansion of the tumour within the pituitary can lead to losses of the follicle-stimulating and luteinizing hormones of this gland and to decreases in libido, potency, menstruation, and fertility. Increased metabolism with high basal metabolic rate readings, usually unrelated to any laboratory evidence of thyroid overactivity, is frequent in active acromegaly. Occasionally, however, thyroid overactivity or underactivity may coexist with acromegaly. The thyroid underactivity could result from losses of pituitary TSH and the overactivity from TSH production by the pituitary tumour. Each of these, however, could occur without causal relationship to the acromegaly. Secondary adrenocortical insufficiency does not appear unless the tumour destroys the ACTH-secreting cells of the pituitary. Indeed, adrenal hyperplasia may be present. Undue facial and other hirsutism may develop in women for reasons that are not clear. Diabetes mellitus appears in 20 to 40 percent of acromegalic individuals because excesses of growth hormone block the action of insulin. If the pituitary tumour enlarges, it can produce visual field defects, blindness or paralysis of the eye muscles, and can injure the posterior pituitary gland or the hypothalamus. Also, hemorrhage into the tumour can produce sudden loss of vision.

Occasionally, acromegalic persons, both male and female, develop a milk-like secretion from the breasts (galactorrhea), probably because growth hormone chemically resembles human lactogenic hormone, or prolactin, which induces lactation. Such galactorrhea, however, could also result from injury to the hypothalamus with loss of the prolactin inhibitor, which prevents pituitary secretion of prolactin.

Acromegaly can be a feature of the so-called Zollinger–Ellison syndrome, in which tumours of the parathyroids, pancreatic islets, thyroid, or adrenal glands may also be present.

An abnormal elevation of serum growth hormone that is not lowered by ingestion of glucose is the best index of active acromegaly or gigantism. Increases in serum inorganic phosphorus, new bone formation at the rib ends, and high levels of the amino acid hydroxyproline in the urine are indices of lesser value.

Treatment of acromegaly

Acromegaly may be treated by surgical removal of the pituitary tumour or destruction of it by X-ray irradiation or liquid nitrogen. Rarely, the pituitary tumour will cease to secrete growth hormone because of a spontaneous hemorrhage or blockade of the blood supply. Decreases in acromegalic manifestations and amelioration of diabetes mellitus have followed therapy with female hormones—estrogen or medroxyprogesterone—which reduce the secretion of growth hormone. For those treated by pituitary surgery, irradiation, or other measures and for those who spontaneously develop deficits of gonadal, thyroidal, or adrenocortical hormones, replacement hormone therapy is necessary.

Gigantism caused by excesses of growth hormone produced by a pituitary adenoma is different from gigantism associated with various types of cerebral injury (cerebral gigantism) or with fractional hypopituitarism (so-called fractional hypopituitarism with gigantism). Growth hormone levels are normal in the last two conditions, in each of which the gigantism may well result from excesses of lactogenic hormone.

Cushing's disease. Cushing's disease is adrenocortical overactivity chiefly, or entirely, of the cortisol-excess type resulting from an ACTH-secreting tumour—usually a basophilic (basic staining), and less often a chromophobe (nonstaining) adenoma or cancer of the anterior pituitary, frequently referred to as pituitary basophilism.

Cushing's syndrome is a more general term indicating that the adrenocortical steroid excesses that characterize

this disorder arise for reasons other than a pituitary tumour—*e.g.*, as a consequence of adrenocortical hyperplasia, adenoma, or cancer in association with cancers secreting ACTH-like polypeptides, and similar conditions. The most common cause today, however, is the administration of pharmacological dosages of cortisol (hydrocortisone) steroids.

The symptoms and signs in both Cushing's disease and syndrome are largely the same and are discussed later. Enlargement of the pituitary fossa or hyperpigmentation are not present, however, in untreated Cushing's syndrome arising in the adrenal cortices.

Causes of hypopituitarism

Deficiencies of pituitary secretions. *Simmonds–Sheehan syndrome.* Hypopituitarism may appear as a consequence of pituitary destruction (infarction) following extensive blood loss and circulatory collapse during delivery of pregnancy (Sheehan's syndrome); it may also result from blood clots, tumour, abscess, syphilis, tuberculosis, or unknown causes (Simmonds' disease); or it may follow pituitary surgery or irradiation, injury to the hypothalamus, influenza, or prolonged cortisol steroid therapy. Destruction of the pituitary need not be complete, and hormonal deficits may be partial. Cachexia, or marked weight loss, is not an essential feature, and even obesity may be present. Wrinkling of the face and decreased sweat and sebaceous gland activity often appear. Pubic and axillary hair is lost, and eyebrows become sparse.

When all of the pituitary hormones (GH, TSH, ACTH, MSH, FSH, LH, and prolactin) are lost, panhypopituitarism is said to be present. Growth hormone deficits produce dwarfism and predispose to low blood-sugar levels. The loss of TSH results in secondary (*i.e.*, pituitary) myxedema or hypothyroidism, with cold intolerance, torpor, slow speech, anemia, low basal metabolic rate, low serum protein-bound iodine, diminished radioactive iodine uptake, and sometimes increased serum cholesterol.

The absence of ACTH results in secondary hypoadrenocorticism, with loss of cortisol secretion and accompanying pallor, weakness, low blood pressure, diminished blood sugar, decreased excretion of water loads, low plasma and urine steroids, losses of adrenal androgens and estrogens with resultant decreases in or absence of sex hair, and adrenal crises (see the section below on deficiencies of adrenocortical hormones). Coma and death may occur as a result of adrenal insufficiency and myxedema.

Loss of FSH and LH is followed by decreased and then absent libido and potentia, amenorrhea, atrophy of internal and external genitalia, absence of sperm and ova production, infertility, and diminished sex hair. Lactation cannot occur in the absence of prolactin.

Treatment consists of replacement of lost hormone secretions—*i.e.*, cortisol, thyroid hormone, and androgens or estrogens. These measures correct these particular hormone deficits but do not restore fertility, which is dependent on adequate supplies of FSH and LH.

Reasons for hypophysectomy

Hypophysectomy. Removal of the pituitary (hypophysectomy) and pituitary stalk section by surgery or destruction by radioactive implants, X-ray or proton beams, ultrasound waves, and coagulation with liquid nitrogen have been undertaken in treatment of cancer and of vision-threatening eye changes in diabetes mellitus.

Hypophysectomy in individuals with cancer of the breast or prostate that has spread to bone often relieves pain and reduces the size of the tumour. In diabetes mellitus, hypophysectomy may stay the course of blindness. Hemorrhages and new blood-vessel formation cease. A type of scar tissue (glial retinitis proliferans) persists when present, and blood clotting and leakage of blood from capillaries (micro-aneurysms) may continue, but visual acuity often improves. High blood pressure, when present, tends to fall. Kidney function, when reduced, is not benefitted. Insulin requirements usually decrease.

Persons with reduced or absent pituitary function must receive daily cortisol and usually thyroid hormones as well. Sex hormones are not ordinarily replaced in persons with cancer of the breast or prostate but may be used in the treatment of diabetic individuals.

Chromophobe adenoma. Chromophobe adenomas make up 60 to 95 percent of pituitary tumours and are the most common cause of enlargement of the sella turcica, the bone "cavity" occupied by the pituitary. Only rarely are these true cancers, and calcium deposits occur in only about 6 percent of the cases. Enlargement of the pituitary fossa may also occur, however, with extra-pituitary tumours in the brain, Hodgkin's disease, or extensions from cancer elsewhere in the body. The space normally occupied by the pituitary may also become enlarged during pregnancy, in myxedema or cretinism, by dilatation of a nearby artery following a localized hemorrhage, or as a result of increases in intracranial pressure.

Chromophobe adenomas are generally considered to be nonsecretory, but this is not always the case. Thus, Cushing's disease, acromegaly, thyrotoxicosis, and galactorrhea have been observed in persons with what appeared to be a chromophobe adenoma. With appropriate staining techniques, such chromophobe tumours often prove to consist of amphophils—cells that are capable of secretory activity. Inappropriately high amounts of vasopressin may also occur in the plasma and urine of individuals with a chromophobe adenoma, perhaps because the tumour is secreting this polypeptide or somehow the rate of its production is increased by the hypothalamus. The signs of this condition are retention of water and dilution of plasma sodium and chloride. The losses of pituitary FSH, LH, ACTH, lactogenic, or growth hormone encountered with chromophobe adenomas result from pressure exerted on the hormone-producing cells by the tumour within the gland, with consequent destruction of those pituitary cells.

As with acromegaly, the possibility of adenomas in other endocrine glands is suspected in anyone with a chromophobe tumour. These tend to be familial and have been observed in various combinations, such as the Zollinger–Ellison syndrome, with chromophobe or eosinophil pituitary adenoma, and in adenomas of the parathyroid, pancreas, and other glands in association with peptic ulcers or diarrhea.

Treatment of chromophobe adenoma

The operative mortality with surgical removal of a chromophobe adenoma ranges between 5 and 15 percent. The appearance of sudden blindness, however, in a person with a chromophobe adenoma is an indication that a hemorrhage has occurred in the tumour and is a signal for instant surgery with the hope of restoring vision. The relative ineffectiveness of irradiation in the treatment of chromophobe pituitary adenomas in the past has been partially cancelled by the use of higher voltages directed more precisely at the tumour. Though results with X-ray irradiation of the adenoma may equal those seen following surgical removal, better results may be obtained by combining the two forms of therapy; *i.e.*, operative removal followed by irradiation.

Hormonal deficits may precede, as already indicated, or follow pituitary surgery or irradiation, requiring maintenance on full replacement dosages of cortisol, thyroid, and male or female sex hormones. So-called pituitary coma appearing in a person with a treated or untreated chromophobe adenoma is usually responsive to massive dosages of cortisol steroids. Low blood-sugar levels and myxedema coma also require therapy when they occur in such individuals.

Craniopharyngioma. Craniopharyngiomas are brain tumours that arise from Rathke's pouch, an embryologic remnant located posterior to, and above, the pituitary. As the tumours grow they encroach on this gland and on the hypothalamus. They can occur at any age, with peak frequencies in the second, third, and fourth decades. The tumour is somewhat more common in males than in females. Calcium deposits within the tumour are not uncommon, and distortion or destruction of that portion of the base of the skull in which the pituitary is housed occurs in 30 to 90 percent of the cases.

Most individuals with a craniopharyngioma report disturbances of vision. Sleepiness and personality changes may also develop; headaches are common and may be accompanied by nausea and vomiting. Rarely, convulsions, fever, and one-sided muscle weakness may devel-

op. Delayed sexual maturation or losses of menstruation, libido, and potency frequently occur.

Severe weight loss and stunting (resulting from malnutrition, losses of growth hormone, or both) take place in approximately one out of four affected individuals. Alternatively, the person may become unduly overweight. Thus, up to one-half of the children meet the criteria for the Babinski–Fröhlich syndrome (hypothalamic injury with obesity and genital underdevelopment). Diabetes insipidus, pointing to damage of the hypothalamus, the supra-optic hypophyseal tract, or the posterior pituitary, appears in a minority. In addition to the deficits of pituitary gonadotropins, which affect sexual maturation and function, underfunction of the thyroid and adrenal cortices may appear as a consequence of the losses of hypothalamic TSH and ACTH releasors or of pituitary TSH and ACTH production.

POSTERIOR PITUITARY

Particular areas in the hypothalamus, the supra-optic nuclei, elaborate the polypeptide neurohormone vasopressin (antidiuretic hormone, or ADH). This substance moves down the supra-optic hypophyseal tract and is stored in the posterior pituitary gland.

The hypothalamus also produces oxytocin, the chief function of which is the stimulation of uterine contractions during labour.

Action of vasopressin

The absorption of water back into the body fluids occurs in the kidney during the formation of urine from the blood plasma. This process is dependent upon vasopressin. When body water must be conserved, vasopressin enhances the ability of the kidney tubules to reabsorb water and enables them to excrete urea, salts, and other metabolic end products in a small volume of concentrated urine. On the other hand, decreases in vasopressin, when excesses of body water are present, permit the excretion of the extra water as a dilute urine. When vasopressin is lacking or ineffective, excessively large volumes of body water are lost in urine. This is termed diabetes insipidus (a flowing through of a nonsweet—*i.e.*, sugar-free—urine). The thirst that develops then leads to an increased intake of aqueous fluids. When diabetes insipidus is present, any increase in the load of solutes to be excreted by the kidney increases the volume of urine. On the other hand, decreases in the solute load by starvation, salt restriction, and reduced liquid intake diminish the urine excesses.

Diabetes insipidus is to be differentiated from psychogenic thirst—*i.e.*, potomania, or a sensation of thirst unrelated to a shortage of water—in which the urine volumes are increased secondary to an increased water intake. Increased urine volumes mimicking diabetes insipidus may also occur in diabetes mellitus, in kidney disease, during discharge of excess body fluids, and in a number of other conditions.

Diabetes insipidus that results from injury to or destruction of the hypothalamic nucleus (where vasopressin is produced), the tract by which vasopressin is conveyed to the posterior pituitary gland, or the posterior pituitary (neurohypophysis; where it is stored), is called supra-optic hypophyseal diabetes insipidus, or SHDI. These conditions can be congenital, familial, or acquired defects produced by brain tumours, injury, or disease. When the supplies of vasopressin are adequate but the kidney tubules are unresponsive—either genetically or because of potassium depletion, high levels of serum calcium, or other disorders—nephrogenic diabetes insipidus (NDI) results. The genetic type is familial and may be fully or partly developed; it may be present in newborn infants and, if untreated, results in mental retardation or death by dehydration.

SHDI can be corrected by injections of vasopressin-like molecules of animal or synthetic origin; such treatment, however, is ineffective in the nephrogenic form of the disorder. In both types, reduction of the solute load presented for excretion to the kidneys or the administration of certain antidiuretic drugs that influence salt and water excretion by the kidneys reduces the volume of urine.

II. The thyroid

The thyroid gland, two shield-shaped structures joined by an isthmus overlying the larynx (Adam's apple) is the source of the two currently recognized thyroid hormones, tetraiodothyronine or thyroxine (T_4) and triiodothyronine (T_3), which participate in the overall control of energy utilization. The thyroid gland is also the site of production of thyrocalcitonin, a hormone secreted by the clear or C cells, which buffers against undue increases in serum calcium.

The thyroid hormones: thyroxine, triiodothyronine, and thyrocalcitonin

EXCESSES OF THYROID HORMONES

Hyperthyroidism. Overactivity of the thyroid gland (hyperthyroidism, thyrotoxicosis, or toxic goitre) with diffuse enlargement of the thyroid gland and excesses of thyroidal hormones (T_4 and T_3, or just T_3 alone) often accompanied by eye protrusion (exophthalmos, or Graves' disease) is an auto-immune disorder. It arises from long-acting thyroid stimulator (LATS), a gamma globulin probably produced as an antibody to thyroid tissue. LATS is unrelated to TSH. Toxic nodular goitre, or Plummer's disease, results from hyperfunction of thyroid adenomata, or tumours. It usually occurs without eye signs, but the eyes often appear larger because of a stare. Thyrotoxicosis may also result from an ovarian struma (a goitre-like, iodine-containing tumour) that secretes thyroid hormones; from a testis, placental, or other tumour that is secreting TSH-like material that stimulates the thyroid gland; or from ingestion of huge amounts of thyroid hormone.

Although hyperactive behaviour, thyroid enlargement, smooth and hot skin, excessive sweating, tremors, rapid pulse, and eye signs are the most frequent features of thyrotoxicosis, virtually any organ or tissue may be involved.

Thyrotoxicosis

Thus, with respect to the central nervous system, emotional and personality disturbances, abnormal brain waves and nerve–muscle transmission, or convulsions may occasionally develop in thyrotoxicosis. Rapid contraction and relaxation of tendon reflexes are frequent.

The circulatory changes in hyperthyroidism can be depicted as a sequence: increased heat production and heat loss, dilatation of blood vessels, decreased arterial tone, increased movement of blood, greater output of blood by the heart, and rise in blood pressure. Blood-plasma volume is probably increased, and extracellular fluid may be, also; insensible loss of fluid is clearly greater than normal. Heart disease associated with hyperthyroidism may be characterized by totally irregular heart action (atrial fibrillation) and congestive failure. These heart conditions are more common in older persons, affecting one-half or more of those over 40 years of age. They may occur, however, in younger individuals with apparently normal hearts. Angina pectoris, a constricting type of chest pain usually associated with a shortage of blood supply and oxygen in the heart muscle is common in thyrotoxicosis, but damage to the heart muscle in these cases is unusual. The elasticity of the lungs and the respiratory excursions may be decreased in hyperthyroidism. Blood flow through the kidneys and various kidney functions are often increased. The intake of food is greater, and food starts to leave the stomach sooner, but emptying of this organ takes longer than normal. The absorption and movement of food through the gut are generally accelerated, and diarrhea may develop. Liver glycogen stores are usually decreased, and infiltration with fat and cells may be present. Weakness of the eye muscles is frequent, and acute or chronic weakness of other muscles may develop. Also, intermittent weakness of the periodic paralysis type, or progressive decrease in muscle strength (myasthenia gravis) as the day goes on, can occur. Thinning of bone (osteoporosis), localized destruction of bone (osteitis fibrosa), and inadequate bone salt deposition (osteomalacia) may take place. In a few individuals, marked clubbing of the fingers, increases in periosteal or surface bone, and arthritic symptoms develop.

Endocrine glands other than the thyroid may be affected when excesses of thyroid hormones are present. Occasionally, urine volumes and thirst are excessive, probably

Effects of hyperthyroidism on other endocrine glands

because increased serum calcium interferes with the action of vasopressin, one of the hormones that regulate water excretion by the kidney. Growth hormone responses to low blood-sugar levels may be diminished. Skin pigmentation increases in some individuals because of increases in ACTH or MSH or in the responses to them. Alternatively, it is possible that the increased pigmentation results from the higher skin temperatures and the increased heat transfers. Adrenocortical function is increased when the thyroid is overactive. This compensates for the shortened survival time of hydrocortisone when thyroid hormone excesses are present. Blood pressure and pulse responses to adrenal medullary hormones of the adrenaline type (epinephrine and norepinephrine) are increased in hyperthyroidism. Loss of menstruation is common when thyroid overactivity is severe, but pregnancy can occur. During such pregnancies, the hyperthyroidism is frequently accentuated, the incidence of toxemia of pregnancy rises, and that of fetal mortality increases. Though, as mentioned, bone mass tends to decrease in hyperthyroidism as a result of thinning and localized destruction of bone, and perhaps deficits of bone salt, the serum inorganic phosphorus generally tends to increase slightly, and serum calcium may increase. When thyroid hormone supplies are excessive, insulin responses to sugar are generally normal, but the blood-sugar levels following glucose may be higher and more prolonged. The frequency of actual diabetes mellitus is increased, especially in older persons; pre-existent diabetes is made more severe, and insulin is probably destroyed more rapidly.

Laboratory diagnosis of hyperthyroidism. A diagnosis of thyroid overactivity is supported by increased oxygen consumption—*i.e.*, elevated basal metabolic rate (BMR) readings—and especially by increases in the levels of serum protein-bound iodine (PBI), butanol-extractable iodine (BEI), thyroxine (T_4) and triiodothyronine (T_3), and the thyroidal uptake of radioactive iodine. Indices of lesser value include the binding of radioactive T_3 to blood cells or resin in vitro, ankle reflex time, and miscellaneous other measurements. No single test is consistently infallible, and, hence, a battery of several is usually employed. Evaluation of the significance of such tests is aided by the history, physical findings, and the responses to therapy.

The BMR is above +15 percent in four-fifths or more instances of thyroid overactivity; it is normal in the rest. High BMR values, however, also occur in illnesses that do not affect thyroid function.

The serum PBI can be a reliable index of thyroid hormone excesses. A value above 7.5 or 8.0 gammas per 100 millilitres of blood serum uncontaminated by iodine compounds is virtually diagnostic. The serum PBI, however, is unduly high during an intake of inorganic or organic iodides; in estrogen excesses; during pregnancy or with use of many of the birth-control pills; in certain forms of goitrous cretinism; some cases of thyroiditis or hepatitis, porphyria, liver disease; during therapy with certain tranquillizer drugs; and after therapeutic dosages of radioactive iodine. The PBI may be below normal during T_3 therapy, when plasma thyroxine binding protein is deficient, or during treatment with androgens or diphenylhydantoin, an anti-epileptic drug. The BEI fails to differentiate organic iodides such as gallbladder and other dyes from circulating T_4 but is not affected by inorganic iodides. It is particularly useful in goitrous cretinism, in which iodinated tyrosines may be released into the circulation. Serum T_4 and T_3 measured by chromotographic techniques identify thyroid overactivity when the serum PBI and BEI fail to reflect the levels of thyroid hormones in the circulation.

The thyroidal uptake of radioactive iodine, when elevated above 40 percent at 24 hours, is also a highly reliable index of thyrotoxicosis. The uptake may be normal, however, despite the existence of an overactive thyroid, and high values occur in iodine deficiency and in certain types of thyroid failure. Refinements of the radioactive iodine uptake test prove useful in equivocal circumstances. Thus, high uptake rates that can be sharply reduced by perchlorate, thiocyanate, or related ions point to defective binding of the trapped iodine to thyroid protein, as in thyroiditis or as in the presence of thyroid blocking agents. Also, when the thyroid is overactive, the residual thyroidal radioactivity is generally low for one week following the test dose of the isotope, whereas it tends to be high in thyroid failure and iodine deficiency. Failure to suppress thyroidal iodine uptake by desiccated thyroid, T_4, or preferably by T_3, is also a useful indication of hyperthyroidism. A high conversion ratio—*i.e.*, the percentage of radioactive iodine present in blood plasma or serum as radioactive PBI two or more days after administration of the isotope—is characteristic of thyroid hyperfunction. It is not specific, however, because similar high values may occur in myxedema, thyroiditis, iodine deficiency, goitrous cretinism, nephrotic syndrome, and in children under the age of five. The removal of iodine from the circulation by the thyroid can be calculated easily and serve as an index of hyperthyroidism.

Diagnostic tests for hyperthyroidism

The binding of radioactive T_3 to a sample of blood or to resin correlates moderately well with the thyroid status. This test may be used when excesses of organic dyes interfere with thyroidal uptake, PBI, or BEI measurements. False low binding of T_3 to blood cells or resin, however, is produced during pregnancy by estrogen excesses because of increased binding of T_3 to the proteins of blood plasma. Also, false elevations in T_3 binding occur in a large variety of nonthyroidal disorders.

Though tendon reflexes are more rapid in hyperthyroidism, measurements of the Achilles reflex time may not demonstrate values outside the normal range.

Other indices employed in the past to support a diagnosis of hyperthyroidism are much less useful. Low serum cholesterol levels, for example, are not regularly present; galactose or xylose tolerance tests are not consistently abnormal; decreased ultrafiltrable magnesium levels have not proved to be a consistent indicator; but total magnesium and radiomagnesium studies are more useful. Under appropriate circumstances, the presence of creatine in the urine of adult males may be a clue to the existence of hyperthyroidism, but this also occurs in various muscle diseases.

A number of less specific and, therefore, nondiagnostic laboratory findings occur in hyperthyroidism. Thus, routine blood studies may reveal an increased lymphocyte percentage and a lowered red blood cell count. The latter is largely a reflection of an expanded blood-plasma volume, but the total red cell mass may also be diminished. In general, the blood sugar, urea, nonprotein nitrogen, and the serum electrolytes remain within normal limits in hyperthyroidism. As indicated earlier, the serum inorganic phosphorus may be slightly increased. Quite regularly, the serum proteins and especially the serum albumin tend to be low in thyroid overactivity, whereas gamma globulins and plasma mucoproteins may be increased. Occasionally, serum calcium may be increased, and calcium may be deposited in the kidneys. Rarely, the serum potassium may be low. The total potassium of the body may also be decreased. This is more likely to occur in men than in women. Chloride and sodium levels in sweat may be low. Urinary and fecal losses of calcium are generally increased in thyrotoxicosis.

Treatment of hyperthyroidism. The selection of a particular treatment for an overactive thyroid—*i.e.*, use of iodide–iodine solution or other thyroid blocking drugs, surgical removal, destruction by radioactive iodine, or a combination of all of these—is determined by a number of variables. These include the age and sex of the person, associated disorders, the size and shape (*e.g.*, it can be nodular or asymmetrical) of the gland, and the ability of the individual to cooperate.

Iodide–iodine treatment

Inorganic iodide and elemental iodine tend to suppress thyroid overfunction. The net result is the same whether one or the other is given. Iodide–iodine is probably more effective in diffuse than in nodular goitres. Their potency, however, is weak compared to that of blocking agents of the thiocyanate, thiouracil, and associated chemical type. Iodide–iodine therapy given as the sole treatment proves

effective in only about 20 percent of the patients. The response is prompt, however, appearing in a few days, and therefore more rapid than that obtained with most of the thyroid blocking drugs or with radioactive iodine. Excesses of iodide or iodine block the binding of this halide to protein within the gland, an effect that interferes with hormone synthesis. The latter presumably is also the basis for the origin of goitres and even of myxedema in persons receiving inorganic iodide, elemental iodine, or related agents for prolonged periods. On rare occasions, iodide–iodine administration in the usual dosages may produce or appear to produce an increase in the severity (exacerbation) of thyrotoxicosis. The notion that iodide–iodine solutions exert effects that last for only a limited period and then cease is erroneous. The action continues, but variations in the intensity of the underlying thyrotoxicosis may give that impression. Also, the view that iodide–iodine solutions interfere with the responses to thyroid blocking agents such as thiouracil is not correct, although they will block the effects of thiocyanate, nitrate, and perchlorate (see below). In fact, inorganic iodide hastens the response to propyl thiouracil and related drugs. Skin eruptions, fever, enlargement of the salivary glands, and respiratory symptoms as a result of increased pulmonary and bronchial secretions may occur during treatment with inorganic iodide or elemental iodine.

Two classes of thyroid blocking agents

In addition to iodide and iodine, there are, as previously mentioned, two general classes of thyroid blocking agents: (1) those that interfere with the "pump" that concentrates iodide in the thyroid cells (*e.g.*, thiocyanate, perchlorate, or nitrate) and (2) those that prevent the binding of iodine within the thyroid (*e.g.*, propylthiouracil with a thiocarbamide grouping). These two classes of thyroid blocking compounds ultimately produce hypothyroidism and a goitre. In the case of the thiocyanate group, the inhibition of the thyroid iodide pump and the subsequent effects can be entirely overcome by excesses of inorganic iodide. With the thiouracil drugs, inorganic iodide and elemental iodine minimize the size and blood supply of the gland but do not prevent or cancel the goitre. In animals, the goitre-producing effects of both classes of compounds can be entirely cancelled by giving thyroid hormones or by hypophysectomy, indicating that enlargement of the gland results from excesses of thyroid-stimulating hormone (TSH) produced by the anterior pituitary.

The thiocyanate drugs have had only limited use in the control of thyrotoxicosis but are of value in selected circumstances. The propylthiouracil and related drugs have been widely employed in the treatment of hyperthyroidism. Experience has been most extensive, and toxicity the lowest, in the case of propylthiouracil. Undesirable side effects have been less than 1 percent. Individuals with diffuse toxic goitres respond more readily than do those with toxic adenomas. During such therapy, hypothyroidism can be avoided by reduction of the dosage or, more simply, by taking full replacement dosages of desiccated thyroid, or equivalent amounts of synthetic thyroid hormones (T_4 and T_3).

Radioactive iodine therapy

Radioactive iodine (I-131) therapy is generally reserved for adults and for those who, at any age, are poor surgical risks, who do not wish to submit to surgical treatment, who have been operated upon previously, or in whom adequate long-term observation during treatment with thyroid blocking agents is not possible. The radioactive iodine is usually given in sessions spaced two or more months apart. The effects of a particular dosage of radioactive iodine may not be maximal for many weeks. Hence, the thyroid status is re-evaluated at intervals following radioactive iodine administration and a decision reached concerning further therapy. Such treatment with radioactive iodine will also reduce the size of disfiguring toxic goitres. Control of the hyperthyroidism with radioactive iodine treatment is more readily accomplished in diffuse toxic goitres than in nodular goitres. Replacement of thyroid hormones is usually prescribed when the hyperthyroidism begins to subside, thus preventing the development of post-treatment myxedema. Radioactive iodine cannot be administered beyond the tenth week of pregnancy because it destroys the fetal thyroid and produces cretinism.

Partial thyroidectomy after the hyperthyroidism has been controlled by thyroid blocking agents is often performed on younger thyrotoxic individuals in whom a permanent remission is desired and especially in those with symptoms produced by the enlarged thyroid. Treatment of hyperthyroidism by surgical removal of a part of the thyroid gland (*i.e.*, by subtotal thyroidectomy) has been brought to a high state of technical perfection. The operative mortality in many institutions is less than 1 percent, and surgical complications (*e.g.*, removal of parathyroid glands) are equally infrequent. To avoid possible shortages, thyroid hormones in replacement dosages are often prescribed following this surgical treatment.

Hyperthyroidism concurrent with pregnancy can be controlled with inorganic iodide or elemental iodine, with thyroid blocking agents alone, or with these agents followed by subtotal thyroidectomy. Every attempt is made to minimize suppression of fetal thyroid function by using the smallest possible dosages of blocking agents. As already noted, radioactive iodine cannot be administered beyond the tenth week of pregnancy because it will be concentrated in and destroy the fetal thyroid.

Thyroid overactivity in children occurs above the age of ten years about 80 percent of the time, but it may be present at birth. As in the adult population, the condition is more prevalent among females than males; weight loss occurs in most of the children, and nervousness, tremor, rapid heart action, and thyroid enlargement are almost always evident. Eye protrusion (exophthalmos) develops in about one-half the cases. The blood pressure may increase. If the overactivity of the thyroid has been present for six months or longer, the child may be unusually tall. In contrast to older age groups, however, exophthalmos is less severe, and the enlargement of the thyroid is diffuse rather than nodular.

Therapy is often confined to thyroid blocking agents of the propylthiouracil type, because in many children a permanent remission can be obtained after several years of treatment. To avoid recurrences and hypothyroidism, and perhaps to cancel or minimize any tendency to the development or progression of exophthalmos, full replacement dosages of thyroid hormones may be given with the propylthiouracil. Such a program prevents exacerbation of the thyroid overactivity and assures a constant normal level of circulating thyroid hormones. Partial surgical removal of the thyroid following production of a remission with thyroid blocking agents is an effective alternative. Treatment with radioactive iodine is generally not used in children because of the question of ultimate effects on the thyroid gland. When circumstances dictate this treatment, it is as effective as in adults.

Thyroid crises. Thyroid crises, or storms, are variably characterized by increases in heart rate, fever, nausea, vomiting, diarrhea, stupor, and coma. They occur spontaneously or following various diagnostic or treatment procedures in persons with incompletely controlled hyperthyroidism. Such thyroid storms can also be produced by taking large quantities of thyroid hormone. These are serious medical emergencies, and failure of the heart, kidneys, adrenals, or the liver and circulatory collapse and death may ensue.

Treatment of thyroid storms

Treatment is based on all potentially beneficial measures, since the mortality, though reduced from its previously high level of virtually 100 percent, may still be as high as 25 percent. Attention is therefore directed toward (1) blocking further release of thyroid hormones; (2) cancelling some of the effects of thyroid hormone excesses by such drugs as reserpine and guanethidine (tranquillizers and anti-hypertensive agents); (3) increasing the supply of protective cortisol-type steroids; (4) treatment of congestive heart failure, shock, dehydration, and electrolyte depletion; and (5) removal of the precipitating cause.

DEFICIENCIES OF THYROID HORMONES

Myxedema and hypothyroidism. The terms myxedema and hypothyroidism refer to thyroid underfunction in

adults or older children with resultant deficits of the two thyroid hormones, thyroxine (or T_4) and triiodothyronine (or T_3). Such underfunction may originate in the thyroid itself (primary type) or in the pituitary or hypothalamus, which control the thyroid (secondary type).

A rough, dry, and yellow skin, coarse scalp hair, and hoarse voice suggest that thyroid hormones may be deficient, but none of these signs need be present. Overweight is not necessarily a characteristic feature. Intellectual functions may become impaired, and personality changes or severe emotional illnesses, including so-called myxedema madness and coma, may appear. The nervous system may also be affected in other ways. Thus, disturbances of gait, of coordination, and of rapid movements may develop. Sensations may be altered and hearing can decrease. Tendon reflexes and brain waves are usually slowed.

Exophthalmos in association with myxedema occurs only in the primary type but is not always present. Drooping of the upper eye lids, often detectable only by measurements before and after thyroid replacement therapy, is frequent. Thyroid underactivity is also accompanied by a constriction of blood arterioles, increased resistance to blood flow, and prolonged circulation time. At times, the heart chambers may enlarge, the output of blood by the heart may decrease, and congestive heart failure may appear. The blood flow to the kidney and various kidney functions diminish, but not to critical degrees. Urinary bladder tone is decreased. In far-advanced thyroid underfunction, breathing may become shallow and slow and may result in carbon dioxide retention (respiratory acidosis) and coma. Body water and extracellular fluid volumes increase during prolonged shortages of T_4 and T_3, whereas the blood-plasma volume diminishes. Abnormal collections of fluid can appear in the extremities, in the abdomen, and in other body cavities. The secretion of gastric juice diminishes, the intestinal absorption of foodstuffs is delayed or reduced, and constipation is frequent. Skeletal muscles may become enlarged and firm as a result of thyroid hormone shortages and of relaxation of muscles after contraction becomes prolonged. The maturation of the skeleton may be affected. Thus, the normal fusion of the growing ends of the bone to the bone shaft may be delayed, and arthritic symptoms reversible by thyroid therapy can occur.

Effect of myxedema on other endocrine glands

Endocrine glands other than the thyroid may be affected by myxedema. Thus, excesses of the antidiuretic hormone (vasopressin) in myxedema may produce low blood sodium (hyponatremia). Other endocrine glands may be affected by the shortage of thyroid hormone: the gonadotropins, ACTH, and growth hormone may be decreased. Ovarian dysfunction in myxedema is different in the primary and secondary types. In the former, menstrual flow may be more frequent, prolonged, and excessive. In the latter, if associated losses of gonadotropins are present, the menses cease. In the male, sexual activity and sperm production decrease. Fertility is reduced in both sexes. Myxedema may also be responsible for delayed sexual maturation (which can be corrected by thyroid replacement), but sexual precocity can also occur. Adrenocortical function may become decreased in myxedema either because of loss of pituitary ACTH, a direct effect of hypometabolism on the adrenals, or as a consequence of decreased utilization of, and therefore decreased need for, cortisol. These are temporary changes. At times, the myxedema is accompanied by permanent primary hypofunction of the adrenal cortices (Addison's disease). This has been called Schmidt's syndrome. Also, circulatory responses to epinephrine (adrenaline) secreted by the adrenal medulla are decreased in hypothyroidism. Finally, it has been suggested that myxedema protects against the development of diabetes mellitus, but the two entities can and do occur together.

Laboratory diagnosis of myxedema. The laboratory diagnosis of primary or secondary myxedema requires the demonstration of decreased thyroid function with a resultant shortage of thyroid hormones. The more reliable of the readily available tests are based on the demonstration of a decreased thyroidal uptake of administered radioactive iodine and low level of T_4 in the blood serum. The T_4 can be measured as such or as the serum protein-bound iodine (PBI). Less reliable and often less readily available or more cumbersome indices include decreased oxygen consumption or basal metabolic rate (BMR), prolongation of the ankle reflex time, increased serum cholesterol, increased binding of radioactive T_3 to serum proteins, decreased conversion of radioactive iodine to radioactive thyroid hormones, increased secretion of inorganic iodide in the saliva, delayed absorption of glucose or other sugars from the gut, decreased creatine in serum and urine, and increased total or ultrafiltrable magnesium in serum. The laboratory differentiation of primary and secondary myxedema depends upon the administration of TSH of animal origin.

Use of routine laboratory tests

Routine laboratory tests may show nonspecific changes during shortages of thyroid hormone. Voltage may be low in heart tracings, for example, or anemia may develop; blood sugar and the serum sodium and inorganic phosphorus may be lowered, while serum cholesterol, calcium, total carbon dioxide content, total protein, gamma globulins, uric acid, and spinal fluid proteins may be raised. The fecal and urinary excretion of calcium is generally decreased; serum alkaline phosphatase tends to be low; and serum creatine phosphokinase is increased.

Therapy of hypothyroidism and myxedema. Desiccated thyroid of animal origin, a mixture of thyroid protein, T_4, T_3, and other iodine compounds, has been extensively employed in the therapy of primary or secondary myxedema. Synthetic thyroid hormones, such as sodium L-thyroxine (T_4) or sodium L-triiodothyronine (T_3), may also be used, but when used alone in full replacement amounts, T_4 produces undue rises in the serum PBI and T_4 values, and T_3 depresses them. These limitations are not encountered when mixtures of T_4 and T_3 in a ratio of approximately 4:1 are employed. In treating hypothyroidism or myxedema, particular care is taken to begin with small dosages, lest adrenocortical insufficiency, heart pain, myocardial infarction, or congestive heart failure be produced; dosage is increased in gradual increments until the full replacement dosages are reached. TSH derived from animals cannot be used to treat myxedema of pituitary origin, because, with continued administration, the human thyroid ceases to respond to this stimulus.

The rapid onset and offset of T_3 effects are of special value in the treatment of life-threatening myxedema coma, myxedema heart disease, or in preparing a hypothyroid person for emergency surgery. In such emergencies, the T_3 must be administered in substantial dosages at intervals of several hours, rather than in the small amounts previously described. The use of these relatively large dosages of triiodothyronine is generally accompanied by the administration of cortisol steroids.

Cretinism and juvenile myxedema. Cretinism is hypothyroidism or myxedema that begins before birth or during infancy or childhood and results in bodily stunting and intellectual retardation. Such thyroid underfunction may be primary in the thyroid or secondary to a deficit of pituitary TSH or hypothalamic TSH releasor. Enlargement of the thyroid (*i.e.*, goitres) may or may not be present in the primary type but is never present in the secondary type. Goitres represent an attempt to restore failing thyroid function to normal by greater output of TSH, which increases the size of the gland. The failure of thyroid function can result from a lack of inorganic iodide or elemental iodine in foods, or increased losses via feces, ingestion of certain foods containing thyroid blocking agents, inflammatory conditions of the thyroiditis type, exposure to radioactive iodine, and apparently spontaneous enzymatic defects. Currently recognized enzymatic defects within the thyroid that can result in a goitre consist of inability of the thyroid to: (1) trap iodide, (2) bind iodine to thyroid protein, (3) deiodinate iodotyrosines, (4) couple iodotyrosines to form T_3 and T_4, and (5) produce and release normal thyroid hormones.

Symptoms and signs of infantile cretinism that appear at birth or several weeks thereafter include: poor ap-

petite, breathing difficulties, low body temperature, dry and thickened skin, and decreased vigour and activity.

Symptoms and signs of cretinism

During the first three to five months of age, the following signs may be indicative of cretinism: delayed union of skull bones, torpid behaviour, slow nursing, spells of choking and cyanosis (*e.g.*, turning blue of the skin) during feeding, excessive sleepiness, hoarseness, constipation, a pallid, cold, gray-mottled or yellow-tinged skin, enlargement of the tongue, umbilical herniation, flabby musculature, a pulse rate of less than 100, or rectal temperatures of 98° F (37° C) or lower. When cretinism develops in the later months of the first year of life or subsequently, any of the above may occur as well as delay or reversal of such activities as sitting, crawling, walking, talking, or head lifting. Also, delays in teething, failure to grow in length at a normal rate, excessive body and facial hair, increased neck or hip fat, or a low pulse pressure may be present. In cretins who remain undiagnosed during the early years of life, the fully developed picture becomes evident. This consists of intolerance to cold, decreased sweating, shortened stature, infantile body proportions with short legs and arms, depression of the bridge of the nose, enlargement of the tongue, retarded behaviour, thin, dry, and scaling skin with "dirty" knees and elbows, slow relaxation of tendon reflexes, and the other symptoms and signs mentioned previously.

Juvenile hypothyroidism and myxedema differ from cretinism in that thyroid hormone supplies, growth, and intellectual development have been normal during infancy and early childhood. Such juvenile hypothyroidism can be spontaneous or may follow surgical removal of the thyroid gland. The shortage of thyroid hormones results in dryness of the skin, delayed eruption and deformation of the teeth, increased dental cavities, excessive body and facial hair, brittle scalp hair, and increased muscle size. Growth may be retarded, but body proportions are not infantile. Although intellectual performance may be decreased, it is not so profound as in cretinism, and, more importantly, it is fully reversible. Sexual maturation may be retarded; occasionally, sexual precocity develops.

Among the nonspecific laboratory indices that may be suggestive of cretinism, juvenile hypothyroidism, or myxedema are the following: anemia, increased blood sedimentation rate, high serum calcium after a calcium load, low serum inorganic phosphorus, low fasting blood sugar, low serum alkaline phosphatase, carotene in the plasma, prolongation of the jaundice of newborns (this jaundice, present in about 70 percent of newborns, usually disappears in about 10 days), decreased urine creatine, lengthening of the Q-T interval of the electrocardiogram, abnormal brain waves, low BMR values, prolonged contraction and relaxation of tendon jerks, delayed skeletal maturation and abnormal X-ray patterns of conversion of cartilage to bone with the development of epiphyseal dysgenesis, and enlargement of the pituitary fossa.

Diagnosis of thyroid underfunction

The diagnosis is usually established by demonstration of low serum levels of PBI and thyroxine and reduced thyroidal uptake of radioactive iodine. In some of the enzymatic disorders, however, serum PBI may be normal or high because abnormal thyroid proteins are being produced and released, but the serum T_4 remains low. Also, thyroidal uptake of radioactive iodine may be misleadingly high at times, but appropriate tests can then establish that the normal conversion of such trapped iodine to thyroid hormones is not occurring.

Prompt diagnosis of thyroid underfunction is essential in infants and children. Adequate replacement with adult dosages of thyroid hormones will prevent further retardation of physical and intellectual development.

Women with a history of previous cretin births or those who have received radioactive iodine or thyroid blocking drugs during pregnancy may be treated with T_3 or other thyroid hormones in the hope of preventing cretinism.

Metabolic insufficiency. Low BMR readings without deficits of thyroid hormones have been described under a variety of terms including: hypometabolism without hypothyroidism, non-myxedematous hypometabolism, pseudohypothyroidism, euthyroid hypometabolism, or metabolic insufficiency.

It has been suggested that such persons suffer from fatigue early in the day, cold sensitivity, impaired intellectual activity, emotional depression, headache, pains and aches, decreased sexual activity, scanty menstruation, and infertility. Obesity, dry nails and hair, tenderness of the muscles, generalized puffiness or edema, slowing of the Achilles tendon reflex, and decreased sperm production or ovulation are stated to be frequent accompaniments. The thyroidal uptake of radioactive iodine and the levels of serum PBI and T_4, however, are within the normal range; serum cholesterol may be normal or elevated.

It has been suggested that therapy with desiccated thyroid or with T_4 in amounts ordinarily adequate to produce a remission in myxedema fails to restore the BMR to normal and does not relieve the above symptoms and signs. Observations that T_3 administration appeared to alleviate the condition have led to the suggestion that hypometabolism without hypothyroidism results from inadequate responses to T_4 or from inadequate conversion of T_4 to T_3.

INFLAMMATORY DISORDERS: THYROIDITIS

Thyroiditis, or inflammation of the thyroid gland, can be acute, subacute, or chronic, but this does not necessarily mean that each of these is a separate and distinct entity. Such a classification merely emphasizes the onset and duration of the symptoms and signs of thyroiditis and relates them to the microscopic appearance of the gland.

Acute, subacute, and chronic thyroiditis

Before the advent of agents for treating infections, acute thyroiditis generally represented a complication of a systemic infection such as typhoid fever or pneumonia. The infecting organism could usually be isolated from the blood or other sites as well as from the gland itself. The death rate was significant, although not necessarily related to the thyroiditis. In current experience, bacterial infection and abscess formation are unusual in acute thyroiditis, either because treatment with antibiotics and related drugs is started early or because the infective agent is frequently a virus. The manifestations of acute thyroiditis include fever, malaise, and pain in the front of the neck particularly on swallowing. The thyroidal uptake of radioactive iodide may be low and uneven, and serum protein-bound iodine (PBI) may be elevated. Despite the latter, circulating thyroxine, or T_4, is not increased. Acute thyroiditis responds promptly to adrenocortical steroid and antibacterial therapy. There is no indication for the use of thyroid blocking agents, unless true thyroid overactivity is present.

Subacute thyroiditis—*i.e.*, a lingering thyroiditis of insidious onset and duration—is correlated with a granulomatous change in the thyroid in which giant cells may be present. This condition has been termed pseudotuberculous, or De Quervain's thyroiditis. Nonspecific symptoms and signs such as fatigue, low-grade fever, and anemia are frequent. The origins of subacute thyroiditis are not established beyond the fact that auto-immunization may play a role. Such persons may, however, have a high level of antibody against the mumps virus, and the virus itself has been isolated from the thyroid glands. The serum PBI in subacute thyroiditis may be normal, low, or high, and the uptake of radioactive iodine by the thyroid is often uneven. Before the demonstration that desiccated thyroid and adrenocortical steroids in adequate amounts produce a remission, the local and systemic complaints of subacute thyroiditis lasted for months or even years.

Many but not all instances of chronic thyroiditis can fall into either the Hashimoto (struma lymphomatosa) type, in which the gland is heavily infiltrated with lymphocytes, or the Riedel (struma fibrosa) type, in which scar tissue replaces most of the gland. These microscopic differences and the natural history of the two forms of thyroiditis originally led to the suggestion that they are separate entities. At present the justifications for such categories seem less valid, since chronic thyroiditis may not meet the histological criteria for either the Hashimoto or Riedel types and must therefore be called nonspecific. Hashimoto's thyroiditis may be familial or hereditary. Disturbances of the thyroidal uptake of iodide

and of the serum PBI levels may be present in various combinations in chronic thyroiditis. Myxedema is especially likely to occur in the Hashimoto type. As with subacute thyroiditis, these chronic types decrease in size with pharmacological dosages of adrenocortical steroids and with replacement thyroid hormone therapy. Reversal of the Riedel type of fibrosis cannot be expected, but X-ray irradiation may reduce the size of the gland. Surgical removal of the enlarged and scarred gland may be necessary for relief of obstruction.

Thyroiditis as an auto-immune reaction

Antibodies to the thyroglobulin or to other components of the thyroid occur in the serum of individuals with subacute, Hashimoto, Riedel, and nonspecific thyroiditis. This suggests that the various types of thyroiditis represent an auto-immune reaction. A predisposition to auto-immunization in such patients is supported by the association of thyroiditis with Addison's disease, various collagen disorders, hemolytic anemia, pernicious anemia, and perhaps myasthenia gravis and cirrhosis, entities in which sensitization of the individual to his own tissues plays a role. Hence, a diagnosis of thyroiditis is supported by the demonstration of circulating antibodies to thyroglobulin or the thyroid gland, but these may also be present in other types of goitres, in thyroid overactivity, myxedema, cretinism, and cancer. Proof of thyroiditis, therefore, must be based on a microscopic examination of thyroid tissue.

TUMOURS

Tumours and cancers of the thyroid gland. Ninety percent of the noncancerous tumours (adenomas) of the thyroid are of the follicular type. The remainder fall into the embryonal, fetal, simple, colloid, Hürthle cell, or papillary types, or into an unclassified category. If the enlargement is nodular and goitre is not prevalent in that geographic area, the possibility of cancer is increased. Cancer of the thyroid makes up slightly more than 0.5 percent of all cancers and is responsible for slightly less than 0.5 percent of cancer deaths among adults.

Although it may affect children and the aged, cancer of the thyroid is largely a disease of middle life. As with all thyroid disorders, women are affected about twice as often as men, but observations on this point vary. Preexisting thyroid nodules and X-ray irradiation of the thymus, neck, face, and other nearby areas predispose to thyroid cancer. Cancer of the medullary type in association with another tumour arising in the adrenal medulla may occur in families. Medullary cancers arise from cells that produce thyrocalcitonin, a hormone that counteracts parathyroid hormone effects in bones.

Cancer of the thyroid is uncommon in childhood, making up about 4 to 6 percent of cancers in that age group. Females outnumber males in a ratio of 2:1. The peak rates in both sexes occur between the ages of seven and 14. A history of irradiation of the chest and neck area for enlargement of the thymus, acne, and other conditions is present in 20 percent or more of these cases. In children as in adults, papillary cancer of the thyroid is the most frequent; the follicular type is next in order, followed by the undifferentiated and the miscellaneous forms.

The most important symptoms and signs of thyroid cancer consist of discomfort on swallowing, anterior neck pain, and voice changes of recent origin, but none of these need be present. Also, although goitre occurs with most thyroidal cancers, the latter may exist in glands that appear grossly normal. Cancer of the thyroid, in general, spreads to the nearby lymph nodes and to the lungs, but such spread varies with the type of tumour. Extension to the skeleton, for example, occurs more often with the follicular than with the undifferentiated cancers and is rare with the papillary type.

Therapy. Localized cancers of the thyroid, particularly the papillary and follicular types, are usually removed by surgery. These well-differentiated cancers also respond well to follow-up local X-ray or to radioactive iodine administration. The undifferentiated cancers such as the small cell or giant cell types grow quickly, are more difficult to remove, do not take up radioactive iodide, and are less susceptible to external X-ray irradiation. In the miscellaneous groups, the epidermoid, the fibrosarcoma, the lymphoid, and the Hürthle cell types respond in part to surgery and to irradiation.

Persons with a history of a treated thyroid cancer are usually maintained indefinitely on full replacement dosages of desiccated thyroid or other thyroid hormones. This suppresses TSH and thereby minimizes recurrence and growth of the tumour. Five-year survival figures in thyroid cancer without regard to cell type or therapy range up to 60 percent. When the cell type is taken into account, survival is shortest in the undifferentiated—*i.e.* giant cell and small cell—cancers. The outlook is more favourable in most of the differentiated forms. Thus, five-year survivals in the papillary or the follicular type reach 70 to 90 percent or more, with the results generally better in the former. Hürthle cell cancers, because of their slow course, are believed to be relatively benign. In the other miscellaneous cancers of the thyroid, such as the epidermoid and the lymphoid, the fibrosarcomas, and the teratomas, survival usually reflects the statistics established for the individual cell types.

Survival after treatment of thyroid cancer

Prevention of cancer of the thyroid consists of limitation of X-ray of the neck and chest areas to procedures absolutely essential for diagnosis and remedial treatment, the removal of any goitres that are increasing in size or producing new symptoms, and the treatment of all noncancerous goitres, diffuse or nodular, with full replacement dosages of thyroid hormones for an indefinite period.

Cancers of nonthyroidal origin present elsewhere in the body can ultimately involve the thyroid gland. These are secondary cancers. The most common tumour that spreads to the thyroid gland from other sites is that originating in the bronchus, with cancers of the breast, stomach, lymph glands, larynx, cervix of the uterus, and esophagus next in order of frequency.

ASSOCIATED DISORDERS

Exophthalmos. Exophthalmos—*i.e.*, protrusion—and other abnormalities of the eye and surrounding structures occur in about one-half of the persons with thyroid overactivity, in a minority of those with thyroid underfunction, and, on rare occasions, in those whose thyroid function appears normal. Women are affected by exophthalmos about twice as often as men. Protrusion of the eyes, usually of a less severe degree, may also develop in Cushing's syndrome and in acromegaly. Spontaneous exophthalmos is to be differentiated from protrusion of the eye resulting from a tumour in the eye socket, clotting of blood in the venous channels within the skull, deformities of the eye socket, and from pseudoexophthalmos, in which the eyes only appear more prominent because of retracted eyelids. Such pseudoexophthalmos is common in individuals with an overactive thyroid and may occur in various illnesses such as asthma, congestive heart failure, and liver disease.

True exophthalmos may be divided into simple and malignant forms. In both, edema is frequent in the outer half of the upper lids (bulging), the lower lids, and adjacent areas. In simple exophthalmos, eye movements are grossly normal, but lid lag may be present, and conjunctival involvement is minimal or absent. Visual acuity is normal, and there is no double vision. In malignant exophthalmos, the eyes are more prominent, conjunctival swelling is marked, the eyes cannot be moved normally, and double vision is frequent; these various abnormalities are progressive, visual acuity becomes reduced, and the survival of the eyes is threatened.

Studies in humans and in experimental exophthalmos in animals have established that the protrusion of the eyes is produced by increases in the volume of tissues—muscles and cells—in the eye socket. Many theories have been advanced to explain the occurrence of exophthalmos in association with thyroid disorders. It has been suggested that exophthalmos is produced by the thyroid-stimulating hormone (TSH) or a closely related principle —*i.e.*, an exophthalmos-producing factor (EPF)—or a long-acting thyroid stimulator (LATS). As yet there is no

conclusive and consistent supporting evidence for any of these hypotheses.

Treatment with cortisol-type steroids often produces a decrease in exophthalmos of recent origin. Irradiation or removal of the pituitary or surgical removal of part of the wall of the eye socket have been successful in treating progressive exophthalmos.

Localized myxedema of the skin. In its usual form, localized myxedema is a firm swelling of the lower legs but may expand to a diffuse thickening of the skin of the legs and arms. The skin over the swelling resembles pigskin leather. It is usually accompanied by thyroid overactivity and exophthalmos but may first appear after hyperthyroidism has remitted or when post-treatment hypothyroidism has developed. Localized myxedema has been reported in thyrotoxicosis without exophthalmos and without thyroid disorder or eye abnormalities.

Features of localized myxedema

Microscopically, localized myxedema is characterized by features identical with, but more marked than, those seen in generalized myxedema. The viscid fluid causing the swelling is high in protein and acid mucopolysaccharides, including hyaluronic acid. Acid mucopolysaccharides may also be increased in unaffected portions of the skin. Localized myxedema probably results from excesses of thyroid-stimulating hormone (TSH) or related principles such as the exophthalmos-producing factor (EPF) or long-acting thyroid stimulator (LATS). As an alternative to the excess TSH, EPF, or LATS hypothesis, it has been suggested that sluggish venous blood brings about a release of hyaluronic acid from certain tissue cells with a production of edema.

Treatment of any underlying thyroid disease usually has no effect on the localized myxedema. The lesion itself is often asymptomatic and does not require therapy. If itching appears or cosmetic considerations are imperative, systemic treatment with ACTH or cortisone or localized injections of cortisol-type steroids can lead to the disappearance of localized myxedema. Success has been obtained with topical adrenal steroids and occlusion by wrapping. Injection of the enzyme hyaluronidase or of T_3 directly into the lesion or simple incision and drainage may produce partial relief. Some improvement after estrogen therapy has been reported. All forms of treatment have at times proved disappointing and surgical excision and grafting may be followed by recurrence.

Goitre. The word goitre refers to any enlargement of the thyroid. An enlarged thyroid gland, smooth or nodular, can represent, among other conditions, a nontoxic goitre, a failing or overactive thyroid, thyroiditis, a benign tumour, cancer, or involvement of this gland by tuberculosis, amyloidosis, or secondary cancer. In regions in which goitres are common, the frequency of thyroid overactivity, thyroiditis, and thyroid cancer is also increased.

Nontoxic goitres are thyroid enlargements that represent an attempt, usually successful, to restore the function of a failing gland to normal by increases in endogenous TSH. The prevalence of nontoxic goitre ranges from approximately 4 percent in the United States to as high as 85 percent in some parts of the world. The occasional cases are termed sporadic, and those that occur frequently in a particular area are called endemic. Nontoxic goitres representing a failing thyroid may result from (1) iodine–iodide deficiency, (2) marked iodine–iodide excesses, (3) defects of enzymes essential for thyroid hormone synthesis, (4) thiocyanate, perchlorate, and other agents that inhibit iodide trapping, (5) thyroid blocking compounds that interfere with the binding of iodine to protein—such as those with a thiocarbamide or related grouping, including propylthiouracil and methimazole—(6) partial destruction of the thyroid by radioactive iodine and other irradiation, (7) inflammatory disorders (thyroiditis), (8) increased urinary losses of iodide produced by high sodium chloride intake, or (9) increased fecal losses of thyroxine, as with high intake of soybeans. The possible role of variables other than iodine–iodide deficiency and environmental anti-thyroid agents in endemic goitre, including diet, water supplies, unsanitary conditions, infections, and ions such as calcium, tellurium, cobalt, fluoride, and cyanide, is incompletely defined. Heredity may play an important role in goitre, as in Pendred's syndrome (high-pitch deafness with goitre), perception deafness, and decreased binding of iodine to thyroid protein.

Causes of nontoxic goitres

Dietary and other iodine or iodide deficiency results in a shortage of thyroid hormones because this element is an essential component of T_3 and T_4. Blockade of thyroid hormone synthesis by excesses of inorganic iodide or elemental iodine is most likely to be encountered in individuals with conditions such as asthma, chronic bronchitis, or bronchiectasis who may be receiving large dosages of Lugol's solution or iodide salts. At one time that therapy was widely used, and it is possible that such goitres appeared only in susceptible individuals. The congenital (*i.e.*, inherited) and the acquired (dietary, drug inflammatory, or other origin) enzymatic defects interfere with the synthesis of thyroid hormones. Thus, any of the following may be decreased or absent: the trapping of iodine or iodide by the thyroid, the binding of trapped iodine to thyroidal tissues with iodination of tyrosine, the deiodination of excesses of monoiodotyrosine and diiodotyrosine, the coupling of iodinated tyrosines to form T_3 and T_4, or the formation of normal thyroglobulin. The rare defects in trapping are most often found among persons with a goitre and low thyroidal uptake of radioactive iodine. The binding defects are detectable by undue dumping of thyroidal radioactive iodine following potassium perchlorate or by thiocyanate administration. Thyroidal auto-immune antibodies are not a specific index of a nontoxic goitre due to thyroiditis, since they occur also in such conditions as colloid goitre, cancer, and thyroid overactivity.

Though, as indicated, most individuals with a goitre are euthyroid—*i.e.*, they secrete normal amounts of the hormone—this is not always true. In evaluating the status of the gland, the major indices of thyroid function (serum PBI and T_4 and thyroidal radioactive iodine uptake) and the minor indices (binding of T_3 to blood cells or resin, Achilles reflex time, lipid levels) must be obtained. Also, the possibility of a failing thyroid or of thyrotoxicosis should be investigated by the TSH stimulation and T_3 suppression tests, respectively.

Prevention and treatment of goitre

The prevention of goitre by administering elemental iodine or inorganic iodide is successful in the iodine deficiency type or in those that result from thiocyanate or other agents that interfere with the iodide trapping mechanisms. In goitres caused by thyroid blocking agents, such as the thiocarbamides, the administration of inorganic iodides or elemental iodine may slightly decrease the size of the gland and correct excessive vascularity. As indicated, however, marked excesses of iodide or iodine may themselves be goitrogenic, at least in certain individuals. On the other hand, desiccated thyroid in full maintenance dosages is highly effective in reducing the size of most nontoxic goitres that represent impending or actual thyroid failure. Long-standing large goitres with fibrosis may well be irreversible, however, even though the cause is removed, and those caused by cancer or tuberculosis will not respond. Also, such therapy is largely ineffective in diffuse toxic goitres because they result from LATS, a product of an auto-immune reaction to thyroid tissue.

III. The parathyroids

The parathyroids, usually four in number, are tiny glands located in the posterior part of the thyroid—either in the capsule or within the thyroid itself. They are the source of parathyroid hormone, a polypeptide that stimulates the removal of bone calcium and regulates the urinary excretion of inorganic phosphorus. Serum calcium and inorganic phosphorus levels are controlled in part by the level of parathyroid hormone. Thyrocalcitonin, a hormone elaborated by the clear cells of the thyroid, counteracts the effect of parathyroid hormone in bone and serves to reduce excesses of serum calcium to normal.

EXCESSES OF PARATHYROID HORMONE

Primary hyperparathyroidism. Primary hyperparathyroidism—excesses of parathyroid hormone without

recognized stimulus—can result from a noncancerous parathyroid adenoma, diffuse enlargement or hyperplasia of the parathyroids, or cancer of these glands, in this order of frequency. Most instances of primary hyperparathyroidism are diagnosed in persons between the ages of 30 and 70. The adenomas are multiple about 10 percent of the time, occurring predominately in females, whereas the sex distribution is about equal in adenomas of childhood and in the cancers. Parathyroid cancers, however, are rare. Increases in serum calcium and other manifestations of hyperparathyroidism have been observed with miscellaneous cancers of tissues unrelated to the parathyroid glands; such cancers produce a polypeptide with parathyroid hormone effects.

Signs and symptoms of hyperparathyroidism

The manifestations of primary hyperparathyroidism are often quite nonspecific; they may include loss of appetite, nausea, vomiting, constipation, fatigue, weight loss, and deafness of the conduction type. Personality changes, stupor, and coma may develop. Kidney stones and duodenal ulcers are common. Occasionally, acute and severe abdominal pain may appear as a result of pancreatitis. The blood pressure may increase and be followed by congestive heart failure or a cerebral hemorrhage. At times, far-advanced kidney failure may supervene.

Early in primary hyperparathyroidism, there is no X-ray evidence of bone change. With long-standing disease, bone density decreases, with loss of clarity of bone spicules or trabeculae first manifested in the shoulder blades and fingers. As the process of bone destruction (osteitis fibrosa) progresses, fractures, especially of the spine, may occur spontaneously or after minimal trauma. Rarely, increased forward curvature of the upper spine as a result of compression fractures of the vertebrae and pseudoclubbing of fingertips as a consequence of bone resorption may be present. Other X-ray findings may include deforming giant cell tumours of bone, calcification of soft tissues and cartilage, and loss of the lamina dura (the hard layer lining the alveoli) of the teeth. Conjunctival and eyelid deposits of calcium and band keratopathy may also appear, as they do whenever the serum calcium is above normal. The electrocardiogram and brain wave patterns may show changes indicative of high serum calcium levels.

High levels of serum calcium and low levels of inorganic phosphate are usually present in primary hyperparathyroidism. That disease, however, is not the sole cause of high serum calcium levels. The differential diagnosis, therefore, must include idiopathic (cause unknown) hypercalcemia of infancy, vitamin D poisoning, sarcoidosis, milk–alkali syndrome, cancers destructive of bone, tumours producing hypercalcemia, and the rare instances of hyperthroidism, myxedema, adrenocortical insufficiency, hypophosphatasia, and idiopathic osteoporosis associated with increases in serum calcium. In general, only the hypercalcemia of hyperparathyroidism fails to decrease after pharmacologic dosages of hydrocortisone and related steroids are administered in the so-called cortisone test; in the others, the serum calcium levels return to normal.

Serum alkaline phosphatase levels are increased in primary hyperparathyroidism only when bone involvement is extensive. Serum potassium and magnesium may be low. Increases in serum calcium impair the urinary excretion of hydrogen ions, and this may result in an alkaline urine. Other body fluids remain normal unless kidney failure supervenes.

Effect of hyperparathyroidism on urinary calcium

Urinary calcium is increased in hyperparathyroidism. This is most clearly demonstrated when the excretion of calcium in urine exceeds the daily intake of calcium and is one of the more reliable tests for hyperparathyroidism. Other indices employed in the diagnosis of this disorder include the urinary excretion of phosphorus in relation to urine creatinine, the effect of injections of parathyroid hormone or of calcium or phosphorus salts, the localization of the parathyroid tumour by its uptake of radioactive selenium, the demonstration of microscopic bone destruction by osteoclast cells, and the measurement of circulating levels of parathyroid hormone.

If stupor, coma, or a hypercalcemic parathyroid crisis occurs during primary hyperparathyroidism, the serum calcium can usually be lowered by administering phosphate, sulfate, or citrate salts. Therapy of primary hyperparathyroidism consists of the removal of the parathyroid tumour or tumours or of excess parathyroid tissue. In rare instances, emergency removal of the parathyroid tumour or glands has been necessary when hypercalcemic coma unresponsive to the usual measures has developed.

The possibility of a Zollinger–Ellison syndrome and variants thereof—comprising peptic ulcer, pancreatic islet tumour, adenomas of parathyroid and other endocrine glands such as the pituitary and adrenal, and other findings—is normally explored in any person in whom a parathyroid adenoma has been found. Hyperparathyroidism caused by a parathyroid adenoma without detected evidence of other endocrine adenomas may occur in brothers and sisters and in successive generations of the same family.

Secondary hyperparathyroidism. Chronic kidney failure (uremia) is the commonest cause of secondary hyperparathyroidism; *i.e.*, of excesses of parathyroid hormone traceable to a known cause. The increased secretion of parathyroid hormone results from the increase in serum inorganic phosphorus, the decrease in serum calcium, and the acidosis that develop in uremia. Under these circumstances, the absorption of calcium from the gut is impaired, and vitamin D is relatively ineffective. Parathyroid overactivity also occurs secondary to other types of chronic acidosis, rickets (osteomalacia), phosphate loading, and calcium deprivation.

Approximately one out of five persons with kidney failure and secondary hyperparathyroidism also shows X-ray evidence of bone destruction, and other bone lesions may also be present. Serum alkaline phosphatase levels are increased, and each parathyroid is increased in size—as differentiated from a primary parathyroid tumour or adenoma, in which usually only one gland is increased in size. Diffuse enlargement of more than one, or all, of the glands, however, can also be a cause of primary hyperparathyroidism. In some individuals with uremia, the secondary hyperparathyroidism raises the low serum calcium levels to or above the normal range. Under these circumstances, calcium is deposited in the soft tissues—especially in the kidney and, thus, further aggravates the kidney failure.

Restoration of normal calcium and phosphorus levels by means of an artificial kidney or by a kidney transplant can correct the secondary hyperparathyroidism. Surgical removal of the excess parathyroid tissue at times has been necessary in persons with renal failure, bone pain, repeated fractures, severe itching of the skin, and in those who develop high serum calcium levels after a kidney transplant.

DEFICIENCIES OF PARATHYROID HORMONE

Idiopathic and acquired hypoparathyroidism. In parathyroid underfunction—*i.e.* in hypoparathyroidism—the control of body calcium and phosphorus is disturbed. In one type of hypoparathyroidism, in which the parathyroid glands are absent, inorganic phosphorus is retained in the serum, which, at the same time, becomes deficient in ionized calcium. This condition frequently follows inadvertent removal of the parathyroids during thyroid surgery, occurs transiently after the excision of a parathyroid tumour, develops only rarely after treatment with radioactive iodine, may become manifest in infants born of hypoparathyroid, hyperparathyroid, or diabetic mothers, or has no identifiable cause.

Decreases in serum calcium can produce contractions and twitching of muscles (tetany), numbness and tingling, convulsions, cataracts, physical and mental retardation, malformation of the nails, faulty dentition, thin and dry hair, fungus infection (moniliasis), diarrhea, and, rarely, congestive heart failure. The tetany may first become manifest after an acute infection or following pregnancy or during lactation. Hypoparathyroidism may occur in various combinations with Addison's disease, pernicious anemia, fatty diarrhea, and moniliasis.

The diagnosis of hypoparathyroidism depends on the

demonstration of decreases in serum calcium and increases in serum inorganic phosphorus, both in the absence of kidney failure. The serum alkaline phosphatase levels are normal. Changes may be present in the electrocardiogram and brain wave tracings, but results of X-ray examination of the skeleton are usually normal. Calcification of the soft tissues or of the basal ganglia of the brain can occur. The individual with acquired or idiopathic hypoparathyroidism responds to injections of parathyroid hormone of animal origin by an exaggerated increase in the excretion of phosphorus in the urine.

Treatment of hypoparathyroidism

Treatment with vitamin D or related sterols (*e.g.*, dihydrotachysterol) and large amounts of calcium salts, such as calcium lactate, can restore serum calcium levels to normal. The drug probenecid, which increases the urinary excretion of phosphorus, may also be used. Diets low in phosphate and the oral administration of aluminum hydroxide gels, which remove phosphate via the gut, may be prescribed to decrease the serum phosphorus levels.

Pseudohypoparathyroidism. Pseudohypoparathyroidism is characterized by all of the features of hypoparathyroidism, such as tetany, convulsions, and cataracts, plus variable somatic and skeletal abnormalities. In pseudohypoparathyroidism, however, the parathyroid glands are present and may even be enlarged, but the tissues do not respond normally to parathyroid hormone. Body build, bone, and soft tissue abnormalities may be present in pseudohypoparathyroidism; these include a short, thickset body, round face, shortening of lateral digits of the hands and feet due to short metacarpal and metatarsal bones, and calcification of the brain basal ganglia and of subcutaneous tissues. The bone changes in pseudohypoparathyroidism are of the chondrodysplastic type and are not recognized to be causally related to the pseudohypoparathyroidism. Females are affected more often than are males by the condition.

Secondary hypothyroidism—*i.e.*, a shortage of thyroid hormones secondary to a lack of pituitary TSH—may also develop.

The criteria for the diagnosis of pseudohypoparathyroidism consist of low serum calcium levels and increased serum inorganic phosphorus without evidence of kidney failure or other disease, normal serum alkaline phosphate, and absent or diminished increase in urine phosphorus after injection of parathyroid extract (the Ellsworth–Howard test). Since injection of parathyroid extract either does not induce an increase in urine phosphorus or produces a poor response by comparison to that obtained in idiopathic or acquired hypoparathyroidism, it is evident that the kidneys are absolutely or relatively unresponsive to this hormone, meaning that an end-organ defect is present.

Therapy of pseudohypoparathyroidism is the same as that employed for idiopathic or acquired hypoparathyroidism—*i.e.*, vitamin D, dihydrotachysterol, calcium salts, low phosphate regimens, aluminum hydroxide gel to increase the fecal loss of phosphate, and probenecid to raise the urinary output of phosphate—but parathyroid hormone, of course, would be of no therapeutic value.

Pseudo-pseudohypoparathyroidism. Individuals who resemble those with pseudohypoparathyroidism in body build and skeletal abnormalities but whose serum calcium and phosphorus values are normal have been called pseudo-pseudohypoparathyroid. This term is confusing, and it is simpler to designate them as chondrodysplastic. The condition differs, however, from ordinary chondrodysplasia in that, as in pseudohypoparathyroidism, soft tissues such as those of the skin, feet, and ears may become calcified, and calcium deposits may be present in the brain. At times excrescences (exostoses) develop in various bones. This type of chondrodysplasia has also been observed in individuals with gonadal dysplasia of the Turner's syndrome type (see also GENETICS, HUMAN).

IV. The adrenals

ADRENAL CORTEX

The adrenal, or suprarenal, glands present at the upper ends of the kidneys produce two completely different sets of hormones. The outer layer, or cortex, is the source of several steroid hormones. The inner portion or medulla secretes amine (NH_2-type) hormones, substances of completely different chemical and physiological properties.

Zones of the adrenal cortex

The adrenal cortex can, in turn, be subdivided into three zones, depending on cell type and arrangement and, more importantly, on the basis of the steroid hormone produced by each of these layers. Thus, the outermost layer, or zona glomerulosa, is the production site of aldosterone, a potent mineral-regulating steroid. The middle, or zona fasciculata, secretes cortisol (hydrocortisone), a steroid hormone. Cortisol is essential for life; it influences the metabolism of foodstuffs, controls sodium, potassium, and chloride exchanges, and participates in a host of other processes. The innermost layer of the cortex, or zona reticularis, is the site of adrenal androgen and estrogen production. The activities of the middle and inner zones of the adrenal cortex are under the influence of ACTH of the anterior pituitary and, in turn, of the ACTH-releasor factor or corticotropin-releasor factor, or CRF, of the hypothalamus. Needs for additional amounts of cortisol evoke increases in CRF and, in turn, of ACTH; on the other hand, the administration of amounts of cortisol in excess of daily secretion suppresses CRF and ACTH. These fluctuations in CRF and ACTH in response to cortisol needs also affects adrenal androgen and estrogen production because, as mentioned, the zona reticularis is controlled by ACTH.

The control of aldosterone secretion from the outer layer of the adrenal centre is mediated largely through extra-hypothalamic and extra-pituitary sources such as the juxtaglomerular apparatus of the kidney (see ENDOCRINE SYSTEM, HUMAN), which elaborates the hormone renin. Renin, in turn, influences the production of substances called angiotensin I and II, and the latter evokes a secretion of aldosterone from the zona glomerulosa. Only under limited circumstances does this outer layer of the adrenal cortex respond to ACTH.

Excesses of adrenocortical hormones. *Cushing's syndrome.* Cushing's syndrome develops as a consequence of a prolonged excess of cortisol or related steroids. It may result from a basophilic tumour of the pituitary gland as originally suggested by Harvey Cushing, the surgeon who first described the condition, and is then called Cushing's disease or pituitary basophilism. Such pituitary tumours produce great excesses of adrenocorticotropin (ACTH), which increase the secretion of cortisol by the outer layer of adrenal cortex and result in hyperplasia (enlargement) of this portion of the adrenal gland. More often, however, spontaneous Cushing's syndrome results from enlargement of the outer layers of the adrenal glands without evidence that a pituitary tumour is present. Less often, a benign tumour (adenoma) or a cancer of the adrenal cortex causes a spontaneous Cushing's syndrome. In rare instances, other cancers—such as those of the lung, colon, or thyroid—produce ACTH, and Cushing's syndrome results. Cushing's syndrome can also be produced by the administration of sufficient amounts of ACTH of animal or laboratory origin or of synthetic cortisol-like steroids.

Bodily effects of Cushing's syndrome

The excesses of cortisol and related steroids that develop in the spontaneous or induced types of Cushing's disease or syndrome affect the following systems and tissues: *Central nervous system:* Serious emotional illnesses (psychoses) in 25 percent and less obvious emotional and personality changes in another 35 percent of the spontaneous instances of Cushing's syndrome; 15 percent, or higher, incidence of personality disturbances in the therapy-induced type. *Body build:* Obesity of the trunk, thin extremities, localized collections of fat (round and moonlike face, "buffalo hump" and supraclavicular —across the shoulder—pads) and increased prominence of the eyes (exophthalmos). *Skin and hair:* Redness of the face (caused by dilated capillaries), purple stretch marks at the hips and shoulders; and baldness or hirsutism in the female. *Lymphoid tissues and blood:* Decreased lymphoid tissue, low blood eosinophil and lymphocyte counts and increased numbers of blood polymorphonuclear leucocytes, possible changes in basophils, platelets,

and sedimentation rate, and shortened coagulation time (see also BLOOD, HUMAN). *Heart, blood vessels, and kidneys:* High blood pressure, increased heart size, possible congestive heart failure; kidney failure, usually, with longstanding disease. *Body fluids:* Sodium and chloride retention, potassium losses, and metabolic alkalosis; lowered concentrations of chloride and potassium in serum (effects are less marked or absent with certain of the newer synthetic cortisol derivatives); increased uric acid, calcium, phosphorus, and possibly magnesium, in the urine. *Gastrointestinal tract:* Abdominal discomfort; increased incidence of ulcers. *Liver:* Increased liver glycogen storage and variable increases in bile flow. *Pancreas:* Decreased volume and change in composition of pancreatic secretions, destruction (necrosis) of pancreatic fat. *Metabolism:* Losses of body nitrogen, blockade of insulin action, increased sugar formation from protein (gluconeogenesis), delayed glucose disposal and diminished response to insulin, and frank diabetes mellitus in 25 percent of spontaneous Cushing's syndromes. *Muscles and skeleton:* Muscle weakness, thinning of the bones (osteoporosis), and compression fractures in advanced stages of the disease.

Treatment and remission of Cushing's syndrome

Many of the manifestations of Cushing's syndrome disappear when the cause of the cortisol excesses is removed. If the cause is from ACTH or steroid treatment, remission occurs when treatment is discontinued. In the spontaneous types, partial or complete surgical removal of the overactive adrenal cortices, or of the ACTH-producing tumour of the pituitary, or cancer located in other tissues or organs, is followed by a remission of the symptoms and signs. Daily treatment with cortisol is essential if the adrenal cortices have been completely removed, but it may also be necessary after partial removal of these glands. Cushing's syndrome is also treated at times by irradiation of the pituitary with a proton beam, by implants of radioactive materials, or by administration of certain chemical agents that block cortisol production.

Normal adrenocortical function is not always the end result of successful therapy of Cushing's syndrome. Removal of a hyperfunctioning adenoma or cancer may leave the patient hypoadrenocortical—*i.e.*, short of adrenal steroids—because of persistent suppression of the remaining adrenal. Also, surgical removal of the entire pituitary or other induced forms of hypopituitarism are followed by deficient gonadal, thyroidal, and adrenocortical function. Most of these can be corrected by appropriate replacement therapy.

Even when the Cushing's syndrome and the untoward effects or side effects of treatment have been eradicated, however, some of the changes produced by the Cushing's syndrome may continue. This is true, for example, of the heart, blood vessel, kidney changes, and the osteoporosis. Finally, pituitary enlargement may follow adrenalectomy as a result of the growth of a previously present or newly developed pituitary tumour.

Primary aldosteronism. Aldosteronism refers to the presence of excess aldosterone, an adrenal steroid with potent mineral regulating and other effects, secreted by the adrenal glands. During good health, the time of day, the position of the body (*i.e.*, upright or horizontal), and the body's supplies of sodium and potassium affect the rate of secretion of aldosterone. Aldosterone secretion is higher during the day and in the upright position. Sodium loading or potassium restriction decreases the aldosterone secretion, whereas sodium depletion or potassium excess raises it. Under ordinary circumstances, the adrenal glands of healthy individuals secrete less than 300 micrograms (millionths of a gram) of aldosterone daily. About one-seventh of this amount appears in the urine, and plasma levels of this hormone are one microgram per litre or less. When these limits are consistently exceeded, either primary or secondary aldosteronism—*i.e.*, hyperaldosteronism—is considered to be present. Primary aldosteronism results from a benign tumour (adenoma) or a cancer of the adrenals and rarely from an unexplained increase in the size of the aldosterone-secreting layers of these glands. The secondary type represents excess aldosterone production in response to various disturbances in the body fluids and in the concentration and amount of sodium and of potassium in particular.

Bodily effects of primary aldosteronism

Primary aldosteronism produces the following manifestations: Weakness, disturbed sensations, and possible tetany; usually persistent high blood pressure (edema is rare); increased losses of potassium in urine; low levels of potassium and high levels of sodium in serum; increased fecal losses of potassium, calcium, and magnesium; inability to concentrate urine with resultant increased urine volumes; vacuole formation in kidney tubules of the type seen in potassium depletion; destruction of fibres; periodic paralysis of the low-potassium type; increased frequency of diabetes mellitus; decreased response to antidiuretic hormone with increased urine volume and increased thirst.

Primary aldosteronism is treated by surgical removal of the excess adrenocortical tissue. The effects of aldosterone excesses on blood pressure and body electrolytes can be cancelled by the drug spironolactone, which interferes with the action of aldosterone.

Secondary aldosteronism. Secondary aldosteronism is most commonly seen in association with congestive heart failure, cirrhosis-like kidney disorders, including the nephrotic syndrome, either as a result of the underlying disease or as a consequence of treatment with diet or medications that produce a shortage of sodium or an excess of potassium. Edema—*i.e.*, swelling of body tissues—is common. Edema that develops during the day in some women while walking about may also be a form of secondary aldosteronism and can be overcome by lying down.

Successful treatment of the heart, or of liver or kidney disorder, and cancellation of the electrolyte abnormalities that produced the secondary aldosteronism returns the supplies of the hormone to normal. Aldosterone antagonists, such as the spironolactones that block the action of the hormone in the kidney tubules, may eliminate the excess sodium and water and other abnormalities that characterize secondary aldosteronism.

Deficiencies of adrenocortical hormones. *Addison's disease.* Historically, the phrase Addison's disease has generally referred to those instances of spontaneous and gradual loss of adrenal steroid secretions that result from destruction of the adrenal cortices by such conditions as atrophy, tuberculosis, or other invasion by micro-organisms, and cancer. An equivalent degree of primary hypoadrenocorticism—*i.e.*, shortage of adrenocortical steroids—however, can result from surgical removal of the adrenals, suppression of adrenal function by prolonged steroid therapy, massive bleeding into the adrenals, or by toxic agents. Adrenocortical insufficiency involving decreased secretion of cortisol as a result of enzymatic abnormalities can also be a feature of the adrenogenital syndrome.

The microscopic findings in the causes (etiology) of Addison's disease appear to be changing: tuberculosis of the adrenals is decreasing, whereas the incidence of idiopathic atrophy is rising. The origins of the latter are not entirely clear. It is conceivable that in some instances specific agents or events, such as chemicals or hemorrhage, respectively, produced the destruction, but most cases of the atrophic type probably result from destruction of the adrenals by circulating antibodies. Addison's disease predominates in adults, becoming manifest most often in early or middle life, but it can also develop in children. Also, the simultaneous occurrence of a fungus infection—moniliasis—and of deficient parathyroid gland secretion in children with Addison's disease has been observed.

None of the manifestations of Addison's disease can be considered diagnostic as they also occur under other circumstances. Combinations of the following, however, point to the diagnosis: fatigue and weight loss; occurrence of increased pigmentation of skin and mucous membranes, especially at pressure points and in scars, dark brown freckles, and areas of depigmentation (vitiligo); variable decreases in body and sex hair; rarely, personality disturbances, or even psychosis; abnormal brain wave tracings; low blood pressure, decreases in

blood pressure on standing; small-sized heart, abnormal heart tracings; circulatory collapse during adrenocortical or Addisonian crises; loss of appetite; nausea and vomiting; increase in blood lymphocytes and eosinophils; depletion of sodium, chloride, and water; retention of potassium; metabolic acidosis; delays in excretion of water; reversibly impaired kidney function and retention of nitrogen as a result of decreased circulatory efficiency; abnormally low values of cortisol in plasma and of cortisol derivatives, 17-ketosteroids, aldosterone, and other adrenal steroids in urine—not corrected by injections of ACTH; decreased liver glycogen stores; unduly low blood-sugar levels in the fasting state or following sugar or insulin administration; inflammation of the thyroid (thyroiditis); associated hypoparathyroidism or diabetes mellitus; decreased steroid requirements during pregnancy.

Partial versus total Addison's disease

Addison's disease may be partial; *i.e.*, steroid secretions are adequate under nonstress situations but cannot be adequately increased during infection, surgery, and other stress because adrenocortical reserves have been lost. Fractional Addisonian-like deficiencies can also occur: loss of cortisol with adequate production of aldosterone and androgens, deficits of aldosterone with other steroids normal, and similar situations.

In total Addison's disease, there is a disturbance or complete loss of (1) maintenance of normal volumes and levels of body water and salts by means of adrenal mineral-regulating steroids, (2) control of metabolism by cortisol, and (3) contributions of adrenal androgens and estrogens to muscle strength, body build, and secondary sex characteristics. Thus, deficits in aldosterone, desoxycorticosterone (DOC), and perhaps of cortisol and related steroids result in inadequate kidney reabsorption and subsequent loss of sodium and chloride ions and water. The end result is potassium retention and circulatory collapse. Collapse or crisis as a consequence of cortisol deficits may, however, occur suddenly before any changes in body fluids become evident, because cortisol is essential for life and normal cell function. Moreover, shortages of cortisol affect the conversion of foodstuffs to glucose and the storage of glycogen. Thus, the formation of glucose from amino acids and glycerol is impaired, and the deposition of glycogen is reduced. At the same time, the blockade of insulin action that cortisol and related steroids exert on the tissues is lost. The blood sugar usually decreases to a moderate degree, but this may or may not produce symptoms. Finally, the loss of adrenal androgens deprives the female of her only significant source of these steroids, decreases sex hair and libido, and aggravates the loss of muscle strength. These changes are less significant in the male because of the continued contribution of androgens by the testes. The deficits of adrenal estrogens do not appear to be clinically significant in either sex.

The increased pigmentation of the skin and mucous membranes of Addisonian patients is attributable to losses of cortisol, desoxycorticosterone, and other steroids with subsequent stimulus (via feedback) to higher regulatory centres, including the pituitary and hypothalamus. This evokes an increased output of adrenocorticotropic hormone (ACTH) and melanocyte-stimulating hormone (MSH), each of which increases melanin pigment production. Aldosterone deficits do not play a role in the hyperpigmentation because, as previously mentioned, the secretion of this steroid is not primarily under the control of ACTH.

The above interrelations illustrate, however, the relative or possible roles of the various classes of adrenal steroids in homeostasis—*i.e.*, the overlap of aldosterone, DOC, and cortisol in the regulation of electrolytes and water; the major contribution of cortisol to intermediary metabolism; and the less significant role of aldosterone and the unimportance of DOC in this respect. The complexity of regulatory mechanisms is further evident in the control of water balance. Thus, in adrenocortical insufficiency, DOC can raise the kidney glomerular filtration rate (GFR) but cannot restore to normal the excretion of a water load; aldosterone also fails in the latter respect, but cortisol and related steroids do accomplish this. Similarly, overlapping of aldosterone, DOC, and cortisol is manifest in the regulation of sodium excretion.

Use of adrenal steroid secretion in diagnosis

Definite evidence of primary—*i.e.*, nonpituitary and nonhypothalamic—adrenocortical insufficiency must be based on the demonstration of the absence of, or restricted, ability of the adrenals to elaborate the chief classes of adrenal steroids under the impetus of adequate stimulation. In the case of cortisol and the androgenic and other precursors of the 17-ketosteroids of adrenal origin, ACTH administration provides a simple and adequate stimulus. (ACTH can also be used to test the adrenal output of estrogen, but this is not a useful test for Addison's disease). In usual practice, ACTH of animal or synthetic origin is injected, and the diagnosis of Addison's disease is established if plasma cortisol-like steroids or the various adrenal-steroid products present in urine do not increase. Aldosterone secreting capacity is tested by measuring the levels of this hormone in response to rigid restriction of sodium, potassium loading, or both, since, as already indicated, aldosterone secretion is not under significant ACTH control.

Other indices employed for the diagnosis of Addison's disease are less conclusive. Thus, sodium and chloride levels may be low in serum and high in sweat and in urine, while potassium is high in serum and low in sweat and urine. All three electrolytes, however, may be normal in these various body fluids. The excretion of water loads is generally delayed in adrenocortical failure, but false positive and, less often, false negative tests occur. Failure to demonstrate sodium conservation and the production of potassium accumulation by diets deliberately low in sodium or high in potassium is an extremely dangerous index of loss of mineral-regulating steroids and, therefore, is no longer used in detecting Addison's disease. Although low blood-sugar levels are usual in Addison's disease, they are not dangerously low, nor are they limited to this disorder; the same is true of the responses to glucose or insulin. Inconsistencies in circulating eosinophil and urinary uric acid–creatinine responses to ACTH render these indices virtually useless as criteria for diagnosis in individual patients.

In the past, sodium supplementation, potassium restriction, and between-meal feedings prolonged survival, but for only a few months, and it did not correct the weakness and fatigue of Addison's disease. Some persons lived for a decade or more, but these may well have had some residual ability to produce cortisol. The advent of desoxycorticosterone (DOC) solved the problem of sodium depletion and potassium retention but did not meet the need for cortisol. Stresses such as infections and surgical procedures still produced crises that often proved fatal. Adrenocortical extracts were employed in the treatment of such crises, but the quantities of steroids, and of cortisol in particular, thus administered were often too small. Expense prohibited use of these extracts for maintenance purposes.

Synthetic adrenal steroids for treatment and maintenance

The current availability of synthetic cortisol-like and related adrenal steroids has greatly simplified the maintenance and treatment of adrenocortical insufficiency. The usual programs are based upon cortisol in daily maintenance dosages, often supplemented by sodium- and potassium-regulating steroids, such as desoxycorticosterone or closely related 9-α-fluorohydrocortisone. In times of stress or in crisis, the dosages of cortisol are increased to quantities sufficient to avert or correct Addisonian crises.

Waterhouse–Friderichsen syndrome. The Waterhouse–Friderichsen syndrome consists of a fulminating infection, usually meningococcal, in the bloodstream (septicemia); hemorrhage; or clotting of the blood in the adrenal glands, in the skin, and in blood vessels in general; adrenocortical insufficiency; and circulatory collapse. Actual hemorrhage into the adrenal may not be present even in fatal cases, although adrenocortical destruction is usually evident. The syndrome is especially likely to occur in children.

The two major problems in this illness are adrenocortical insufficiency and the extensive hemorrhage and

clotting of blood in tissues. Treatment consists of massive dosages of cortisol (there is not an expert consensus on this point), sulfa drugs, fluids, and support of the blood pressure as needed.

Adrenalectomy. Removal of the adrenal or adrenals (adrenalectomy), partial or complete, is highly effective in the treatment of tumours of the adrenal cortex or medulla, in Cushing's syndrome, and in primary aldosteronism. It has also been successfully employed in attempts to prevent the growth of spreading cancer of the breast. In the past it was used, with questionable benefit, to treat prostate cancer, and with some effect in the control of hypertension unresponsive to drugs and none in the slowing or stopping of vision-threatening changes in the eyes of diabetic individuals.

Preoperative and postoperative care is based on the administration of adequate quantities of cortisol: 500 milligrams or more, or physiologically equivalent amounts of related steroids are given during the first day. These quantities are decreased on successive days. Ultimate maintenance dosages, when needed, are the same as those employed in other types of adrenocortical insufficiency.

Secondary hypoadrenocorticism. Decreases in or losses of pituitary adrenocorticotropic hormone (ACTH) as a consequence of nonfunction of the pituitary or the hypothalamus result in secondary hypoadrenocorticism and a shortage of adrenal steroids other than aldosterone.

Causes of secondary hypoadrenocorticism

This may result from destruction of the pituitary by a tumour, for example, or hemorrhage, surgery, or from damage to the ACTH-releasor areas of the hypothalamus. Such adrenocortical underfunction may also be a feature of myxedema with resultant depression of the pituitary. The most common form, however, is that induced by prolonged therapy with cortisol and related steroids. In theory, secondary hypoadrenocorticism is easily distinguished from the primary type by the absence of increases in urine and plasma steroids in the latter following ACTH injection. Responses to ACTH, however, may also be decreased or absent in long-standing secondary hypoadrenocorticism.

Patients with secondary hypoadrenocorticism do not show the pigmentary or sodium, chloride, and potassium symptoms of pituitary adrenocortical insufficiency of the Addison's disease or related type, but all the other manifestations may be present. In addition, patients with damage of the pituitary or hypothalamus will often have FSH, LH, TSH, or growth hormone deficits and, at times, diabetes insipidus, obesity, and other dysfunctions. ACTH in plasma or serum is difficult to measure, and such data are not readily available. Indirect evidence of ACTH deficiency may be obtained with the drug metyrapone, which blocks cortisol synthesis. In the normal person it evokes a rise both in ACTH and in certain easily measured plasma and urine steroids.

There is at present no satisfactory method of avoiding the adrenal suppression produced by large dosages of cortisol-type steroids taken for many months or years. The best preventive is the limitation of steroid therapy to intermittent courses of several days or administration of a single large dose of steroids every 48 hours. Also, intermittent injections of ACTH may serve to keep the adrenal cortices responsive until internal secretion of ACTH is resumed.

Therapy of secondary hypoadrenocorticism of the irreversible type consists of replacement dosages of cortisol with provision for increases during stress. Aldosterone or desoxycorticosterone (DOC) are not needed, either because endogenous production remains adequate despite losses of ACTH or because mineral regulation is effected by the cortisol. If gonadotropins and TSH have also been lost, androgens or estrogens and thyroid are given. Care is taken to meet the cortisol needs before administering thyroid hormones in myxedema and thereby avoid the possibility of an adrenocortical crisis.

Combined deficits and excesses of adrenocortical hormones. *Adrenogenital syndromes.* The phrase adrenogenital syndromes refers to specific disorders of the adrenal and either the ovaries or testes, often accompanied by abnormalities of the genital apparatus. The adrenogenital syndromes of infancy and childhood are characterized by compensated, impending, or actual deficits in the enzymatic synthesis of cortisol by the adrenals that are genetically determined, and by an associated excess of androgen production by the adrenal glands. The actual or impending shortage of cortisol evokes an increased secretion of adrenocorticotropic hormone (ACTH) by the pituitary, which increases cortisol production toward or to normal but simultaneously raises above normal the production of adrenal androgens.

Symptoms of the adrenogenital syndrome

The recognized symptoms and findings in the several types of adrenogenital syndromes stem from varying combinations of ACTH-induced androgenic excesses, deficits of cortisol, and the production of other steroids in disproportionate amounts.

Excesses of androgens in certain types of adrenogenital syndrome present before birth produce male-like genital changes (pseudohermaphroditism) in genetic females and excessive genital development with small testes in genetic males. Masculinization of body build then occurs in both sexes, and sex hair appears prematurely, but breast development and menses are absent. Body growth is accelerated, but the ultimate height is shortened. Acne may appear in both sexes, and facial hair increases in the females. In persons with the 3-β-hydroxyl dehydrogenase deficiency or those with defects in conversion of cholesterol to pregnenolone, the developmental and genital changes are variable. In the former, the newborn infant may die, and, in the latter, undescended testes have been noted. The 17-α-hydroxylase type of adrenogenital syndrome has been characterized by absence of menses, high blood pressure, and swelling of the extremities and other tissues.

The deficits of cortisol that occur in the adrenogenital syndrome produce pigmentation of skin and mucous membranes as a result of the excesses of ACTH. Life-threatening Addisonian or adrenocortical crises may occur with or without increases in serum potassium, decreases in serum sodium, acidosis, and kidney failure. Blood-sugar levels tend to be low; the disposal of sugar loads is accelerated; and the blood-sugar responses to insulin are increased. In addition to the cortisol shortages, actual or impending, present in all types of adrenogenital syndromes, the levels of other steroids are also altered. Another enzymatic defect of steroid production by the adrenals results in a shortage of aldosterone. This has not been associated with changes in the genitalia nor with a shortage of cortisol.

Treatment of adrenogenital syndromes

The adrenogenital syndromes are treated by replacement–displacement amounts, usually 20 to 30 milligrams per day on a maintenance basis, of cortisol. This results in the return of normal steroid patterns; *i.e.*, the shortage of cortisol is cancelled, and the excessive production of androgens and of steroids, which results in masculinization, high blood pressure, and related manifestations, ceases. The adrenocortical crises are treated as in Addison's disease. Therapy of amenorrhea requires estrogen administration. Plastic surgery on the genitals and reorientation of the individual's social sex are then undertaken after careful review of the psychological and emotional factors involved.

Adrenogenital syndromes can be mimicked by excesses of various steroids with actual or potential androgenic activity administered during pregnancy and resulting in abnormal development of the genitals prior to birth. Also, the administration to growing children of medications with male sex hormone produces changes in the body build, sex hair, and genitals of the type that develop spontaneously in the adrenogenital syndrome. In these, however, there are no subsequent developmental abnormalities of the type seen in the enzymatic forms of adrenogenital syndrome, such as accelerated body growth, retarded or accelerated sexual maturation, and others previously described. Also, the tendency to cortisol deficits and the abnormal steroid patterns encountered in the spontaneous adrenogenital syndrome are not present, and the urinary 17-ketosteroids are not elevated.

Adrenogenital syndromes may also be mimicked by

tumours of the adrenals, the testes, the ovary, or other tissues that can produce excesses of male sex hormone and thus modify the primary and secondary sex characteristics of males and females, while cortisol production and supplies are normal.

Finally, adrenogenital type changes of unexplained origin may be present at birth. Some of these are associated with gross abnormalities of the gastrointestinal and urinary tracts.

Adrenogenital syndrome in adults. Adrenogenital syndromes based on deficiency of certain hydroxylase enzymes may occur in adults. In each of these syndromes, the shortage of cortisol can result in an Addisonian crisis. Hypertension can be a feature of the 11-β-hydroxylase and the 17-α-hydroxylase types. Increases in body hair other than the scalp and masculinizing changes develop in the 21-β-hydroxylase and 11-β-hydroxylase disorders but not with 17-α-hydroxylase deficiency. Loss of menses occurs in all three.

Changes in blood serum content

The blood and serum solute changes, when present in the 11-β-hydroxylase and 21-β-hydroxylase types, are the same as those that occur in the childhood form. They may include low blood sugar and increased serum urea nitrogen, depletion of sodium, and increases in serum potassium. In the 17-α-hydroxylase type, an excess of desoxycorticosterone results in retention of sodium, losses of potassium and chloride, and increases in serum carbon dioxide. The plasma and urinary steroid patterns are identical with those observed in children. In all three forms, treatment consists of cortisol in replacement–displacement dosages, and estrogen is also given in the 17-α-hydroxylase type.

Increases in facial and body hair (hirsutism) may be present in females who do not have one of the recognized forms of adrenal or ovarian disorders—such as the adrenogenital syndrome, Cushing's syndrome, adrenal or ovarian tumour, or polycystic ovarian change—that increase androgen supplies and support hair growth. Some represent unrecognized disorders. In others, hirsutism may be a genetic trait or may result from undue sensitivity of hair follicles to ordinary supplies of androgen. In such hirsutism, a trial of cortisol-type steroids in small dosages may be undertaken. Estrogens or ovulation suppressants may also be effective.

ADRENAL MEDULLA

The adrenal medulla produces three simple amines or catecholamines: epinephrine or adrenaline, norepinephrine or noradrenaline, and isoproterenol, which, in addition to their endocrine functions, also serve as chemical mediators of autonomic nerve impulses in various organs.

Responses to epinephrine and norepinephrine

The responses to epinephrine and norepinephrine differ. The former produces a net dilatation of peripheral blood arterioles, decreases the diastolic blood pressure, and increases the pulse pressure (*i.e.*, the difference between the systolic and diastolic blood pressure readings), the pumping action of the heart, and the pulse rate. Norepinephrine induces constriction of peripheral vessels, increases the diastolic blood pressure, narrows the pulse pressure, decreases the output of blood by the heart, and slows the pulse rate.

Both epinephrine and norepinephrine raise the blood glucose by mobilizing glucose from liver and muscle glycogen and restricting the output of insulin by the pancreas. These two amines also mobilize fatty acids from body fat depots.

Isoproterenol is the most active of the amines that act almost exclusively on beta receptors: the heart, the smooth muscles of the bronchi, the blood vessels of skeletal muscle, and the alimentary canal. The hormone lowers peripheral vascular resistance, raises cardiac output, relaxes smooth muscle, and causes the release of free fatty acids. It causes less hyperglycemia (high blood sugar) than does epinephrine but has almost the same calorigenic value.

Pheochromocytomas are catecholamine-secreting tumours that develop in the adrenal medulla or at nerve endings. Those within the adrenal secrete epinephrine and norepinephrine. Those that arise from nerve endings produce only norepinephrine. They occur in the population at a rate of 1 per 1,000 and are more likely to involve the right adrenal. They may be large or small, are bilateral about 10 percent of the time, and one out of ten are cancers. The tumour may occur in locations outside the abdominal cavity; *e.g.*, the chest, the pelvis, or the urinary bladder. The possibility of a pheochromocytoma is especially likely in relatives of proved cases and in persons with conditions such as neurofibromatosis.

Transient or persistent elevation of the blood pressure is almost an invariable hallmark of the tumour, but the following may also be encountered: *General:* thin body build, pallor, and sweating; *central nervous system:* headache, nervousness, anxiety, personality change, convulsions, and eye ground (fundus) signs indicative of high blood pressure; *other heart and blood changes:* heart pain, shortage of breath, enlargement of the heart, congestive heart failure, cyanosis, and blood pressure decreases on assuming upright position (postural hypotension); *effects on urine and kidneys:* protein and sugar in the urine; kidney failure; *gastrointestinal:* abdominal pain, vomiting; *endocrine and metabolic:* increased water exchanges, diabetes mellitus, loss of subcutaneous fat, and elevated basal metabolic rate (BMR).

The availability of analytical technics or assays for the measurement of epinephrine and norepinephrine and their derivatives in urine and other body fluids has greatly facilitated the identification of such tumours. A diagnosis may also be made by various X-ray techniques, by provoking a rise in blood pressure by massage of the tumour area, or by the intravenous injection of test substances such as histamine when the blood pressure is not elevated. If the blood pressure is high, such drugs as phentolamine or piperoxan, which cancel the blood pressure rises produced by epinephrine or norepinephrine excesses, are injected and are used as screening tests for pheochromocytoma. Anesthesia, sedatives, drugs that lower the blood pressure, and kidney failure increase the incidence of false results with these screening tests.

Successful surgical removal of the tumour usually brings about a disappearance of all the reversible manifestations of a pheochromocytoma, but in some persons high blood pressure and diabetes mellitus persist. Cortisol steroids must be given during and following surgical procedures in which both adrenals are removed.

V. The gonads

PREMATURE DEVELOPMENT

True sexual precocity—*i.e.*, early maturation of ovaries and testes with production of ova before the age of nine years in girls, or sperm before ten years in boys—occurs without evident cause (constitutional or familial type) or as a manifestation of cerebral disease (cerebral type). In such true sexual precocity, the maturation of genital structures and body growth are also accelerated, and early reproduction is possible—*i.e.*, as early as the age of six years. After a period of accelerated growth, premature fusion of the growing ends of the bones, the epiphyses, occurs with ultimate stunting. The manifestations in the cerebral type are the same as those in constitutional or familial true sexual precocity. In addition, intracranial disease (brain tumour, for example, or fibrous dysplasia, abscess, encephalitis, or hydrocephalus) is present. Laboratory indices of ovarian or testicular function such as gonadotropin levels (the follicle-stimulating and the luteinizing hormones) and the excretion of male or female sex hormones reaches but does not exceed the range of values observed in normal young adults.

Causes of sexual pseudoprecocity

Sexual pseudoprecocity results from excesses of sex hormones (of androgens in boys and of estrogens in girls) from the adrenal cortex, testis, ovary, or from other sources, including extragonadal tumours. Such pseudoprecocity may also result from gonadotropin-secreting non-endocrine tumours of the liver or other tissues that stimulate androgen or estrogen production by the testes or ovaries. It can also be produced by the administration of androgens or estrogens.

Rare instances of sexual precocity or pseudoprecocity

have been observed in conjunction with insulin excesses, with dwarfism, and with hypothyroidism.

Sexual pseudoprecocity in boys as a consequence of excesses of male sex hormones occurs in the adrenogenital syndrome or with androgen-producing tumours of the adrenal or testis. Boys with a Cushing's syndrome caused by an adrenal tumour may also show this change. In all of these, enlargement of the penis, accelerated appearance of sexual characteristics—pubic and axillary hair, masculinization, faster body growth, and ultimate stunting—are present. The testes, however, are smaller than usual, a cardinal clinical sign, and do not produce spermatozoa. The urinary androgens and derivatives, such as the 17-ketosteroids, are elevated and may reach levels above the adult range, but follicle-stimulating and luteinizing hormones are low or absent.

Sexual pseudoprecocity in girls arises from increased supplies of female sex hormones secreted by tumours of the gonads or adrenals. The external manifestations of sexual maturation—breast formation and appearance of sex hair, for example—appear early, but the maturation and discharge of ova do not occur. Body growth and skeletal maturation are accelerated, but the final height attained is below normal. Estrogen levels are increased, while measurements of gonadotropin are below normal.

In some female infants or children, premature breast development (thelarche) may be the sole clinical manifestation of pseudoprecocity, with subsequent sexual maturation occurring at the appropriate time. Similarly, pubic and axillary hair can appear prematurely followed by normal sexual development at the usual time. Such precocious hair has been attributed to premature increases in adrenocortical secretions in the course of adrenarche (menstruation induced by adrenal activity).

Removal of the cause, if identifiable, is the first goal of therapy. In idiopathic complete sexual precocity, menstruation and other manifestations of premature sexual development may be held in abeyance or even made to regress by medroxyprogesterone. Such therapy does not slow the accelerated maturation of the skeleton and therefore will not avert ultimate stunting. Thyroid hormone administration in those with true sexual precocity in association with hypothyroidism will cancel or partially reverse the accelerated sexual development.

DISORDERS OF THE TESTES

Cryptorchidism. Terms such as undescended testes, cryptorchism, or cryptorchidism refer to testes that are incapable of descending spontaneously to the usual position in the scrotum. Such testes are located within the inguinal canal or in the abdominal cavity. The frequency of undescended testes at birth is as high as 10 percent, decreases to 1 percent at one year of age, and is below 1 percent among adults. The condition is bilateral in approximately one out of five persons with this problem. Rarely, it occurs in more than one member of a family. Undescended testes are to be differentiated from those that are temporarily retracted into the inguinal canal. At times what appears to be cryptorchidism really represents complete absence of the testes.

Frequency of cryptorchidism

Sperm production is decreased or absent in untreated cryptorchidism, but male hormone production is normal. A cryptic testis is more likely to be injured, and the blood supply is more likely to be cut off by twisting; also, cancer develops about 20 times more often in such an organ.

Gonadotropin, androgen, or surgical therapy of undescended testes is undertaken in the hope that the deleterious effects of the higher temperature of extrascrotal locations, injury, torsion, and cancer can be avoided. Human chorionic gonadotropin therapy evokes maturation of the external genitals, and in 20 to 70 percent of the cases of gonads located in the inguinal canal, the testes move down into the scrotum. Even when the testis reaches the scrotal position, its ability to produce spermatozoa tends to be low. Fertility may be higher, however, if such gonadotropin therapy is undertaken when the patient is five or six years of age. Male sex hormone therapy also produces a more pendulous adult scrotum, increases the size of the penis, and may induce the descent of testes. The testes, however, become or remain small, and fertility is low. Androgens are not used before the usual time of pubescence, or sexual maturation, because they accelerate skeletal growth and produce premature fusion of the epiphyses and stunting.

Male hypogonadism. Normal development and function of the male reproductive organs and sex-related or sex-influenced structures, such as the penis, prostate, larynx, and pubic, axillary, and facial hair, depends on release of pituitary gonadotropins at the appropriate time and in adequate amounts—of LH, which stimulates the Leydig, or male-sex-hormone-producing, cells of the testes, and of FSH, which stimulates the Sertoli, or sperm-producing, cells of these sex glands. Such release of FSH and LH marks the beginning of sexual maturation, or pubescence, and usually occurs between nine and 12 years of age. Earlier—*i.e.*, during prepubescence—the secretion of pituitary LH and FSH does not occur because the hypothalamus has not yet begun the production of the LH- and FSH-releasor hormones.

Types of hypogonadism

Defects in, or injury to, the hypothalamus, the pituitary, or the testes results in hypogonadism. Such defects can result from an absence of normal prenatal and postnatal development of these structures, including those instances associated with abnormalities of the sex chromosomes or from damage to the hypothalamus, pituitary, or testes prior to or after maturation.

Male hypogonadism can consist of deficient androgen production (Leydig cell hypofunction), deficient sperm formation (decreased Sertoli cell function), or both. Such hypogonadism can result from a shortage of pituitary gonadotropins (LH, FSH, or both) or can represent primary testicular failure. The persistent absence of either or both gonadotropins, with or without other evidences of pituitary insufficiency, identifies the hypothalamic—hypopituitary hypogonadotropic type of hypogonadism. In primary testicular failure, the target tissues of the gonadotropins in the testes—either the Leydig cells or the Sertoli cells, or both—are incapable of responding to LH, to FSH, or to both of these pituitary hormones. Ultimately, one or both gonadotropins rise above normal, depending on which target cells in the testes are defective. These are examples of hypergonadotropic hypogonadism.

The time of onset of the androgen deficiency and the presence or absence of associated pituitary hormone deficits determine the effects on body build. When a shortage of all the pituitary hormones, including FSH, LH, and growth hormone, occurs prior to or during puberty, the development of genitalia and secondary sex characteristics is retarded (hypogonadotropic hypogonadism), and dwarfism is present. If the same disorder occurs after maturity, previously developed primary and secondary sex features regress, but stature is normal.

If LH is lacking during the growth years with or without a shortage of FSH, the development of the genitalia and secondary sex characteristics is retarded and incomplete. When LH decreases or is lost after maturity, adult male sex features partially regress. If FSH supplies are normal, spermatogenesis can occur despite Leydig cell deficiency.

If only FSH is lacking, Leydig cell and somatic development are normal, but spermatogenesis does not occur, and the testes are small.

Primary testicular failure with lack of Leydig cell function prior to maturity also results in immaturity of genitalia and deficient secondary sex manifestations. If Leydig cell function becomes inadequate in adult life, sexual activity, the size of the testes and penis, and secondary sex characteristics decrease, but sperm formation may continue. If only spermatogenesis is absent because of primary testicular failure, the testes are small, but sexual function, the size of the penis, and secondary sex characteristics are not affected.

CHROMOSOMAL DISORDERS

Klinefelter's syndrome. Klinefelter's syndrome is one type of primary testicular failure usually associated with

visible abnormalities of the sex chromosomes. It is variably characterized by enlargement of the male breasts (gynecomastia), small testes with scarred sperm-producing tubules and reduced sperm formation, tall eunuchoid body build, and a tendency to intellectual retardation. The syndrome can occur without obvious gonadal and body abnormalities, and some individuals with Klinefelter's syndrome may be fertile.

Symptoms and signs of Klinefelter's syndrome

The majority of Klinefelter individuals have one extra female sex chromosome, resulting in an XXY pattern and a total of 47 chromosomes instead of 46. Other less frequent chromosomal patterns include XXXY, XXXXY, XXYY, XXXYY, mixtures (mosaics), and partial losses or deletions. About 40 percent of the diagnosed examples of Klinefelter's syndrome have a normal XY pattern. One variant of Klinefelter's syndrome, the XXXXY type, is characterized by fusion of the forearm bones and other skeletal anomalies, underdevelopment of the penis and scrotum, incomplete descent of the testes, and marked mental retardation.

Diabetes mellitus and goitre incidence are more common, and various cancers may be more prevalent, among Klinefelter's syndrome patients. Thyroidal trapping of radioactive iodine and the responses of the thyroid to injections of TSH may be low. Testosterone production tends to be below normal, with urinary 17-ketosteroids in the normal or low-normal range. Serum and urinary gonadotropins (FSH and LH) are usually, but not necessarily, increased.

Turner's syndrome. Gonadal dysgenesis (Turner's syndrome) is a disorder of development characterized in the extreme form by small "streak" gonads without secretory or germinal cells. Internal and external genitalia are female but undeveloped; body build is suppressed female in type, with absence of breasts and sex hair. Other abnormalities, such as stunted growth, web formation of the neck, shieldtype chest, coarctation of the aorta, increased carrying angle, short lateral metacarpals, kidney abnormalities, and other signs may be present. Chromosomes of body cells are normal, but sex chromosomes are abnormal in the majority of such individuals with XO, XX, and X ring patterns and mosaics; so-called super-female patterns with three or more X chromosomes may be present. In the remainder, the chromosomal pattern is normal female, and in these the sexual and body-build aberrations are less common. Though such persons usually are infertile, pregnancy, on rare occasions, may be possible.

Pure gonadal dysgenesis refers to Turner's type of streak gonads with a minimum of body abnormalities. Thus, the patient may have immature breasts, and height may be average or above average.

Mixed, or atypical, gonadal dysgenesis is characterized by an undifferentiated streak gonad on one side and an intra-abdominal testis on the other. This combination is related in part to true hermaphroditism and in part to testicular feminization. The clitoris may be enlarged. Chromosomal pattern may be XO/XY, XY, or one of a number of other combinations.

Hormone levels in gonadal dysgenesis

Growth hormone levels are normal in gonadal dysgenesis, though it has been suggested that they may be high and ineffective in stimulating growth. Thyroid hormone supplies are normal, but thyroidal trapping of iodine may be abnormal, the frequency of thyroiditis is increased, and myxedema can result. Cortisol supplies are normal, but there is evidence of decreased production of adrenal androgens after ACTH administration. Urinary 17-ketosteroids tend to be low-normal because the ovarian contribution is low or lacking. Urinary and serum FSH and LH measurements are generally, but not always, high. Estrogen and progesterone excretion is diminished. The prevalence of diabetes mellitus appears to be increased.

Turner's syndrome may occur with a male body type. In general, these persons are characterized by underdeveloped testes, decreased or absent sperm formation, stunted growth, webbing, skeletal abnormalities, and variably sparse body hair. Information on sex chromosomes is limited, but XO patterns have been found. (See also BIRTH DEFECTS AND CONGENITAL DISORDERS.)

INTERSEXES

In addition to the Klinefelter and the Turner types of defective testis and ovarian development usually associated with abnormalities of sex chromosomes and the female pseudohermaphroditism of the adrenogenital type already described, several other aberrations of gonadal and genital structures have been recognized. These include male pseudohermaphroditism, true hermaphroditism, ill-defined abnormalities of genitalia in association with a tumour of the gonad, and total absence of testicular or ovarian tissues.

Male pseudohermaphroditism. In male pseudohermaphroditism, the individual is a genetic and gonadal male, but the body build and external genitalia have female features. In its most marked form, this is classic testicular feminization, but other less severe types also occur. In testicular feminization, immature and infertile testes secrete normal androgens, but the tissues are unresponsive to the male sex hormone; body build is feminine with short vagina and breasts, but sex hair is absent. In the lesser forms of the disorder—the Lubs, the Gilbert–Dreyfus, and the Reifenstein types—body build is female in the first, and male body and genital configurations are present in the other two. Breasts develop and sex hair is present in all three of these subgroups.

Dysgenetic male pseudohermaphroditism is a nonhereditary subtype. Such persons have poorly differentiated but identifiable testicular tissue, internal ducts may be male or female, external genitals are poorly developed male, body build may be male or female, and chromosomal pattern is XY or XO/XY. Some of these individuals may be fertile.

Hermaphroditism. In true hermaphroditism, in about one-half of the individuals, separate collections of ovarian and testicular tissue are present, usually in the abdomen and scrotum, respectively. In the rest, an ovotestis is found, and, more often than not, such a combined organ is accompanied by another discrete gonad. The latter is usually an ovary. The ovarian tissue generally appears more normal than the testis, and ovulation may occur. A uterus is generally present, but the ducts usually correspond to the adjacent gonad. The external genitals are largely male, but hypospadias (*i.e.*, the urethra opening on the underside of the penis) is frequent. Breasts usually develop at puberty, and menstruation occurs in about one-half of these individuals. The chromosomal pattern of persons with true hermaphroditism is generally of the female type; male patterns and mosaic or other abnormalities, however, do occur.

Predominance of female organs in hermaphroditism

Gonadal neoplasm. It has been suggested that gonadal tumours such as the dysgerminoma (an ovarian or testicular tumour with undifferentiated germinal cells) also produce hormonal factors that affect normal female sexual development. Thus, precocious puberty has occurred in some instances of dysgerminoma, and increased levels of urinary gonadotropins have been reported. Masculinization has also been observed.

Absence of genital tissues. Genital structures, external and internal, may be entirely absent in persons with a female or male body type. Even in such instances, however, there may be some scattered cells from the primitive genital tissues that secrete sex hormones.

OTHER DISORDERS OF THE OVARY

Stein–Leventhal syndrome. The polycystic ovarian, or Stein–Leventhal, syndrome consists of cystic changes in the ovaries variably associated with absent, scant, or excessive menses, lack of ovulation, infertility, increased facial and body hair, and overweight. It has not been established that such polycystic changes represent a single entity, since they can be produced in animals by transplantation, irradiation, or subtotal removal of the ovaries and can occur spontaneously in humans. At least some instances, however, represent enzymatic disorders of steroid hormone production in the ovaries, resulting in the production of excess male hormones. In one of these, the enzymatic defect interferes with the production of estrogen by the ovaries, and the precursor steroids are converted to hormones with androgenic effects.

Signs and symptoms of polycystic ovary

The polycystic ovary is usually, but not always, enlarged, and the capsule is thickened. The ova do not mature, and neither ovulation nor corpus luteum formation occurs. The endometrium of the uterus and the vaginal cells fail to develop the normal post-ovulation progesterone effects (see also EMBRYOLOGY, HUMAN; MENSTRUATION; PREGNANCY). The pituitary fossa of such persons is not unusual. The serum calcium, phosphorus, and other solutes are within the normal range. Thyroid function is not disturbed. Occasional diabetic-like glucose tolerance curves are encountered. Serum LH is often intermittently increased, while FSH is normal or low.

Urinary 17-ketosteroids and 17-hydroxycorticosteroids are normal or slightly increased. Plasma cortisol and urinary cortisol derivatives are normal or, if altered, assume patterns seen in simple obesity. Urine pregnanediol, derived from progesterone produced by the corpus luteum, is low in those who are not ovulating.

The diagnosis of polycystic ovaries is certain only with laparotomy (exploratory surgery), but pelvic examination, culdoscopy (pelvic endoscopy), or gynecography (X-ray diagnosis) can also provide satisfactory evidence.

The excesses of ovarian androgenic steroids can be reduced by oral contraceptive steroids. Polycystic ovarian syndrome is also treated by cortisol-like steroids that reduce adrenal androgen and may restore menses and fertility. Wedge resection of a part of each ovary often corrects amenorrhea and infertility. The gonadotrophic inhibiting drug clomiphene or human menopausal gonadotropin (FSH) and human chorionic gonadotropin (HCG) may produce ovulation.

Hyperthecosis ovarii. Hyperthecosis ovarii, characterized by lutein-like cell proliferation deep within ovaries away from the sites of ovum maturation, is a disorder related to the polycystic ovarian syndrome. This change has been described chiefly in women in the second and third decades of life with or without diabetes mellitus, hypothyroidism, pelvic irritative disease, Brenner tumour (nodular ovarian tumour), cancer of the breast, or related disorders. Individuals with hyperthecosis ovarii have androgenic manifestations similar to those seen with polycystic ovaries.

TUMOURS

Origin of testicular tumours

Tumours of the testes. About 97 percent of the tumours of testes arise from the germinal epithelium. Apart from a variable tendency to increase gonadotropin secretions, patients with germinal cell tumours do not have any discernible endocrine changes. The remaining 3 percent of the tumours of the testes arise from the interstitium and secrete gonadal or adrenal steroids. Rarely, the testis may also be the site of other tumours, such as fibroma and sarcoma.

Of the germinal tumours of the testes—the seminoma, embryonal carcinoma, teratoma, teratocarcinoma, and choriocarcinoma—only the last may secrete large quantities of gonadotropins. Increases in urinary gonadotropins, however, are present in about one-third of the individuals with seminomas, embryonal carcinomas, and teratomas. It is likely that these tumours contain elements that elaborate gonadotropins, but other sources are not excluded. The gonadotropins have FSH properties in the case of seminomas and HCG effects in the case of carcinomas. The teratomas may also secrete estrogens and placental lactogen, and TSH has been found in choriocarcinomas.

Tumours originating from the interstitial cells of the testis usually produce androgens. The excesses of such steroids accelerate masculinization, decrease the size of the uninvolved testis, induce breast enlargement, and decrease spermatogenesis. Occasionally, such tumours produce excesses of estrogens. Rarely, adrenal rest (enclave) tumours develop in the testis, resulting in excesses of cortisol steroids and Cushing's syndrome.

Tumours of the ovary. About 10 percent of the tumours occurring in the ovary secrete male or female sex hormones. The feminizing tumours of the ovary exceed the masculinizing types in a ratio of about 4 to 1. The granulosa cell, the theca cell tumours, and occasionally others, produce feminization; the malignant arrhenoblastomas, the adrenal rest tumours, and most hilus cell (Leydig) tumours result in masculinization. Exceptions can be cited, however, in most categories. Some few miscellaneous tumours—such as the dysgerminoma (seminoma) associated with abnormalities of the genital and the thyroid-tissue-containing dermoid cyst or teratoma known as struma ovarii—also belong in the hormone-secreting category. At various times it has been suggested that others such as the Brenner tumour also secrete hormones.

The remaining 90 percent of ovarian tumours are of the types that do not produce hormones (the cystadenoma or adenocarcinoma category of the pseudomucinous and, especially, the serous types).

THE BREAST IN ENDOCRINE DISORDERS

Gynecomastia. In true gynecomastia—*i.e.*, enlargement of the breast in males—the ductules and surrounding tissues are increased in quantity. Pseudogynecomastia, on the other hand, merely represents an excess of fat in the breast area. True gynecomastia with and without hypogonadism usually occurs sporadically, but it has also been reported in father–son–cousin and other familial distributions. Pseudogynecomastia is common among overweight males.

True gynecomastia and pseudogynecomastia

The most frequent form of true gynecomastia, the subareolar node seen or felt in boys during puberty or adolescence, is usually a feature of normal development. It is present in minor degree at some time in all normal boys and need not be symmetrical. Such physiological gynecomastia appears to result from increases in androgens, estrogens, adrenal steroids, gonadotropins, or other hormones that normally occur during pubescence.

The most common endocrine disorder in older boys and adults with marked true gynecomastia is Klinefelter's syndrome, discussed previously. Increases in breast tissue have also been described in various other endocrine disorders, including pituitary tumours, acromegaly, hyperthyroidism, adrenal tumour, Cushing's syndrome, Addison's disease, tumours of the testes, cryptorchidism and other genital abnormalities, and diabetes mellitus. Also, androgen or estrogen therapy can produce true gynecomastia.

Gynecomastia that occurs in untreated cirrhosis or during the treatment of cirrhosis with a high-calorie–high-protein diet probably results from excesses of estrogens. Also, resumption of a normal caloric intake after prolonged starvation may be associated with enlargement of the male breasts and tenderness of the breasts and testes.

Enlargement of the breasts has been noted at birth, after surgical removal of the prostate, in ulcerative colitis, congestive heart failure, skin disorders, vitamin D poisoning, rheumatoid arthritis, cancer of the lung or kidney, Hodgkin's disease, leukemia, myotonia dystrophica, arthritis, paralyses, and during treatment with digitalis, amphetamine, and other drugs. Most of the above associations could be fortuitous, but a causal relationship was probably present in those patients with a history of estrogen, androgen, or other drug therapy.

Removal of the cause, such as estrogen excesses, is followed by some recession of the gynecomastia. In others, surgical excision may be dictated by symptoms, cosmetic reasons, adverse effects on the personality, fear of cancer, and other reasons.

Galactorrhea. Galactorrhea, or secretion of milk-like fluid from the breasts, can occur in either sex. Though rare, the possibility of galactorrhea as a result of manipulation or breast cancer must always be raised. In males, galactorrhea has been associated in all but a few instances with either an intracranial tumour (acromegaly, pineal tumour with sexual precocity, angiosarcoma of the pituitary) that presumably removes the prolactin inhibitor of the hypothalamus and permits prolactin secretion by the pituitary or with a cancer (of the testis, adrenal, or kidney) that presumably secretes a prolactin-like hormone or related materials or increases androgen or estrogen secretion. Galactorrhea in females may follow pregnancy or, in women who have not been pregnant,

may develop in association with brain lesions, enlargement of the pituitary gland, other endocrine disorders, or appear after surgery or drug therapy.

Galactorrhea in females

In females, persistent lactation without suckling, which follows upon a recent pregnancy, is called the Chiari–Frommel syndrome. Malnutrition, psychiatric disturbances, and amenorrhea can be present, and the uterus and ovaries may be small.

Galactorrhea in a female who has never been pregnant is termed the Ahumada–del Castillo or the Argonz–del Castillo syndrome. Such galactorrhea appears to result from excesses of secretion from the pituitary eosinophils. This implies that excessive secretion by the anterior pituitary eosinophilic cells can result in two syndromes, acromegaly or lactation, and that both may occur in the same individual. This possibility is supported by the demonstration that human growth hormone does possess prolactin activity and that injections of growth hormone increase galactorrhea. On the other hand, growth hormone levels may or may not be increased in the plasma or serum of individuals with galactorrhea.

When a pituitary tumour is found in a female with galactorrhea as the sole manifestation, an Albright syndrome is said to be present. In addition to, or perhaps as a subdivision of, the Albright syndrome, galactorrhea in women has been associated with acromegaly or with other brain tumours, including craniopharyngioma and basophil adenoma. Endocrinopathy other than acromegaly may also be found in females with galactorrhea. This may be hypopituitarism, Cushing's syndrome, Addison's disease, goitre, hyperthyroidism, myxedema, or hypoparathyroidism.

Galactorrhea in females has also been reported following encephalitis, pituitary stalk section, various surgical procedures, pneumoencephalography, and chlorpromazine, reserpine, birth-control pill, progesterone, or other drug therapy.

Therapy of galactorrhea in both sexes includes removal of the cause when it is known and administration of estrogen–progesterone combinations (see further REPRODUCTIVE SYSTEM DISEASES).

VI. The pancreas: diabetes mellitus

Levels of blood sugar are in great measure controlled by two hormones, insulin and glucagon, both produced by the pancreatic islet cells. An excess of insulin produces low blood-sugar levels, and excesses of glucagon raise them. Losses of insulin secretion or of its effect result in diabetes mellitus. Tumours of pancreatic islets of the insulin-secreting type (insulinoma) produce low blood-sugar levels. Those made up of glucagon-secreting cells result in high blood sugars and even diabetes.

Cells within the islet tissues of the pancreas elaborate a protein hormone, insulin, that lowers blood glucose levels and stimulates the formation of fat and protein. Diabetes mellitus is usually a genetic, and less frequently an acquired, disorder of the metabolism of carbohydrates, fats, and proteins resulting from a relative or an absolute shortage of insulin or of insulin action. It is most readily recognized by the presence of unduly high blood-glucose levels in the fasting state or following a glucose or food load. In uncontrolled diabetes, however, fat formation is also diminished with an increase in lipolysis; *i.e.*, the amount of fat mobilized from the tissues. Similarly, in severe untreated diabetes, protein formation is decreased, and negative balances of tissue proteins may ensue.

Prevalence of diabetes mellitus

As a result of community and other surveys, the prevalence of diabetes mellitus in North America is estimated at about 2 percent, but in the course of an individual's lifetime, the probability of developing diabetes is higher. Almost all children with diabetes, and most of those adults with the more severe insulin-dependent type of diabetes, have probably inherited the trait from both parents. Those who have the trait from only one parent are carriers and may or may not develop the disease. Acquired, or nongenetic, diabetes is much less frequent than the genetic type. It is most clearly identifiable when it follows upon removal or destruction of the pancreas or is caused by such disorders as Cushing's syndrome, acromegaly, pheochromocytoma, hyperthyroidism, or other disorders.

The diagnosis of diabetes mellitus is most often based on the finding of elevated blood-glucose or serum glucose levels obtained in the fasting state or during a glucose tolerance test. The glucose is administered by mouth or intravenously in a variety of dosages, and blood or serum samples are analyzed for glucose drawn during one of each of several time schedules. Interpretation of the results of an oral glucose tolerance test in diagnosing glucose intolerance or diabetes mellitus is simplified when the results are expressed as the Glucose Tolerance Sum for two or for three hours ($GTS_{0\text{-}2\ hr}$ and $GTS_{0\text{-}3\ hr}$).

The $GTS_{0\text{-}2\ hr}$ value is the sum of the 0-, ½-, 1-, and 2-hour glucose levels in venous blood when 1.75 grams of glucose are administered per kilogram of body weight. When $GTS_{0\text{-}2\ hr}$ is 500 or less, there is a 98 to 100 percent probability that the test is clearly normal by the standards of the World Health Organization, the British Diabetic Association, and the United States Public Health Service. On the other hand, when the $GTS_{0\text{-}2\ hr}$ is 801 or higher, the test virtually always fulfills each of the sets of criteria cited above for diabetes. Since values between 501 and 800 are variably classifiable as normal or abnormal by these same standards, this range represents the equivocal zone.

Tolerance tests with GTS values in the equivocal zone should be repeated at intervals of months or years.

CLASSIFICATION OF DIABETES MELLITUS

Diabetes mellitus occurs in man as a spectrum with seven stages based on the intensity of the carbohydrate and fat disturbances.

The seven stages of diabetes mellitus

Stage I is prediabetes—*i.e.*, the phase preceding the appearance of any carbohydrate disturbance. It is more likely to occur in individuals with the following individual or family histories: diabetes, obesity, abortions, stillbirths, newborn deaths, birth of large babies, or pregnancies with high blood sugars, excesses of amniotic fluid, and toxemia.

Stage II is stress diabetes. It is present during a period of stress and usually disappears after the stress is removed. Such stress may include infection, trauma, adrenal steroid excesses, or drug therapy.

Stage III is the most common undiagnosed form of glucose intolerance and is the earliest phase of what may turn into fully developed diabetes mellitus. It is characterized by normal fasting blood-glucose levels with unduly high blood-sugar levels after a meal or after glucose administration.

Stage IV is the most commonly recognized form of diabetes mellitus. The fasting blood glucose is elevated, and the after-meal and glucose tolerance test blood-sugar concentrations are also unduly increased. In Stages III and IV, there are no ketone bodies (products of fat metabolism) in the urine or blood plasma. This cardinal feature sharply delimits these two stages from Stages V, VI, and VII of diabetes, each of which is characterized by the presence of ketone bodies.

Stage V, or ketosis, is characterized by ketone bodies in the urine and high blood-sugar levels. There is, however, no accumulation of ketone bodies in the plasma. Respirations are normal.

Stage VI is diabetic ketoacidosis. Ketone bodies are present in the urine and are accompanied by ketone accumulation in the plasma and reduction of the serum carbon dioxide content. Slight to moderate decreases in serum pH (7.3 to 7.1) are present (pH is a measure of acidity and alkalinity, a reading of 7.0 being neutral, values above are alkaline, values below are acidic). Rapid respirations are common, and deep (Kussmaul type) overbreathing may develop.

Stage VII is diabetic acidosis coma with marked accumulation of ketone bodies in plasma and urine, decrease in serum carbon dioxide to less than ten milli-equivalents per litre, and reduction of serum pH to values as low as 6.8 or 6.9. Deep overbreathing is almost always present.

In the mildest type of glucose intolerance or early diabetes (Stage III), the fasting blood-sugar levels are nor-

mal, and marked rises in blood glucose appear only after a carbohydrate or food load. In other words, in the basal state the amounts of sugar released from glycogen stores or produced from protein can be metabolized without raising the blood sugar. In a more advanced phase (Stage IV), the residual ability to dispose of carbohydrate of endogenous origin is further limited, so that the blood sugar is high even in the fasting state. In Stages III and IV, however, there is still considerable residual ability to utilize glucose as a fuel. Indeed, by virtue of the law of mass action, the high blood-sugar level enhances the disposal of glucose. In these mild stages of diabetes, day-to-day survival is not in jeopardy.

Fat utilization in advanced stages of diabetes

In the more advanced stages of diabetes (Stages V, VI, and VII), the tissues cannot utilize glucose in amounts adequate to meet energy needs because insulin supplies or effects are insufficient. Hence, fatty acids become the fuel because they can be used when insulin action is absent or insufficient. Marked shifts to fat as a fuel create problems, however, in addition to those attributable to insufficient utilization of glucose. Fatty acids are mobilized from fat depots of the body in excess of the ability of the tissues to use them. Derivatives of these fatty acids (β-hydroxybutyric and acetoacetic acids) accumulate in body fluids with acetone, a derivative of acetoacetic acid, present as an easily identifiable indicator. This indicator is the indisputable signal of undue reliance on fat as fuel because of inability to utilize carbohydrate in adequate amounts. It means that the diabetes has reached the most advanced stages, and survival is threatened. In addition to the underutilization of glucose and overmobilization and underutilization of fat, diabetes in these advanced stages is characterized by disturbances of protein metabolism. Protein formation is decreased, protein catabolism (breakdown) is increased, and the resultant amino acid excesses are converted to sugar or ketones.

OCCURRENCE OF DIABETES

Diabetes in adults. Most instances of adult-onset diabetes occur in overweight persons and are relatively mild (Stage III and IV), with unduly high blood-sugar levels as the chief manifestation; ketone bodies are absent. The shortage of effective insulin is relative rather than absolute. Despite the high blood sugar, there is enough insulin action and sufficient residual ability to extract energy from carbohydrate; ketosis, therefore, does not occur. The complete absence of symptoms, or even signs in the earliest phase of diabetes, has resulted in such terms as latent or chemical diabetes. Diabetes of maturity onset, however, may also produce enough of an increase in blood glucose to increase urine volumes (polyuria), thirst (polydipsia), food intake (polyphagia), weight loss, and other symptoms. Most individuals with adult-onset diabetes are overfed, underactive, and overweight. Significant alterations of one, two, or all three of these variables can eradicate laboratory evidence of diabetes. Treatment begins with close scrutiny and adjustment of the diet in terms of calories and foodstuffs, administration of insulin as needed for maximal control of diabetes; use has been recommended of certain oral drugs, such as the sulfonylureas or phenformin, to facilitate glucose utilization in those in whom such medications are indicated.

Rarely, under the stress of illness, trauma, pregnancy, and drug or adrenal-steroid therapy, the ability to extract energy from carbohydrate decreases, and the tissues turn to fat as a fuel. Fatty acids are then overmobilized, and ketone bodies may appear. Some few individuals with maturity-onset diabetes are ketoacidosis prone and insulin dependent, even in the absence of stress or illness. Such persons are generally not overweight. They cannot be treated by caloric restriction and weight reduction; they must receive insulin.

Less usual forms of diabetes observed in the adult include the lipoatrophic type, in which body fat is decreased, serum cholesterol and triglycerides are high, cirrhosis is common, and insulin is mandatory. Mild and generally non-insulin dependent diabetes may develop with the adult type of premature aging—a condition referred to as Werner's syndrome.

Diabetes in children. In children, obesity is not a clue to diabetes, as it may be in adults. Diabetic children generally are not overweight, nor is there a history of overweight. Respiratory infections are common antecedents of symptoms and signs that lead to diagnosis of childhood diabetes. This sequence, of course, is of little practical value because of the frequency of such infections in this age group. As in adults, increased urine volumes and increased thirst are the two most common symptoms, but they occur more frequently in juvenile diabetes. Increased food intake, weight loss, and a tendency of the patient to tire easily are next in the order of frequency. Irritability, onset of acidosis or coma, vomiting, abdominal pain, and recurrence of bed-wetting are other indices that may be indicative of juvenile diabetes. The diabetes of childhood differs from that in adults in that it is more severe and almost always requires insulin in addition to diet therapy.

Examples of transient diabetes have been observed in infants. In some of these, respiratory or skin infections were present. The origins of such transient diabetes are obscure.

CAUSE AND TREATMENT OF DIABETES

Diet and insulin therapy. A consistent intake of calories, predictably distributed among carbohydrates, fats, and proteins and divided into the three main meals and between-meal and before-bedtime snacks, is the keystone of successful regulation of diabetes with diet alone, with insulin, or with other hypoglycemic agents.

Types of insulin

In North America and in many countries of western Europe, some or all of the following types of insulin are available for the treatment of diabetes: Protamine Zinc, which has low intensity but long action; NPH Ultra-Lente, Lente, Semi-Lente, and Globin, which exhibit intermediate intensity of action and duration; Regular or crystalline, which gives rapid action of short duration.

Diabetic ketosis, ketoacidosis, and acidosis coma must be managed with the last of these. Also, this is the insulin used in multiple daily injections for the treatment of "brittle" (unpredictable or labile) diabetes. Prophylaxis of ketosis, ketoacidosis, and coma in insulin dependency and in attempts at optimal treatment of all types of diabetes is effected with the intermediate types in a variety of programs: NPH, Lente, or Globin alone or with Regular; Lente and Semi-Lente or Ultra-Lente; Lente and Regular, and other combinations. Protamine Zinc Insulin (PZI) is still used by persons and physicians who found this form suitable before the other types became available. PZI, however, has a cumulative effect that, in the course of time, can produce dangerous hypoglycemia.

Special types of insulin may be employed to treat allergic reactions and undue resistance to insulin, *viz.*, pure pork instead of the usual beef–pork preparation, or dealanated pork insulin from which the terminal amino acid, alanine, has been removed, leaving a 50-amino acid moiety that is chemically indistinguishable from the same segment of human insulin.

The chief mechanism of sulfonylurea action is an increased secretion of endogenous insulin and a decrease in endogenous glucagon, the pancreatic hormone that raises blood-sugar levels. It can be used as the sole blood-sugar lowering agent in conjunction with a diabetic diet in non-insulin dependent diabetes. Phenformin, on the other hand, does not exert its effects by increasing insulin output but does enhance glucose utilization, provided that insulin supplies and insulin action (endogenous or exogenous) are adequate. It also controls the rate at which foodstuffs are absorbed from the gut. Hence, it is rational to employ a diabetic diet and to use phenformin as the sole blood-sugar lowering agent in non-insulin dependent diabetes, as an adjuvant in insulin therapy of non-insulin dependent diabetes, and as an aid in the regulation of diabetes dependent on the administration of exogenous insulin. Also, it is reasonable to use phenformin and a sulfonylurea drug in diabetes of Stage III or IV, which are non-insulin dependent.

Diabetic acidosis and coma. Inadequate diabetic regulation resulting from a relative or absolute shortage of

effective insulin is the ultimate cause of diabetic ketoacidosis or coma. Such a degree of inadequate regulation of diabetes may be a consequence of an insufficiency of insulin or its action in previously untreated diabetes, omission of the usual injection of insulin, or failure to increase the dosage when the requirements for insulin rise. Increases in the need for insulin occur during inactivity, major or minor infections, insulin shocking, insulin resistance, emotional trauma, pregnancy, surgical operations, thyroid crises, and other stresses. If the increased need for insulin is not met, the blood sugar rises, and then mild ketosis, ketoacidosis, and acidosis coma ensue.

Periods requiring increased insulin

A history of vomiting can be obtained in three-quarters or more of the children and adults suffering from diabetic ketoacidosis. Abdominal pain may be a result of, or occur with, the nausea and vomiting in 40 or more percent of these individuals and can be intense enough to mimic abdominal disease. Deep overbreathing (of the Kussmaul type) is one of the manifestations of diabetic ketoacidosis. It generally appears when the blood serum carbon dioxide is reduced to about ten milli-equivalents per litre. Shallow and rapid respirations, however, precede the deep overbreathing. Coma can then develop. Circulatory collapse may result as a consequence of dehydration and salt depletion. Weakness, paralyses, or disturbed sensations resulting from high or low serum potassium levels or other causes may characterize early acidosis or coma; these symptoms and signs can also develop during the early hours of treatment.

Several types of non-ketoacidotic coma may occur in diabetes mellitus. These include: (1) coma as a result of a cerebral hemorrhage or occlusion, drug excesses, kidney failure, or advanced liver disease; (2) coma caused by extremely low blood-sugar levels due to excess insulin or other anti-diabetic therapeutic measures; (3) lactic acidosis following tissue-oxygen shortages; (4) hyperosmolar coma as a consequence of high levels of glucose, of sodium, or both; and (5) cerebral edema with coma following treatment of ketoacidotic coma.

Long-term concomitants and complications of diabetes. Diabetes mellitus, usually of many years' duration, may be accompanied by thickening of the basement membrane of small blood vessels and by accelerated and exaggerated atherosclerosis of the larger vessels. Basement membrane changes in capillaries are responsible for the small blood vessel disorders in the kidneys, eyes, and nerves and are present in other tissues such as the conjunctiva, skin, and muscle. The atherosclerosis may result in such conditions as myocardial infarction, gangrene, or stroke. Diabetes also predisposes to infections, necrobiosis, atrophy of fat, contractures of the hands, cancer of the pancreas, and pregnancy complications.

Effects on blood vessels

The appearance and intensity of the small blood vessel basement changes are definitely related to the duration of the diabetes and less clearly to the adequacy of diabetic regulation. As far as lesions of the large blood vessels are concerned, atherosclerosis tends to appear earlier, is more severe, and produces myocardial infarction, stroke, gangrene, and other serious consequences more frequently. Two variables predispose to the increased atherosclerosis. One is genetic: high blood pressure and accelerated atherosclerosis develop more frequently in nondiabetic relatives of persons with diabetes than in the general population. The other variable is metabolic: unduly high serum fat or lipids—*i.e.*, triglycerides and cholesterol—levels occur in diabetes and predispose to atherosclerosis. Such lipid excesses are treated by regulation of diabetes, caloric restriction, diet manipulation, and administration of certain drugs that lower serum lipids.

Pancreatic islet cell tumours. Pancreatic islet cell tumours that secrete excesses of insulin (insulinomas) may be multiple and can become cancers. In persons with an islet cell tumour, unduly low blood-sugar levels develop spontaneously or may follow upon prolonged fasting or exercise. Stimuli such as tolbutamide, leucine, or glucagon are used to support a diagnosis of an islet cell tumour. Administration of each of these agents produces a rise in serum insulin and a fall in blood sugar in normal individuals and excessive responses in those with an insulinoma. Therapy consists of surgical removal of the tumour.

Various combinations of islet cell tumours with tumours of the pituitary, parathyroids, adrenal cortex, adrenal medulla, and thyroid have been recorded in sporadic or familial instances.

Glucagon in diabetes. Since glucagon raises blood-sugar levels by releasing liver glycogen, it is not surprising that pancreatic islet cell tumours made up of glucagon-producing cells can result in diabetes mellitus. Also, levels of glucagon in persons without such a tumour play a role in determining the concentrations of blood sugar. As already cited, the sulfonylurea drugs are used in the treatment of diabetes mellitus to reduce glucagon and to increase insulin levels in the serum.

VII. Other endocrine glands

THE PINEAL GLAND

In man, the pineal is a pear-shaped midline structure located at the back of the base of the brain. It becomes calcified by the sixth decade in about 70 percent of the population. In the past, a variety of roles has been tentatively assigned to the pineal, including the regulation of growth, the size of the endocrine organs, sexual maturation, blood-sugar levels, and skin pigmentation. Attempts have also been made to implicate it in psychoses and epilepsy.

Effects of pineal extracts

More recent studies indicate that a number of principles or effects can be elicited, some in animals and some in man, with preparations derived from the pineal, but their physiologic significance, if any, in man is not established. These include a melanocyte contracting principle (melatonin), an accelerator of adrenocortical aldosterone secretion (glomerulotropin), and an inhibitor of estrus in rats (anestrin) that may be the same as melatonin. Also, serotonin and metabolites such as 5-hydroxyindoleacetic acid (5-HIAA) and 5-methoxyindoleacetic acid, histamine, and 8-lysine vasopressin have been demonstrated in extracts of the pineal.

In the rat, light striking the retina sets up nervous impulses that reduce the melatonin in the pineal. The decrease in melatonin reduces its inhibiting effect on the rate of estrus cycle. Hence, this represents a transducer system whereby nervous impulses affect humoral or hormonal secretion akin to the supra-optic–hypophyseal system that controls vasopressin output and the hypothalamic releasor and inhibitor systems that act on the anterior pituitary.

Tumours of the pineal may be associated with accelerated and, less often, with retarded sexual maturation. True parenchymatous pinealomas are generally present in those with repressed sexual maturation. On the other hand, sexual precocity occurs in conjunction with nonparenchymatous tumours, such as gliomas and teratomas, that destroy the pineal. The sexual precocity occurs almost exclusively in males.

THE THYMUS

Studies in animals have established that the thymus, a collection of cells behind the breastbone, must be present in the first few days of life for the subsequent appearance of certain immune responses and antibodies. The thymus may be the source of lymphocytes that migrate to the lymph nodes, spleen, and other sites and participate in immunologic maturation. Simultaneously or alternatively, the thymus may elaborate a principle that stimulates lymphocyte formation in other body sites. Normally, the thymus decreases in size beyond the neonatal period but continues to play an immunogenic role. If the thymus is removed shortly after birth, the ability to develop normal lymphoid tissue, small lymphocytes, and plasma cells and to reject tissue transplants from donors of foreign strains or different species is reduced or lost.

The thymus as a source of lymphocytes

A number of auto-immune disorders present in animals have been related to a defective or absent thymus. These include auto-immune hemolytic anemia and systemic lupus erythematosus. In man, agammaglobulinemia and the increased frequency of certain types of leukemia, rheumatoid arthritis, and Hodgkin's disease may also be

related to defective immune responses in which the thymus plays or has played a role. Excesses of cortisol hormones, as in Cushing's syndrome, are associated with decreased immune responses and increased frequency and severity of infections. This is mediated, in part at least, by the lymphopenic action of cortisol excesses, an effect akin to the thymolytic effect of cortisol.

Several disorders in man have been associated with enlargement of the thymus, including myasthenia gravis, deficiency of gamma globulins with increase in infections, decreased red blood cell formation, increased destruction of red blood cells by hemolysis, and shortage of circulating white blood cells of the small lymphocyte type. An auto-immune component has been identified or postulated in each of these. Also, several endocrine disorders—including thyrotoxicosis, thyroiditis, Addison's disease, and hypoparathyroidism—have auto-immune components in which the thymus at one time or another has played, or still plays, a role. Thus, there is evidence that, in thyrotoxicosis and thyroiditis, thymic parenchyma may be enlarged. The exact mechanisms, however, have not been defined.

It is possible that some of the sudden deaths in infancy initially attributed to enlargement of the thymus may have represented adrenocortical insufficiency of the Addisonian crisis type, since in the absence of cortisol-type steroids, the thymus enlarges. Thus, removal of the pituitary in animals results in atrophy of the adrenal glands and increases the thymus and lymphoid tissues elsewhere in the body. Also, the enlarged thymus present in infants born with massive brain abnormalities is probably attributable to the associated adrenal atrophy.

In animals, extracts of the thymus may increase or decrease sugar metabolism, promote or retard normal growth, regulate fertility, and accelerate or halt the spread of cancer, but the factors and mechanisms involved require verification.

BIBLIOGRAPHY. T.S. DANOWSKI, *Outline of Endocrine Gland Syndromes*, 2nd ed. (1968), presents in summary and reference form the findings in the various endocrine disorders that occur spontaneously or are induced in man. The same author's *Diabetes Mellitus, with Emphasis on Children and Young Adults* (1957), contains detailed information on the features of the type of diabetes mellitus that usually occurs in the younger years. T.S. DANOWSKI et al., "Utility of Equivocal Glucose Tolerances," *Diabetes*, 19:524–526 (1970), is a discussion of the resolution of much of the uncertainty in reconciling the interpretation of glucose tolerance tests in normal and diabetic persons by use of the Glucose Tolerance Sum principle. R.C. KHURANA *et al.*, "Insulin and Glucose Patterns in Control Subjects and in Proved Insulinoma." *Amer. J. Med. Sci.*, 262:115–128 (1971), is an article that answers questions concerning the presence or absence of excess circulating insulin levels, by reference to control studies in healthy subjects.

(T.S.D.)

Endocrine Systems

Most animals have specialized organ systems, each of which is adapted for a particular function and usually situated in a particular part of the organism. The functions of organ systems such as the digestive system, reproductive system, and excretory system must be integrated; in addition, the composition of the fluid medium in which all tissues are bathed must be regulated with some degree of constancy, or the complex organisms that depend on them cannot survive.

Two systems, consisting of definable pathways that extend throughout the body of an animal, allow the integration of widely separated functions; they are the circulatory system and the nervous system. The nervous system, by means of nerve-cell pathways, integrates sense organs, which detect changes in the environment, with internal organs such as glands and muscles. The circulatory system, by means of blood-vessel pathways, conveys blood containing a variety of substances to all parts of the body.

The endocrine system is a group of glands the secretions of which, called hormones, integrate the activities of organ systems and help to maintain a constant internal environment. A hormone is secreted from the gland in which it is formed and travels in the blood to a specific tissue or organ, commonly called a target, where it regulates in time, amount, or quality the function of that organ or tissue. The activities of certain endocrine glands may be influenced either indirectly by the secretions of specialized regions of the nervous system called neurohemal organs or directly by nerves. Thus, there is a very close relationship between the endocrine system, which secretes hormones; the nervous system, which regulates endocrine gland activity; and the circulatory system, which conveys hormones to their targets.

Many examples of endocrine regulation have been established in vertebrate and invertebrate animals. During the digestive process, for example, when food enters the intestine, a hormone, secretin, is released from the lining layer (mucosa) of the intestine into the blood. When secretin reaches the pancreas in sufficient quantity, it evokes the secretion of pancreatic juice into the intestine to aid in digestion. The regulatory function of secretin is to coordinate the functions of the intestine and the pancreas during digestion of food. The secretin mechanism, initially described in 1902, was among the first endocrine phenomena to be recognized.

Various other kinds of blood-borne substances (for example, oxygen and minerals) can influence the function of particular organs, yet they are not hormones. Carbon dioxide, for example, which is produced in many parts of the body, may, after transport in the blood, affect certain functions. It is termed a parahormone since it lacks specificity; that is, it has no specific source and occurs generally.

Neurohormones and neurohemal organs

Neurohormones are substances manufactured not by glandular structures, but by nerve cells. The axons of neurosecretory nerve cells—that is, the long processes that usually direct nerve impulses away from the cell body—generally end in a group upon a set of specialized blood vessels. The unit so formed is known as a neurohemal organ. Neuroendocrine phenomena are important in that they provide a mechanism by which external changes, perceived by sense organs, can initiate responses in many organ systems and processes by affecting endocrine glands (see Figure 1). There are three ways in which this can occur:

1. Nerves may lead from sensory cells to neurosecretory cells that release a neurohormone into the blood. In vertebrates, this type of neurosecretory cell is located in a neurohemal organ called the pars nervosa, a region of the pituitary gland, or hypophysis.
2. Nerves may lead from sensory cells to neurosecretory cells that terminate in a neurohemal organ called the median eminence below the brain (hypothalamus). Neurohormones known as releasing hormones, released from the median eminence, are carried by the blood to the pars distalis region of the pituitary gland, where they bring about the release of hormones, which in turn affect a broad variety of processes.
3. Sensory nerves, via a chain of nonneurosecretory nerve cells, may directly lead to pituitary cells, as, for example, in the pars intermedia region of the pituitary gland (Figures 1, 3).

This article describes the endocrine systems of vertebrate and invertebrate animals. The endocrine system of man is dealt with in ENDOCRINE SYSTEM, HUMAN; the article HORMONE contains discussions of the individual hormones secreted by endocrine glands.

GENERAL FEATURES OF VERTEBRATE ENDOCRINE SYSTEMS

Relationships. *Of endocrine organs to each other.* The number of endocrine glands and the variety of their roles may, at the outset, make any discussion of endocrine function seem complex. Figure 3 shows in schematic fashion the functions and relationships of endocrine glands. At the upper left is the pituitary gland, the function of which is controlled by the brain. On the right side of the figure are the other major endocrine glands of vertebrates. Targets (*i.e.*, sites of action) of the pituitary gland hormones may be separated into two types: targets that are themselves endocrine glands, such as the thyroid,

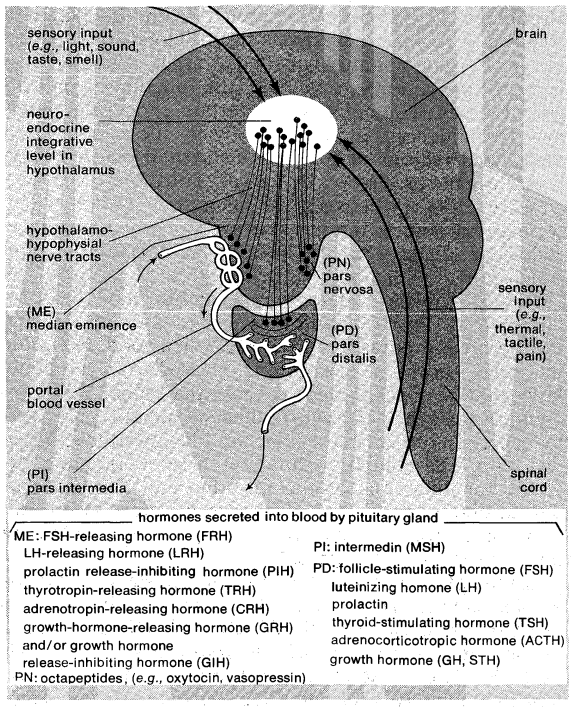


Figure 1: The mechanism by which environmental changes perceived by sense organs initiate responses through neuroendocrine phenomena.

the gonads (ovaries and testes), and the adrenal cortex; and targets that are not—*e.g.*, kidney, uterus, pigment cells, mammary gland, skeleton. Of the major endocrine glands shown in Figure 3, some are under pituitary control; others—for example, the parathyroids, the ultimobranchial gland, the pancreas, the adrenal medulla, and endocrine tissue of the digestive tract—are not. The pancreas, parathyroid glands, and ultimobranchial gland respond to changes in blood levels of particular substances; the adrenal medulla responds to nervous stimulation; and the endocrine tissue of the digestive tract responds to local stimuli not yet identified.

Endocrine glands thus regulate various processes by the hormones they secrete; their respective functions, in turn, are regulated by a nerve supply, by blood levels of metabolic compounds (metabolites), or by other hormones.

Of endocrine organ systems to the blood. The most common arrangement of endocrine gland cells is as branching cords of cells (Figure 2), an arrangement particularly conducive to the efficient release of hormones. Each cord is usually no more than two cells thick; and between the cords are blood-filled spaces, usually of the type known as sinusoids. Only the thyroid gland and neurosecretory systems have cellular arrangements other than this one. The thyroid is made up of follicles, or minute bags of cells (Figure 2), with walls one cell thick. Surrounding these follicles is a close network of blood capillaries. In neurosecretory systems, the secretory ends of nerve cells terminate on or very close to blood vessels (Figure 2), forming neurohemal organs. Neurohemal organs include the median eminence and the pars nervosa of the pituitary gland, the urophysis in fishes, and the sinus gland in crustaceans.

Functions. Endocrine functions are divisible into two phases: (1) function of the gland itself, which is to manufacture and release a hormone in response to a stimulus; (2) function or action of the hormone in regulating activities of particular targets. These functions are dealt with more fully in succeeding sections of this article and in the article HORMONE.

General properties of hormones. *Chemical nature.* Fatlike and protein-like hormones

Generally, vertebrate hormones are either substances called peptides, which consist of two or more nitrogen-containing compounds called amino acids, or they are steroids, which are complex fatlike substances. Although the chemical structures of steroid hormones are similar, slight structural variations determine whether the hormone acts as a corticoid—a name given to hormones formed in the outer region (cortex) of the adrenal gland—as a male hormone formed in the testis, or as a female hormone formed in the ovary. The active principle of the hormone of the thyroid gland is the amino acid thyroxine. The hormone adrenaline, synthesized in the inner region (medulla) of the adrenal gland, is a relatively simple molecule derived from the amino acid tyrosine. The neurohormones whose structures are known are relatively simple molecules; *e.g.*, thyrotropin-releasing hormone contains three amino acids, the hormones from the pars nervosa region of the pituitary gland contain eight or nine amino acids. The gastrointestinal hormones and the hormone intermedin (MSH), formed in the pars intermedia of the pituitary gland, consist of about 19 amino acids; the hormones from the pars distalis of the pituitary gland, the parathyroid gland, and the pancreas are protein in nature, containing larger numbers of amino acids.

The structures of two invertebrate hormones, both found in arthropods, have been completely established. Ecdysone, or molting hormone, is a steroid, and juvenile hormone is a member of a group of fatlike compounds called terpenoids.

Mechanisms of action. The functions of hormones are diverse: some cause target organs to synthesize various substances, including enzymes and other proteins, for activities such as secretion and growth; a few evoke movements such as muscular contraction or pigment redistribution; many regulate the metabolism of particular substances (*e.g.*, glucose, calcium, phosphate, oxygen) either in specific organs or generally throughout the organism; others influence behaviour by affecting the nervous system.

Hormonal function involves interaction with responding cells either at the cell surface or within the cell. In many instances, the hormone acts at the cell surface to cause the release, just inside the cell membrane, of the compound cyclic adenosine monophosphate (cyclic-AMP), which actually stimulates the characteristic hormonal actions. The cyclic-AMP system at the cell surface includes a receptor substance, probably a protein, that binds the hormone; cyclic-AMP, therefore, is formed only in response to a specific hormone.

Some hormones evoke a change—generally an increase—in the permeability of the membrane of the responsive

From A. Gorbman and H.A. Bern, *Textbook of Comparative Endocrinology* (© 1963); by John Wiley & Sons, Inc.

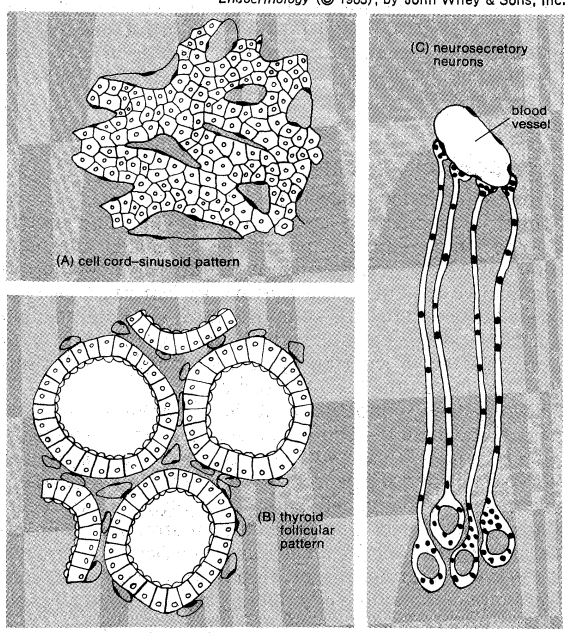


Figure 2: The three most common types of endocrine cell arrangements; each has a different relationship between secretory cell and blood channel into which its hormonal product is released.

cell; others stimulate the transport across the membrane of particular metabolites—*e.g.*, insulin accelerates the transport of glucose; vasopressin increases the transport of sodium and water in the kidney; aldosterone promotes the transport of potassium. In some instances, these actions are mediated through a second messenger, such as cyclic-AMP or through stimulating the synthesis of particular enzymes. Some hormones are known to evoke the synthesis of particular cell products such as structural proteins and enzymes by a mechanism that involves activation of the genetic material of the cell.

VERTEBRATE ENDOCRINE SYSTEMS

Neurosecretory systems. *The hypothalamus and the neurohypophysis (nervous pituitary).* In the lower part of the diencephalon of the brain (hypothalamus) are a number of assemblages, or nuclei, of nerve cell bodies where neurosecretory granules are produced. These granules appear to contain, in addition to the hormones, a protein that acts as a hormonal carrier by attaching to them. The granules can be seen under the light microscope after staining. Stainable neurosecretory material travels principally to the pars nervosa of the pituitary gland and is, therefore, primarily related to the hormones produced by the pars nervosa—vasopressin, oxytocin. Nonstainable neurosecretory material similarly travels from other nuclei of nerve cell bodies in the anterior and posterior hypothalamus to the median eminence. In lower vertebrates some of the neurosecretory material in the median eminence is stainable, even if it is not stainable in the nerve cell bodies from which it arises.

Releasing hormones of the median eminence

The median eminence neurosecretions are called releasing hormones. That at least one releasing hormone exists for each of the six hormones of the pars distalis has been established in the rat and rabbit. The releasing hormones are: thyrotropin-releasing hormone (TRH); adrenocorticotropin-releasing hormone (CRH); somatotropin- (growth hormone-) releasing hormone (GRH) and/or somatotropin release-inhibiting hormone (GIH); luteinizing hormone-releasing hormone (LHRH); follicle stimulating hormone-releasing hormone (FRH or FSHRH); and prolactin-release-inhibiting hormone (PIH). Possibly there may be no separate FRH. The substance that was purified as LHRH releases both follicle stimulating hormone (FSH) and luteinizing hormone (LH).

Five of these brain hormones stimulate the release of their respective hormones from the pars distalis of the pituitary gland; two (PIH and GIH) inhibit the release of hormones from the pars distalis. Destruction of the hypothalamic area where PIH arises increases the amount of prolactin secreted by the pituitary. These relationships have been established for a few mammals that have been investigated. Birds appear to have a prolactin-releasing hormone, and bony fishes appear to have a thyrotropin release-inhibiting hormone in place of, or in addition to, a thyrotropin-releasing hormone.

A special blood vessel supply, the hypophysial portal system, carries blood directly from the median eminence to the pars distalis (see Figures 1 and 3). This arrangement permits the releasing hormones to exert their typical action in the pars distalis before becoming diluted in the general circulation.

The pineal organ. The pineal organ, or body, is one of the most variable in structure of the endocrine glands. It develops from the roof of the diencephalon of the brain, whereas the hypothalamus develops from the lower part and floor of the diencephalon. The role of this organ is not yet fully understood. In some lower vertebrates (*e.g.*, lampreys, reptiles) the pineal has a well-developed eye-like structure. In others, some of the pineal cells, though not organized as an eye, have a photoreceptor type structure, and experiments show that the organ is in fact a light receptor. In lower vertebrates nerves connect the stalk, or base, of the pineal to other nervous centres. In birds and mammals this nervous connection is lacking, but the pineal is connected with the sympathetic part of the autonomic nervous system.

Two principal cell types in the pineal organ are the parenchymal and the glial. The parenchymal cells, which apparently are the endocrine units, contain granules thought to contain the hormone melatonin. Since the parenchymal cells develop from embryonic brain cells, they and their hormone are considered neurosecretory. The formation of melatonin, which is derived from the amino acid tryptophan, depends on an enzyme found only in the pineal organ.

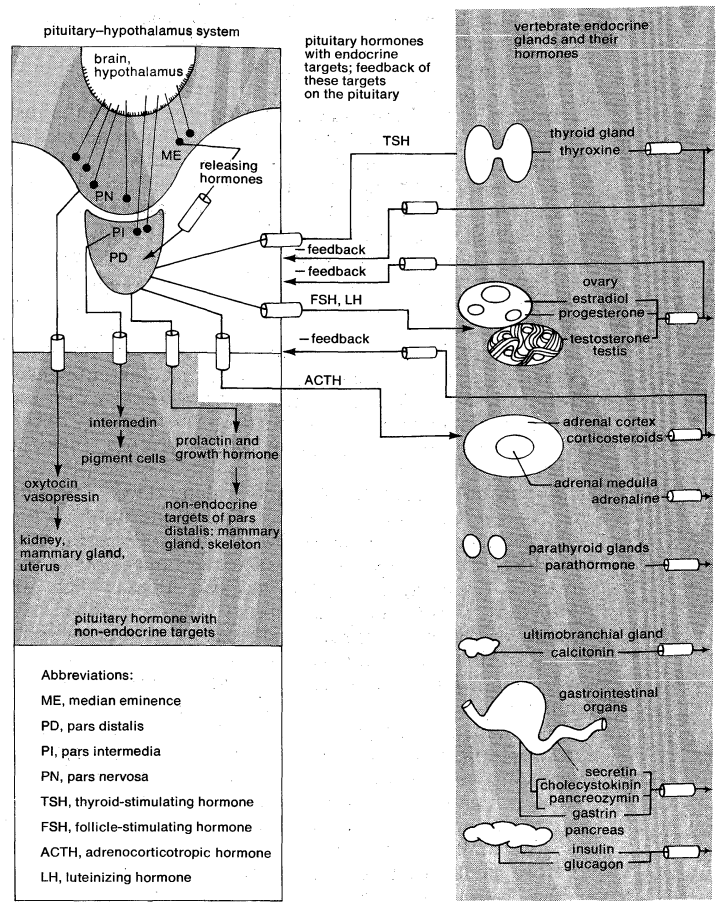


Figure 3: The functions and relationships of the components of the vertebrate endocrine system.

The adenohypophysis. The adenohypophysis develops as a finger of gland tissue that separates from the roof of the mouth cavity and becomes attached to the neurohypophysis and the hypothalamus. In most vertebrates it differentiates into three recognizable zones of tissue, the largest of which is the pars distalis; the next largest is the pars intermedia; and the smallest is the pars tuberalis. The last two may be absent in certain vertebrate groups. All three zones consist of cell cords, but only the pars distalis has an extremely rich blood sinusoid network, supplied by the portal blood vessels from the median eminence that carry the hormone releasing factors.

Hormones of the pars distalis

Six types of cells occur in the pars distalis, each identifiable with one of the six protein hormones found there: thyroid-stimulating hormone (TSH); adrenocorticotropic hormone (ACTH); follicle-stimulating hormone (FSH); luteinizing hormone (LH); growth or somatotropic hormone (GH or STH); and prolactin. Because the target of FSH and LH is the gonads, which are themselves endocrine glands, these hormones are called gonadotropins.

The pars intermedia usually has one type of secretory cell. It contains and apparently secretes the melanocyte-(pigment-cell) stimulating hormone (MSH), a peptide. There is some uncertainty concerning regulation of the release of MSH. Certain neurons from the hypothalamus, abundantly distributed throughout the pars intermedia, terminate directly upon the secretory cells. The compound norepinephrine appears to be liberated at the endings of most of these neurons (*i.e.*, they are adrenergic), but some are neurosecretory. The control of MSH secretion is of interest because the mechanism is the basis for the ability of lower vertebrates (fishes, amphibians, reptiles) to match the colour and sometimes even the pattern of their background.

The pars tuberalis has no known function or hormone. When present, it is generally near the median eminence. In reptiles and birds, in fact, the portal vessels from the median eminence pass through the pars tuberalis to the pars distalis, and the pars tuberalis, therefore, forms a bridge.

The thyroid gland. The thyroid gland develops in the embryo from tissue that grows out of the floor of the pharynx, becomes detached, and eventually completes its differentiation in the neck near the upper end of the trachea. In man, the thyroid has two lobes, which are joined by a bridge (isthmus) that crosses the trachea. An abnormality occurs in some individuals—a duct, or solid mass of tissue, is retained, joining the thyroid to its point of origin, which, in the adult human, is at the back of the tongue.

The shape of the thyroid gland is variable; however, its microscopic units always are follicles (Figure 2) with walls one cell thick. The centre of the follicle is a viscous fluid, which contains the stored hormone, thyroxine, attached to a protein (thyroglobulin). Thyroglobulin does not normally leave the follicle and must be dissociated before the thyroxine bound to it can be released.

The thyroid follicles are among a small number of animal tissues that can selectively take up iodine from the blood. The distinctive functions of thyroid tissue include the ability to accumulate iodine, to couple it with the amino acid tyrosine, and, in a few relatively simple steps, to synthesize thyroxine in the thyroglobulin form. Thyroid-stimulating hormone (TSH) stimulates several steps in this process: the movement of the iodine into the follicle and the manufacture of enzymes important in the iodination of tyrosine in the thyroidal protein and in the digestion of the thyroglobulin to free the hormone (see also the discussion in HORMONE).

Negative feedback

The thyroxine circulating in the blood has actions in many tissues; they are outlined in succeeding sections. An important self-regulatory action, however, is dealt with at this point, that of negative feedback—an important control mechanism in endocrine systems generally—on the hypothalamus and the pituitary (see Figures 3 and 4).

Figure 4 illustrates the negative feedback system of thyroxine and indicates those elements in the system that are stimulatory and those that are inhibitory. It can be seen that introducing thyroxine into the system augments the inhibition on the hypothalamus and pituitary by thyroid hormone already present. The rate of formation of TRH and TSH thus would decrease. This is the way by which extraneous thyroxine reduces the activity of the thyroid gland. Removal of the thyroid, on the other hand, leaves the hypothalamus-pituitary TSH system completely unopposed, causing the pituitary uselessly to secrete TSH. This also happens when the diet lacks iodine, and the thyroid makes insufficient thyroxine. The pituitary, unopposed by negative feedback of thyroxine, again secretes large amounts of TSH, which stimulates excessive enlargement of the thyroid and, although it is capable of responding, still lacks the iodine necessary to make the hormone. Such feedback (or more precisely, lack of feedback) results in enlargements of the thyroid, known as goitre, in man and other animals in geographical areas where dietary iodine is very low.

Ultimobranchial gland. The ultimobranchial gland is so-named because during embryonic development it separates from the most posterior of the branchial (pharyngeal) pouches. The hormone produced by this gland is calcitonin, which reduces the amount of calcium in the blood.

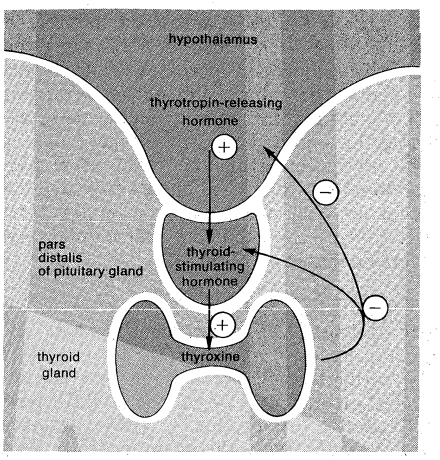


Figure 4: Hypothalamus–pituitary–thyroid self-regulatory system.

Features of ultimobranchial tissue

In most vertebrates, the embryonic ultimobranchial tissue migrates from its pharyngeal points of origin to the level of the heart. In fishes and amphibians, the tissue is sometimes in contact with the membranous sac (pericardium) surrounding the heart. In many vertebrates other than mammals, the ultimobranchial glands in the adult are of unequal size, and in many species only one gland persists. In mammals, the ultimobranchial tissue fuses during embryonic development with the thyroid and then differentiates as thyroid-like follicles, or as scattered interfollicular C cells within the adult thyroid gland. In mammals, the ultimobranchial thyroid follicles may resemble thyroid follicles, but they do not metabolize iodine.

Whether the ultimobranchial tissue is separate or intrathyroid, it appears to be responsive to blood levels of calcium: that is, high levels of blood calcium appear to activate ultimobranchial tissue and cause it to secrete calcitonin, a relatively small peptide hormone; low levels of blood calcium appear to inhibit calcitonin production.

Parathyroid glands. Like the ultimobranchial glands, the parathyroid glands are derivatives of the pharyngeal pouches of the vertebrate embryo. Because of their multiple origins from opposite sides of the pharynx, they may be quite widely separated in the neck region of the adult. They vary in number from one to two pairs; rarely are there three pairs. Only in the mammals are they commonly fused with the thyroid gland. This fusion is not as intimate as in the case of the mammalian ultimobranchial gland, however; in man and the rat, for example, the parathyroids are small round bodies just under the fibrous capsule that surrounds the thyroid gland.

The two kinds of mammalian parathyroid cells

Microscopically, the parathyroids are organized in the cell cord–sinusoid pattern (see Figure 2). In mammals, two kinds of cells are distinguishable, the chief cells and the oxyphil cells. The more numerous chief cells appear to manufacture the parathyroid hormone parathormone, a peptide. The function of the less numerous oxyphil cells in the mammalian parathyroids has not yet been established.

The microscopic structure of parathyroid cells under different conditions of blood calcium content suggest that parathyroid cells respond to differences in blood calcium level. They are activated to secrete parathormone when the blood calcium content is low and are quiescent when the content is high. The ultimobranchials and parathyroids together form a system sensitive to altered blood calcium levels, and they reciprocally regulate these levels.

Gastrointestinal endocrine system. Thyroxine-producing, calcitonin-producing, and parathormone-producing glands are derived from the developing wall of the upper digestive tract. A larger group of hormone-secreting glands is derived from the embryonic midgut, or abdom-

inal level of the digestive tract: gastrin is secreted by the stomach wall; insulin and glucagon are secreted by the pancreas, which in the embryo becomes separated from the intestine; secretin, cholecystokinin-pancreozymin (CCK-PZ), and possibly others are secreted by tissues that remain in the intestinal wall. Studies have indicated that cholecystokinin and pancreozymin are merely different phases of action of the same hormone. For convenience the two names, usually combined and hyphenated, are abbreviated CCK-PZ.

Several of the hormones of the midgut, all of which are relatively small peptides, resemble each other chemically. There are, for example, numerous duplications of sequences of the amino acids constituting the hormones. Secretin and glucagon especially resemble each other —14 of the 27 amino acids of secretin are found as duplicates in the glucagon molecule. Gastrin is of particular interest since four amino acids at one end of the molecule, which consists of 17 amino acids in chainlike arrangement, have all of the actions, though weaker, of the entire molecule. The augmenting action of the first 13 amino acids in the gastrin molecule can be simulated by attaching through chemical methods various other groups to the last four amino acids; for example, such a compound named pentagastrin has generally higher physiological activity than gastrin itself.

The amino acid sequence at one end of the intestinal hormone cholecystokinin-pancreozymin molecule (CCK-PZ) is the same as that of gastrin. Although CCK-PZ contains 33 amino acids, its chemical resemblance to gastrin not unexpectedly means that some of the actions of gastrin and CCK-PZ overlap; *e.g.*, CCK-PZ has 4 percent of the activity of gastrin on frog stomachs. The differences in action between CCK-PZ and gastrin apparently reside in the part of the molecule preceding the five identical terminal amino acids. Interestingly, the skins of certain hylid frogs contain a peptide called caerulein, consisting of ten amino acids; the last five are identical with those at one end of gastrin and CCK-PZ. In pharmacological tests, caerulein has some of the effects ascribable to gastrin and CCK-PZ.

Hormonal effects during digestion

When food enters the stomach, the distension of the organ stimulates nerves leading to gastrin-secreting cells scattered through the mucosa (inner lining layer). The gastrin then released by the stomach into the blood affects the stomach itself, principally stimulating the enzyme-forming cells to secrete pepsin, so-called parietal cells to secrete acid, and the stomach musculature to contract. Even before food enters the stomach, however, some secretion of gastrin and acid may be started by nervous stimuli originating in the brain. This is known as the cephalic phase of gastric stimulation. The increased quantity of acid in the stomach has a direct negative feedback action on the acid-secreting and gastrin-secreting cells, retarding their function after conditions favourable for gastric digestion have been achieved.

The effects of gastrin on liver, pancreatic, and intestinal function, though relatively weak, have meanwhile begun. When the acidified food enters the intestine from the stomach, two more hormones, secretin and CCK-PZ, are released into the blood by intestinal mucosal cells and activate the flow of pancreatic secretion. In addition, these two hormones affect the function of the liver, the gall bladder, and the intestine itself. Secretin also has an inhibiting effect on gastric acid secretion and on gastric muscular movements. CCK-PZ has a slight stimulatory action on gastric functions but reduces gastric acidity by competing with gastrin for sites of action on responsive cells in the stomach. Displacement of gastrin by CCK-PZ, a weaker stimulant, moderates the actions of gastrin on the stomach. The inhibiting effect of secretin and CCK-PZ on gastric acid secretion was once thought to be the action of an intestinal hormone called enterogastrone. The existence of enterogastrone as a separate chemical entity, however, is no longer considered valid.

The three principal hormones of the gastrointestinal wall, and possibly glucagon from the pancreas, have interlocking and interrelated actions, forming a system for endocrine regulation of the digestive process. It is thought that this interrelatedness derives from the evolutionary divergence of both the hormones and the glandular and muscular receptors. Comparison of gastrin molecules from six mammalian species shows that, of the total number (17) of amino acids, these vary by only one or two, which are never at the "active" end of the molecule. The variation is caused by single changes (mutations) in the genetic code. When similar comparisons are available for gastrins and other gastrointestinal hormones of lower vertebrates, an interesting evolutionary picture should appear.

The gastrointestinal endocrines were the first of the endocrine systems to be demonstrated experimentally, and although the chemical structure of the hormones has been established, the identity of the cells secreting these hormones has been elusive. Gastrin does not originate in a separate organ, but in isolated endocrine cells scattered through the mucosal layer of the gastric wall. It is thought that the cellular sources of the intestinal hormones may be a similar type of cell of the intestinal mucosa. The function of these cells is not yet fully understood, but they resemble, in relative position and structure, the gastrin-secreting cells.

Pancreatic endocrine tissue. In most vertebrates, the tissue that eventually becomes the insulin-glucagon secretory system originates in the embryo with the pancreas as single or multiple outpocketings of the upper part of the intestine. As the pancreas differentiates into secretory units, part of it remains connected with the pancreatic ducts and, in the adult, functions directly in the digestive process. A part of the developing pancreas, however, differentiates into islands of tissue isolated from the ducts; these islands are responsible for the endocrine function of the pancreas. In vertebrates with the pancreatic development described above, the endocrine cell units are called islets of Langerhans. In some vertebrates—bony fishes for example—the endocrine pancreatic tissue may become partly separated from the exocrine pancreas. In cartilaginous fishes, the endocrine tissue, which retains an anatomical relationship to the ducts, presumably does not secrete into them. In the most primitive vertebrates, the cyclostomes, the endocrine pancreatic tissue is in the intestinal wall near the point of attachment of the bile duct.

Role of alpha and beta cells in mammals

In all vertebrates, the endocrine pancreas is organized as cell cords separated by sinusoids. Two principal cell types are distinguishable, and as many as three more minor cell types may also occur. In mammals, the most numerous cell type, recognizable by its position and by details of its intracellular structure, is the beta cell, which has been identified by a variety of experimental procedures as the source of insulin. The alpha cell, which is usually found in greatest numbers near the edges of mammalian islets, is considered to be the source of glucagon. Functions of the other cell types are not clear, but one may produce gastrin; the human disease known as the Zollinger-Ellison syndrome, for example, is characterized by pancreatic islet cell tumours and high levels of gastrin in the blood.

The proportions of islet alpha and beta cells relative to each other vary widely and characteristically among vertebrate species. In birds and reptiles, alpha cells are more numerous; in amphibians, beta cells predominate.

If blood or saline fluids high in glucose are experimentally introduced into the pancreatic artery, insulin is immediately released. Similar introduction of fluids low in glucose stops insulin release and activates the release of glucagon. Such experimental results indicate that the alpha and beta cells are directly sensitive to differences in glucose concentration in the blood. As in the case of blood calcium regulation (see below *Parathyroid–ultimobranchial system*), the alpha and beta cells form part of a regulatory system for maintaining a relatively constant blood glucose level. This role is further described in the article HORMONE, as is the molecular variation among vertebrates of the insulin molecule.

Adrenal medulla. In the higher vertebrates—reptiles, birds, and mammals—the adrenal glands are generally rounded or angular organs composed of two basic kinds of tissue: one secreting steroid hormones, the other se-

creting adrenaline (epinephrine) and noradrenaline (norepinephrine). In mammals, the adrenaline and noradrenaline-secreting cells are in the core, or medulla, of the organ, and the steroid-secreting tissue forms the outer shell, or cortex (Figure 5). The cells of the medulla are sometimes referred to as chromaffin cells because they can be stained by chromate-containing tissue preservatives.

In lower vertebrates, chromaffin cells are found not only in association with those tissues equivalent to the adrenal cortex, called the interrenal tissue, but also on the kidney, where they are even more widely distributed, and along the roof of the body cavity, where they are associated with nerve ganglia of the sympathetic nervous system. The distribution of the chromaffin-adrenal medullary cells partly reflects their embryonic origin and their association with sympathetic ganglia—they originate from the neural crests of the embryonic neural tube, which also give rise to the sympathetic ganglia. The higher vertebrates show a more intimate association between adrenal cortical and medullary tissues than do the lower vertebrates. No distinct functional value has yet been recognized for this anatomical association (see Figure 5).

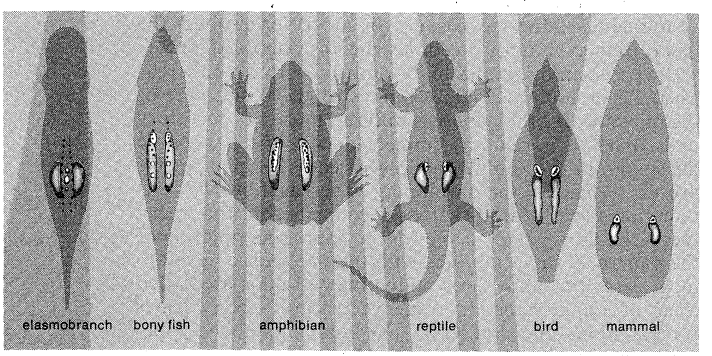


Figure 5: *Patterns of vertebrate adrenal glands.* White, interrenal or cortical tissue; black, chromaffin or medullary tissue; shaded area, kidney.

Regardless of the type of organization, the chromaffin cells secrete into the blood noradrenaline and the closely related adrenaline. Among the vertebrates, there is a wide variation in the proportions of these two hormones in the blood and in the adrenal medullary tissue. In the adrenal medulla of mammals, the cells are arranged in a branching cell cord-sinusoid pattern. Chromaffin cells, whether in a glandular mass, in a single group, or in small groups, are innervated by nerve cells from the central nervous system. They are thus endocrine glandular counterparts for the nerve cells that release noradrenaline as a chemical transmitter of impulses between cells.

Range of effects of adrenaline and noradrenaline

Adrenaline and noradrenaline have a wide variety of target organs, including the heart and blood vessels and most of the organs with smooth muscle in their walls (*e.g.*, the hollow viscera, the uterus); in addition, these hormones affect the metabolism of glucose and the dispersion of pigment in skin pigment cells. Because of their chemical similarity, the physiological actions of noradrenaline and adrenaline also are similar; however, their effects may vary considerably in degree, and in a few instances oppose each other. Furthermore, the precise actions of adrenaline and noradrenaline on a given organ or physiological process may differ in different animal groups.

One of the many as yet unexplained phenomena involving the function of adrenaline and noradrenaline is the extremely large concentrations of these hormones found in the heart muscle of hagfishes. The hagfish heart, however, is not innervated, nor does it appear to respond to additional quantities of adrenaline or noradrenaline.

Adrenal cortex. Adrenal cortical tissue and the gonads are the only steroid hormone-secreting endocrines in vertebrates. The embryonic origin of these glands is also similar—that is, they develop, as do the kidneys, from condensations of loosely organized cells called mesenchyme cells in the body wall above the abdominal cavity. In elasmobranch fishes, the condensed masses of mesenchyme cells commonly differentiate as median glands between the two kidneys (see Figure 5) and are known as interrenals. In bony fishes, the interrenal tissue is associated with the kidneys much further forward and may emerge from within the kidney, attached to the walls of veins carrying blood toward the heart. In addition, bony fishes have two or more so-called Stannius corpuscles attached to or embedded in the kidney. These develop from the mesonephric region of the embryonic kidney. Studies indicate that the Stannius corpuscles may secrete adrenal steroid hormones. In amphibians, interrenal or adrenocorticoid tissue, arranged either as a single mass or in segmented units, is embedded in or on the kidney and mixed with chromaffin cells. In reptiles, birds, and mammals, the adrenal mesenchymal masses generally form the pair of discrete glandular organs typically associated with chromaffin tissue at the upper ends of the kidneys.

The embryonic origins of these tissues explain the general anatomical association of the adrenocortical tissue, the kidney, and the gonadal tissue, even though details of the association are highly variable; for example, accessory adrenocortical cells commonly are found among the testicular ducts, which develop in the embryo from the early kidney.

The usual arrangement of adrenocortical tissue is as cords or columns. The adrenals of lower vertebrates appear to have fewer and smaller blood vessels among the cords. The mammalian adrenal cortex is regularly structured in a concentric pattern, the cell cords in the outermost zone (zona glomerulosa) being arranged in whorls, with the next zone (zona fasciculata) arranged in long parallel columns. There is some evidence that the zones secrete different steroids.

A small amount of steroid hormone is stored in the steroid-secreting cells as lipid (fat) droplets. Corticoadrenal cells, therefore, stain well with fat stains, unless adrenocortical hormone (ACTH) from pituitary is injected; ACTH stimulates release of the hormone stored in the cells.

Cortical adrenal gland cells and interrenal gland cells produce a variety of hormones related in structure; a relatively small number of them predominate, namely cortisol, corticosterone, and aldosterone. In certain fishes, a related steroid, 1-alpha hydroxycorticosterone, is produced in relatively high concentration. Corticosteroids belong to one of two general classes, depending on their principal or strongest action: mineralocorticoid and glucocorticoid. Mineralocorticoids (*e.g.*, aldosterone) affect the movement of water and salts through organs of exchange with the environment such as the kidneys (skin and gills in aquatic animals). Glucocorticoids (*e.g.*, cortisol, corticosterone) chiefly affect metabolic formation of carbohydrates, such as glucose from noncarbohydrate substances (gluconeogenesis); and the general activity of connective tissue cells, called fibroblasts; and of cells involved in the immunological response. Glucocorticoids may have some mineralocorticoid action and vice versa.

In mammals, the synthesis of glucocorticoids in the zona fasciculata of the adrenal cortex is regulated by ACTH. ACTH release is in turn controlled by the hypothalamus; the glucocorticoids thus are steroid hormones that participate chiefly in neuroendocrine reflexes and that serve as the chief negative feedback controls on ACTH secretion.

Aldosterone is formed mainly in the zona glomerulosa of the mammalian adrenal cortex. Its formation is controlled to a minor degree by ACTH, but it is influenced by the renin-angiotensin system, a combination of a kidney-produced enzyme and a peptide hormone that affects blood pressure, sodium concentration, and water movements, (see HORMONE for a more detailed description).

The stress response

Experiences harmful to an organism, such as pain, poisoning, general trauma, starvation, and disease, can excite nerves that lead eventually to a common terminus in the hypothalamus, where adrenocorticotropin-releasing hormone (CRH) is formed. Stimulation of CRH enhances ACTH stimulation of high levels of glucocorticoid secretion; the physiological consequences—the utilization of

proteins to form carbohydrate and an increase in the amount of sugar and protein in the blood—are prominent features of a phenomenon known as the stress response.

The stress response of interrenal and adrenal cortical tissues is an important adaptive system. The conversion of noncarbohydrates such as protein and fat to carbohydrates results in a source of energy for animals that are starving or in some other adverse condition. Destruction of lymphoid tissue and lymphocytes as a result of high levels of glucocorticoid secretion releases proteins, including antibodies, into the blood. The increased amounts of blood protein temporarily overcome large losses in blood volume resulting from hemorrhage. The increased antibody level helps the organism to resist infections. Because of a stress response, the Pacific salmon arrives at the end of his spawning migration with a tremendously enlarged interrenal, as well as high blood levels of glucocorticoids; both changes are adaptive, since, in moving upstream, the salmon does not stop to feed. The intensive gluconeogenesis during this time results in the depletion of tissue proteins; the glucose formed from the protein provides the energy for the muscular work involved in upstream migration.

Ovary and testis. The gonads (the ovary and the testis) are double glands in the sense that their external secretion, the sex cells (ova and spermatozoa), are released into ducts, and their internal secretions, the hormones, are released into the blood. Early in development a pair of urogenital structures develops from the body wall, above the body cavity near the adrenal cortex, and splits lengthwise into an embryonic urinary system and a genital ridge. The genital ridge contains elements capable of differentiating into either a testis or an ovary, depending on the genetic sex of the individual. In a few vertebrate species, the differentiated gonad contains, either simultaneously or successively, both male and female structures and hormones.

The ducts of the mesonephric kidney, which remains functional in lower vertebrates but usually degenerates in reptiles, birds, and mammals, are used to convey spermatozoa toward the external male organs. The female ducts and passages do not have such a close relationship to the urinary system except at the opening leading outside the body.

The endocrine elements in both the ovary and the testis, the interstitial tissues, are so-called because they are found between and among the germinal tissues, which are the seminiferous tubules in the testis and the ovarian follicles in the ovary. Cellular layers, called the theca, in the wall of the ovarian follicle also are endocrine. In many vertebrates, a structure, the corpus luteum, develops from the follicular wall after the ovum is released. The corpus luteum, composed of cords of large cells separated by sinusoids, is an endocrine gland and secretes the steroid hormone progesterone.

The typical hormone of the testicular interstitial cells is the steroid testosterone; in certain species and under particular circumstances, other androgens (male hormones) may be secreted. The typical female sex hormone, or estrogen, presumably secreted by the ovarian interstitial and thecal tissues, is estradiol.

Reproduction, one of the most highly adaptive functions possible in an organism, is geared to the presence or absence of a sexual partner, to temperature, to season of the year, and to various aspects of the life cycle of a particular animal species. Accordingly, the cycles of sex cell production and of sex hormone secretion are controlled by mechanisms involving the pituitary gland, environmental sensors, and neuroendocrine mechanisms (Figures 1 and 3). The functions of these interlocking endocrine controls, their feedback relations, and their cyclic activity are described in greater detail in HORMONE.

Role of the hypothalamus in reproductive regulation

The hypothalamus integrates the complex factors necessary for a successful reproductive system. It receives important neurosensory information—*e.g.*, season of the year, temperature, availability of a mate—is subject to feedback from the sex hormones, and exerts control over both the release of gonadotropic hormones from the anterior pituitary (adenohypophysis) and sexual behaviour. The integration of these many factors is complex and only imperfectly understood thus far. A beginning has been made, however, in understanding the difference in hypothalamic activities in male and female rats. The pituitaries of adult female rats, and many other female mammals, alternately release follicle-stimulating hormone (FSH) and luteinizing hormone (LH) to maintain cyclic ovarian function. The pituitary is stimulated by the hypothalamus to release the hormones. The noncyclic secretion of FSH and LH in the male is balanced by feedback action of testosterone from the testis. The explanation for the differentiation of variable sex hormone feedback function in the female hypothalamus and a static feedback equilibrium in males apparently lies in the sexual differentiation that occurs in the embryo and newborn. Sexual differentiation can be altered experimentally; for example, giving male hormone to a late embryo or a newborn female rat causes its hypothalamus to differentiate in the male (noncyclic) pattern. Adult female rats that have been altered this way are sterile; although they have normal ovaries, they do not have cyclic ovarian function because the hypothalamus has been "programmed" to stimulate a noncyclic supply of gonadotropins from the pituitary. Such females also have abnormal malelike sex behaviour.

INVERTEBRATE ENDOCRINE SYSTEMS

Excepting the arthropods, it can be said that invertebrate endocrine systems are virtually unknown. Much of the lack of information results from the difficulty of applying to small organisms the usual procedures and criteria for proving that a given mechanism is endocrine. The procedures involve: (1) surgical removal of the suspected gland and observation of consequences; (2) elimination of the consequences of removing the gland by reimplantation of glands or by injection of gland extracts; and (3) extraction, isolation, and chemical characterization of the responsible hormone from glands.

Methods used to determine endocrine function

In most invertebrate groups, even though it has not been possible to establish the existence of endocrine mechanisms by these methods, some evidence has been obtained for neurosecretion. Nerve cells with a presumed neurosecretory function and containing granules of the proper size and structure have been found in coelenterates, flatworms, roundworms, nemertean worms, annelid worms, mollusks, arthropods, and protochordates. In the majority of invertebrates, however, there is little evidence to indicate the nature of the function of these cells. In annelid worms it has been shown that regeneration of certain body parts and maturation of sex cells in the gonads may not occur when the brain is injured or removed; this has been taken as evidence for an endocrine relation of the brain to these processes. Even if such a relationship does exist, however, it remains to be proved that the nerve cells are involved.

In echinoderms such as starfishes, an egg-shedding hormone can be extracted from the radial nerves in the rays, or arms. The precise source of this substance, which may be considered more a pheromone—a chemical substance by which individuals of a species communicate—than a hormone, is not yet known. In a variety of mollusks, both the anatomy and the seasonal variations in staining of large neurosecretory cells have been established. A direct relation to internal functions, however, is not yet known, although the electrical activity of some of the cells has been studied. In the octopus, a discrete nonneurosecretory, but innervated structure, the optic gland, is believed to have a gonadotropic function; *i.e.*, it influences the gonads.

Endocrine and neuroendocrine functions in the arthropods have been better established than in any other invertebrate groups. Not only are anatomical features well defined but functions are also fairly well understood, and the chemical nature of some of the arthropod hormones is known.

Insects. Figure 6 summarizes the general structure of the central nervous system and known glands of an insect. Although individual differences exist in the disposition of the parts in different insects, all parts are usually

represented in all insects, at least in the immature (larval) forms.

Neurosecretory neurons

The brain, or cerebral ganglion, of insects contains groups of the cell bodies of neurosecretory neurons. Each group, referred to as a pars intercerebralis, sends secretion-laden axons toward organs called the corpus cardiacum and the corpus allatum on each side (see Figure 6). Nerves connect the brain and the corpora. In addition, discrete neurosecretory neurons occur in the chain of nervous ganglia behind the brain. Several kinds of neurosecretory cells occur in the brain and ganglia, possibly with different functions, since they stain differently. The two corpora cardiaca presumably release neurosecretory factors important in various insect functions. One of these neurosecretory hormones, known as prothoracotropic or ecdysiotropic hormone, stimulates the prothoracic gland, or ecdysial gland, in the larva to secrete an important hormone—ecdysone, or molting hormone. The prothoracic gland degenerates in the adult insect.

Some of the neurons between the corpora cardiaca and the brain traverse the latter and terminate in a glandular structure, the corpus allatum, on each side. The corpora allata secrete juvenile hormone, which, in the larva, has a role in development and the periodic shedding of skin (molting) and, in the adult, appears to be gonadotropic.

Analogies and similarities exist between the brain–thoracotropic–prothoracic gland system of insects and the brain–pituitary–pituitary target organ systems of vertebrates. The same can be said of the brain-corpus cardiacum-juvenile hormone-target organ system. As is explained below, similar brain–neurosecretion–primary target gland-secondary target organ systems occur in crustaceans. The adaptive value of these systems is the same in all animals in which they have been demonstrated. The systems provide a way for sense organ-detected changes in the environment to influence, or to trigger, an internal functional adjustment. It is dangerous, for example, for some animals to molt under conditions of dryness. In such cases low humidity or high humidity, as perceived through appropriate sense organs, inhibit or favour release of molting hormone through a nervous input at the neurosecretion level in the brain. In a similar manner, it is most favourable and efficient if certain gonadotropic functions occur only when a sexual partner ready for mating is present. Internal reproductive changes (*e.g.*, ripening of gametes, ovulation) often are triggered either by pheromonal signals from the opposite sex or by mechanical or tactile signals such as touching particular external structures—*e.g.*, vagina of the rabbit, genital structures in the roach. Sensory signals from a possible mate are integrated in the central nervous system and eventually result in release of a gonadotropin from the appropriate gland (pituitary or corpus allatum).

From (bottom) A. Gorbman and H.A. Bern, *Textbook of Comparative Endocrinology* (© 1963); by John Wiley & Sons, Inc.

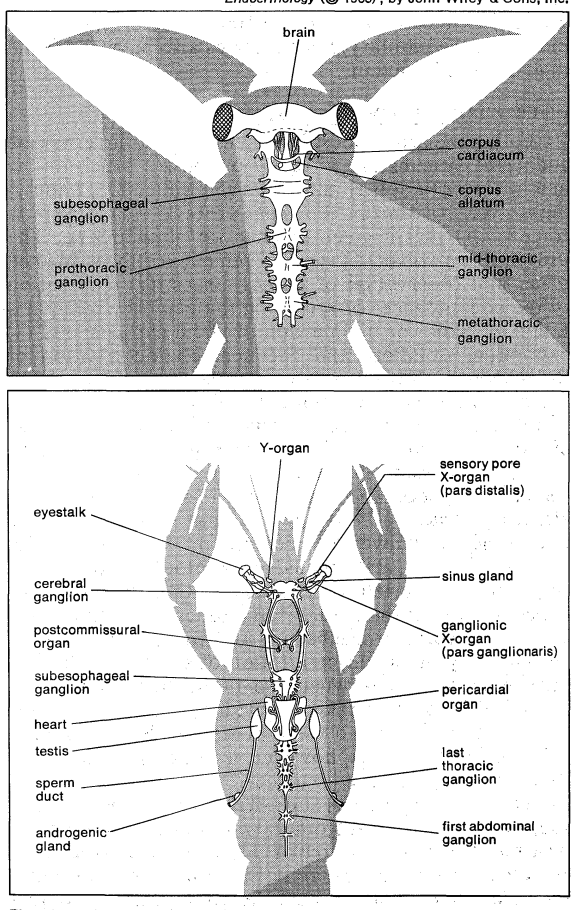


Figure 6: ***Invertebrate endocrine systems.***
(Top) Central nervous and endocrine structures of a silkworm moth; (bottom) endocrine system of a generalized male crustacean.

Crustaceans. In crustaceans, as in insects, the primary neurosecretory structure is the brain, or cerebral ganglion; other groups of neurosecretory cells are located in nervous ganglia apart from the brain (Figure 6). In addition, there is a separate group of neurosecretory cells known as the X-organ in the eyestalk. The neurosecretory cells terminate in several neurohemal organs in order to discharge their hormones into the blood; perhaps the best known neurohemal organ is the sinus gland in the eyestalk. The postcommissural organ serves a neurohemal function for nerve cells in nearby thoracic ganglia. Crustaceans have two known nonneurosecretory glands, the Y-organ and the androgenic glands (Figure 6); both types are under the control of blood-borne neurosecretory substances from cells in the central nervous system.

It has been established that the neurosecretory cells of the brain and the eyestalk are involved in the control of molting through inhibition of the Y-organ, in gonadal function through inhibition of the androgenic gland, in the metabolism of carbohydrate and probably other substances, in the dispersion of body pigment, and in the rhythmic movement of pigment in the retina of the eye.

There is strong evidence that the neurohormone released from the pericardial organs accelerates the action of the heart. A simultaneous loss, or discharge, of neurosecretory granules from the neurohemal organ and an increase in heart rate can be brought about by electrically stimulating the neurons leading to the pericardial organs.

The crustacean molting hormone, called crustecdysone, secreted by the Y-organ, is a steroid very similar chemically to ecdysone, the molting hormone of insects. Molting in insects is a phenomenon only of larvae, but it occurs throughout the life of crustaceans; both groups, however, have similar endocrine mechanisms.

The structure of the presumed hormone of the crustacean androgenic gland has not yet been fully characterized. But the androgenic gland and the corpus allatum of insects appear to have analogous gonadotropic functions. There are, however, greater dissimilarities in gonadotropic functions between crustaceans and insects than in molt-control systems between the two groups.

COMPARATIVE, ADAPTIVE, AND EVOLUTIONARY ASPECTS OF ENDOCRINE SYSTEMS

It is valuable to compare endocrine mechanisms among animal groups, to recognize consistent evolutionary trends, and to define the adaptive values or purposes of these trends.

The neurosecretory cell. Although it has been commonly said that neurosecretory cells illustrate the most widespread and common endocrine phenomenon, accumulating knowledge about neurosecretory cells has made clear the fact that comparisons among neuron-derived secretory cells cannot be made more readily than comparisons among secretory cells derived from other sources. Secretory nerve cells may exemplify the most plastic of all endocrine systems; that is, they are readily modified, or adapted, to serve a variety of particular functions. It may not be any more valid to compare two different kinds of neurosecretory cells—*e.g.*, in a fish and in an insect, in the hypothalamus and in the spinal cord,

in the cerebral ganglion and in an abdominal ganglion—than it is to compare two glandular endocrine secretory cells, one in the thyroid and one in the pancreas.

Properties of neurosecretory cells

It is useful nevertheless to summarize the properties of neurosecretory cells, in order to show the many ways by which they may differentiate and thus adapt to particular functions.

The neurosecretory cell (see Figure 7) has a neuron-like form; that is, the cell is composed of a cell body, projections called dendrites (D, in Figure 7), a long axon (A, in Figure 7), and synaptic contacts with other neurons at least at one end. Some neurosecretory cells do not have all of these structural features; others have additional ones. Evidence indicates that some hypothalamic neurosecretory cells may have special fibres, called recurrent collateral fibres (labelled possible RC in Figure 7), extending from a point on the axon near the end of the cell (neuron terminal) to the cell body (perikaryon, PK). Recurrent collateral fibres allow a variety of feedback relationships between the neuron terminal and the cell body. Some neurosecretory cells vary widely from Figure 7; for example, the axon in some cells—*e.g.*, molluscan brain, Speidel cells in the spinal cord of elasmobranch fishes—is extremely short, and the size of other cells may be large enough to require blood from small blood vessels (capillaries).

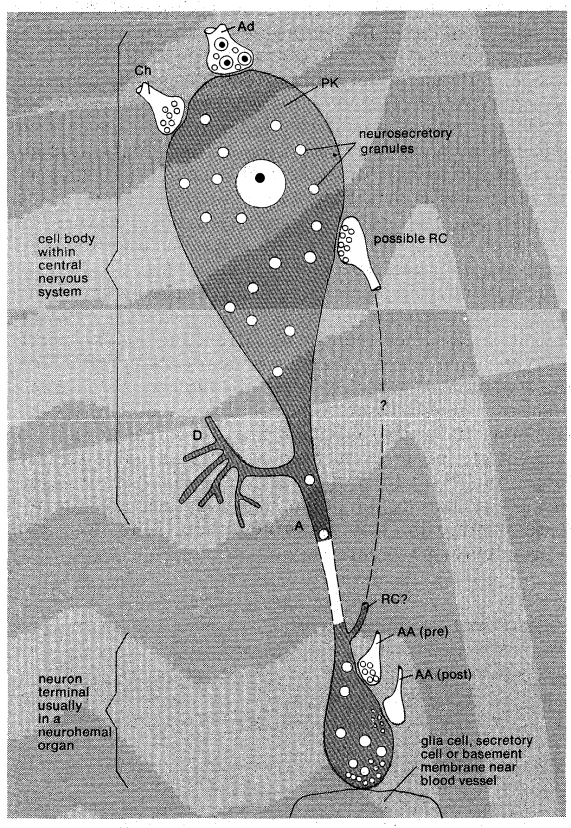


Figure 7: Generalized neurosecretory cell (see text).

One important structural variable of neurosecretory cells, and one that determines the precise character of their function, is the nature of the incoming (afferent) neurons that form contacts, called synapses, across which nerve impulses are transmitted to the cell body. Considerable progress has been made in the identification of afferent neurons for brain neurosecretory cells of the nudibranch mollusk *Aplysia*. Drugs that block certain classes of neuronal function and techniques used to study cellular components reveal that neurosecretory cell bodies differ in the relative numbers upon them of various kinds of nerve fibre endings, such as those that liberate the compound acetyl choline at the synapse (cholinergic, Ch in Figure 7) and those that liberate norepinephrine (adrenergic, Ad in Figure 7).

A second important variable of neurosecretory cell function is the point at which the cell ends. The most common termination is in a neurohemal organ; as has been mentioned, there are numerous types. Generally, neurohemal organs consist of secretory endings of one type of neurosecretory cell and the blood vessels upon which they lie. An exception is the vertebrate median eminence, which contains at least six types of neurosecretory endings; the number of types has not yet been resolved for the crustacean sinus gland. Some invertebrate neurosecretory cells and the elasmobranch Speidel cell may have no apparent neurohemal organs; the endings of these cells are dispersed in connective tissue that contains blood vessels. Some neurosecretory cells end in direct synaptic contact with glandular secretory cells, as for example, in the corpus allatum of insects, in the pars intermedia of most vertebrate pituitary glands, and in the pars distalis of the pituitary of fishes.

Another variable in neurosecretory cells is the type of connection of their neuron terminals. Formerly it was thought that the axons of neurosecretory cells do not synapse with other neurons at their neuron terminals; however, more recently inconspicuous axon endings have been found upon neurosecretory endings. These are either pre-synaptic (AA pre in Figure 7) or post-synaptic (AA post in Figure 7); in the first instance, the impulse is coming from the neurosecretory cell, in the second, it is going to the neurosecretory cell.

Still another important variable in neurosecretory cells is the chemical nature of the neurosecretory secretion. Chemical differences have been shown by staining techniques among almost contiguous neurosecretory cells in the brains of arthropods. Neurosecretions whose chemical nature has been established include several from the vertebrate pituitary gland—thyrotropin releasing hormone (TRH), luteinizing hormone-releasing hormone (LRH), growth hormone release-inhibiting hormone (GIH), oxytocin, and vasopressin—all of which are peptides. Based on relatively crude chemical evidence, some hormones (*e.g.*, ecdysiotropic hormone of insects) are believed to be small proteins. It is probably safe to conclude that no particular uniformity will be found in the chemical nature either of the neurosecretory hormones or of their intracellular protein carriers, called neurophysins.

Finally, it should be pointed out that the overt physiological roles of neurosecretions are extremely broad; *i.e.*, they exert direct effects on general metabolism, control the functions of important endocrine glands such as the pituitary gland and the Y-organ, and regulate the function of particular organ systems—the kidney in excretion, the uterus in reproduction—and adaptive pigmentary changes. The location of the neurosecretory cells in the central nervous system—that is, their contact with sensory systems on one hand and numerous effector systems on the other hand—is highly strategic. It is no wonder, then, that they have been subjected to many kinds of adaptive pressures, and, in response, have evolved in many directions, in the ways outlined above, to subserve a variety of vital functions.

Hypothalamus–pituitary control systems. In most vertebrate groups the structural features of the hypothalamus–pituitary system seem comparable; that is, there is a neurohemal organ, the median eminence, from which blood vessels carry neurosecretory hormones directly to sinusoids in the pars distalis region of the pituitary gland (see Figures 1, 3). This relationship, as has been pointed out, makes possible the regulation by the brain of hormonal secretions of the pituitary gland. Among the vertebrates, however, are several exceptions, undoubtedly of functional significance, to the arrangement of blood vessels between the brain and the pituitary gland. The evolutionary or adaptive functional significance of these anatomical variations has not yet been established.

The most primitive living vertebrates, the Agnatha, apparently have no blood vessel connection between the brain and the pituitary gland. One cyclostome group, the hagfishes, lives deep in the ocean under nearly constant physical conditions. Since no seasonal changes occur, it may be that there is no functional advantage in the regulation of pituitary gland functions to outside conditions

through hypothalamic neurosecretion; hagfishes, for example, breed continuously. Another group of Agnatha, the lampreys, however, breed and migrate seasonally, so that there must be some regulation of functions of the sex organs (among others) through neurosensory systems affected by environmental seasonal cycles. But lampreys have no hypothalamus-pituitary blood system, so that the mechanism by which seasonal changes affect seasonal reproductive behaviour is not known.

Because Agnatha lack a hypothalamus-pituitary blood vessel connection, even though such a system would be advantageous to the lampreys, it is possible that this relationship first evolved in the vertebrate line between the Agnatha and other vertebrates.

The regulation of the pars distalis region of the pituitary gland of cartilaginous fishes also is of interest. One part of the pars distalis, called the ventral lobe, secretes several hormones, including gonadotropins, but has no blood vessel connection to the median eminence; yet gonadotropic hormone secretion in elasmobranchs is regulated according to season or other external environmental events.

In birds and in elasmobranchs, the median eminence differentiates into at least two regions, a specialization that may permit the secretion of specific releasing hormones into specific portal blood vessels that go to a particular part of the pars distalis. Whether or not such a regional segregation of the hypothalamus-pituitary system exists has not yet been proved.

In bony fishes, the hypothalamus and pars distalis are intimately associated. In fact, many hypothalamic neurosecretory neurons cross the boundary between hypothalamus and pars distalis, and terminate on or very near the secretory gland cells. It would appear, then, that bony fishes have a system whereby the hypothalamus can directly regulate secretion by the pars distalis. The relative advantages or disadvantages of such direct control remain to be evaluated.

Parathyroid-ultimobranchial system. The study of the evolution of anatomical features of vertebrate endocrine systems, with the exception of the hypothalamus–pituitary system, usually involves subtle changes. The parathyroid–ultimobranchial system, however, involves relatively important evolutionary differences in structure among the various groups. The hormones and their functions—maintenance of a constant blood level of calcium—on the other hand, appear to be more constant, apparently an exception to the generalization that it is not the hormones that vary, but the uses to which they are put.

Variation in blood calcium regulating mechanisms

The parathyroid gland, for example, is the only vertebrate endocrine organ that is not present in all vertebrates. None of the fishes has a parathyroid. In addition, the hypercalcemic (blood-level increasing) hormone of higher vertebrates, parathormone, is relatively inactive in altering blood calcium levels in fishes. It is reasonable to assume that fishes must regulate blood calcium levels. Their diets vary in calcium content depending on the environment, and, unlike most vertebrates, the skeletal reservoir of calcium is relatively slight, and some bone (acellular bone) is not readily available for metabolic purposes. The ultimobranchial body of fishes, on the other hand, is an excellent source of the hypocalcemic (blood-level lowering) hormone, calcitonin. In fact, calcitonin from salmon and dogfishes is so readily available, and so potent in humans, that it is used in medicine.

Amphibians have parathyroids, but their structure is unlike that of higher vertebrates, consisting, in frogs and salamanders, of whorls of cells (see Figure 2). In reptiles, birds, and mammals, parathyroids are composed of the familiar cell cords and sinusoids. In amphibians and reptiles, who have a well-developed calcitonin-secreting ultimobranchial gland; the parathyroids commonly undergo degenerative changes in the winter and regenerative changes in the spring.

The mechanism of blood calcium regulation in mammals involves the interaction of both parathormone and calcitonin. The mechanisms of regulation in fishes, in which only calcitonin is available, or in amphibians and reptiles, which have a seasonally developed parathyroid gland, clearly indicate an evolution of endocrine structures regulating calcium metabolism in vertebrates.

The two glands that regulate the blood level of calcium through hormonal secretion are derived from the lateral pouches of the pharynx. The glands separate and remain separated in the heart and neck regions of most groups of higher vertebrates. In mammals, however, the structures that will form the ultimobranchial and parathyroid glands consistently migrate into and fuse with the thyroid gland, which has a different embryonic origin. Generally, when such a modification of embryonic development is consistently evolved in a given group of organisms, some functional advantage (not yet established in this case) exists for the evolutionary trend.

Thyroid glands and thyroid function. The vertebrate endocrine organs are present in some form, with the exception of the parathyroids, in the lowest vertebrates, but are lacking in the invertebrates. Thus, the evolutionary beginnings and ancestry of vertebrate endocrine organs have been lost in those extinct prevertebrate animals that have disappeared and whose fossil record is either absent or of no help in understanding endocrine evolution.

The thyroid gland and its function seems to be an exception, at least in part. The formation of the relatively simple hormone thyroxine involves the addition of iodine to proteins containing the amino acid tyrosine. Two iodotyrosines may, even in the test tube, couple to form thyroxine. Although iodotyrosines form easily, the coupling of iodotyrosines to form thyroxine is a much slower process. Naturally-occurring iodotyrosines have been found in numerous invertebrates. In fact, the first iodoproteins and iodotyrosines were obtained in the 19th century from corals before they were known to occur in the thyroid. Naturally-occurring nonthyroidal thyroxine has been reported much more rarely, but it has been found in some invertebrates, especially mollusks.

Thus, the thyroid hormone probably existed without any known function long before the gland itself was evolved. It must be assumed that the occurrence of iodotyrosines and thyroxine was a natural chemical accident, but that the evolution of the thyroid gland followed the development of a physiological role for thyroxine. The special functions of the thyroid gland are: (1) to extract and to concentrate iodine from the blood to add iodine rapidly to a protein; (2) to provide enzymes that favour rapid protein iodination, diiodotyrosine coupling, and the release of thyroxine from its protein; and (3) to store and release thyroxine as necessary.

Special functions of the thyroid

From A. Gorbman and H.A. Bern, *Textbook of Comparative Endocrinology* (© 1963); by John Wiley & Sons, Inc.

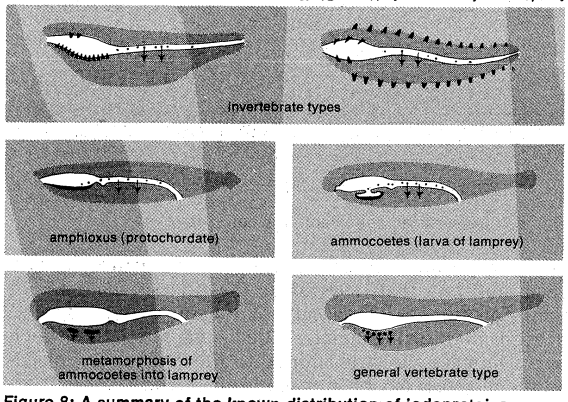


Figure 8: A summary of the known distribution of iodoproteins (shown as solid black) in the animal kingdom, suggesting a pattern of evolution of thyroid function.

Groups of cells in the floor of the pharynx of protochordates (*e.g.*, amphioxus, sea squirts) that concentrate iodine and convert it to thyroxine have been described. This is significant for two reasons: (1) the protochordates are the invertebrates most closely allied with the earliest vertebrates; and (2) the floor of the pharynx is the em-

bryonic source of thyroid tissue. In developing lampreys, an analogous region of the pharynx concentrates iodine; later, this same region develops into thyroid follicles. The fact that, in the larval lamprey, thyroid function—the formation of thyroxine—begins before the thyroid gland has developed from the pharynx may provide a link between the condition in protochordates and the fully evolved condition in the follicular thyroid of all other vertebrates (see Figure 8 for the presumed evolutionary progression of the thyroid).

BIBLIOGRAPHY. E. BAJUSZ (ed.), *Physiology and Pathology of Adaptation Mechanisms* (1969), a well-integrated series of essays and treatises on aspects of the neuroendocrine and endocrine mechanisms involved in organismic adaptation to changes and stresses in the environment; E.J.W. BARRINGTON and C. BARKER JORGENSEN (eds.), *Perspectives in Endocrinology: Hormones in the Lives of Lower Vertebrates* (1968), emphasis on adaptive and evolutionary aspects of endocrinology; U.S. VON EULER and H. HELLER (eds.), *Comparative Endocrinology*, 2 vol. (1963), a comprehensive multi-author account of endocrinological principles, requiring some basic understanding of the subject; P.J. GAILLARD, R.V. TALMAGE, and A.M. BUDY (eds.), *The Parathyroid Glands* (1965), the best available text in this field, but a bit dated; A. GORBMAN and H.A. BERN, *A Textbook of Comparative Endocrinology* (1962), a college-level textbook, now a bit dated because of rapid development of the field, but still the best text source of information concerning comparative morphology of the endocrines; G.W. HARRIS and B.T. DONOVAN, *The Pituitary Gland*, 3 vol. (1966), the most thorough treatise on the subject; K.C. HIGHNAM and L. HILL, *The Comparative Endocrinology of the Invertebrates* (1969), an excellent student-level text that summarizes endocrine regulation in invertebrates; K. LISSAK and E. ENDROCZI, *The Neuroendocrine Control of Adaptation* (1965), an excellent short work on neuroendocrine control; L. MARTINI and W.F. GANONG, (eds.), *Neuroendocrinology*, 2 vol. (1966), a treatment of the field of neuroendocrinology with emphasis on mammals; K.W. MCKERNS (ed.), *Functions of the Adrenal Cortex* (1968), a comprehensive summary of information on corticosteroid hormone formation and control, and the actions of the corticosteroids; S. TAYLOR and G. FOSTER (eds.), *Calcitonin* (1970), a summary, in symposium form, of a rapidly evolving field, written for the specialist, but containing articles understandable by the interested amateur; A. VAN TIENHOVEN, *Reproductive Physiology of Vertebrates* (1968), an excellent brief compilation of facts concerning endocrine control of reproduction in animals; C.D. TURNER and J.T. BAGNARA, *General Endocrinology* (1971), a good introductory college-level textbook, with a biological rather than a clinical orientation; S.C. WERNER (ed.), *The Thyroid* (1968), a compendious treatise on all phases of thyroid structure, function, and clinical pathology; R.H. WILLIAMS (ed.), *Textbook of Endocrinology*, 4th ed. (1968), a thorough and well-written textbook, with a medical and paramedical orientation; R.J. WURTMAN, J. AXELROD, and D.E. KELLY, *The Pineal* (1968), a good source book on the pineal gland.

(Au.G.)

Energy

The concept of energy is most useful in the understanding of the physical world, and, at the same time, the general rule or law developed for describing the role of energy is one of the most abstract ideas in the formulation of that understanding.

Energy is usually and most simply defined as the equivalent of or capacity for doing work. The word itself is derived from the Greek *energeia: en*, "in"; *ergon*, "work." Energy can either be associated with a material body, as in a coiled spring or a moving object, or it can be independent of matter, as light and other electromagnetic radiation traversing a vacuum. The energy in a system may be only partly available for use. The dimensions of energy are those of work, which, in classical mechanics, is defined formally as the product of mass (m) and the square of the ratio of length (l) to time (t): ml^2/t^2. This means that the greater the mass or the distance through which it is moved or the less the time taken to move the mass, the greater will be the work done, or, the greater the energy expended. More popularly, energy is expressed in such work units as foot-pounds, joules, ergs, British thermal units, kilowatt-hours, or tons of TNT-equivalent explosive energy.

GENERAL CONSIDERATIONS

The concept of energy conservation. A fundamental law that has been observed to hold for all natural phenomena requires the conservation of energy; *i.e.*, that the total energy does not change in all the many changes that occur in nature. The conservation of energy is not a description of any process going on in nature, but rather it is a statement that the quantity called energy remains constant regardless of when it is evaluated or what processes go on between successive evaluations.

Application of law of conservation of energy

The law of conservation of energy is applied not only to nature as a whole but to closed or isolated systems within nature as well. Thus, if the boundaries of a system can be defined in such a way that no energy is either added to or removed from the system, then energy must be conserved within that system regardless of the details of the processes going on inside the system boundaries. A corollary of this closed-system statement is that whenever the energy of a system as determined in two successive evaluations is not the same, the difference is a measure of the quantity of energy either added to or removed from the system in the time interval between the two evaluations.

Energy can exist in many forms within a system and may be converted from one form to another within the constraint of the conservation law. These different forms include gravitational, kinetic, heat, elastic, electrical, chemical, radiant, nuclear, and mass energy. It is the universal applicability of the concept of energy, as well as the completeness of the law of its conservation within different forms, that makes it so attractive and useful.

Development of the concept of energy. The term energy was not applied as a measure of the ability to do work until rather late in the development of mechanics. Indeed, the development of classical mechanics may be carried out without recourse to the concept of energy. The idea of energy, however, goes back at least to Galileo in the 17th century. He recognized that, when a weight is lifted with a pulley system, the force applied multiplied by the distance through which that force must be applied (or, by definition, the work) remains constant even though either factor may vary. The concept of *vis viva*, or living force, a quantity varying as the mass multiplied by the square of the velocity, was introduced in the 17th century. In the 19th century, the term energy was applied to the concept of the *vis viva*.

Newton's first law of motion recognizes force as being associated with the acceleration of a mass. It is almost inevitable that the integrated effect of the force acting on the mass would then be of interest. Of course, there are two kinds of integral of the effect of the force acting on the mass that can be defined. One is the integral of the force acting along the line of action of the force, or the spatial integral of the force; the other is the integral of the force over the time of its action on the mass, or the temporal integral.

Evaluation of the spatial integral leads to a quantity that is now taken to represent the change in kinetic energy of the mass resulting from the action of the force and is just one-half the *vis viva*. On the other hand, the temporal integration leads to the evaluation of the change in momentum of the mass resulting from the action of the force. For some time there was debate as to which integration led to the proper measure of force, the German philosopher-scientist Gottfried Wilhelm Leibniz arguing for the spatial integral as the only true measure, while earlier the French philosopher and mathematician René Descartes had defended the temporal integral. Eventually in the 18th century, the physicist Jean d'Alembert of France showed the legitimacy of both approaches to measuring the effect of a force acting on a mass and that the controversy was one of nomenclature only.

To recapitulate, force is associated with the acceleration of a mass; kinetic energy, or energy resulting from motion, is the result of the spatial integration of a force acting on a mass; momentum is the result of the temporal integration of the force acting on a mass; and energy is a measure of the capacity to do work. It might be added that power is defined as the time rate at which energy is transferred (to a mass as a force acts on it, or through

transmission lines from the electrical generator to the consumer).

Conservation of energy was independently recognized by many scientists in the first half of the 19th century. The conservation of energy as kinetic, potential, and elastic energy in a closed system under the assumption of no friction has proved to be a valid and useful tool. Further, upon closer inspection, the friction, which serves as the limitation on classical mechanics, is found to express itself in the generation of heat, whether at the contact surfaces of a block sliding on a plane or in the bulk of a fluid in which a paddle is turning or any of the other expressions of "friction." Heat was recognized as a form of energy by the English physicist James Prescott Joule, who also proved experimentally the fixed equivalence between mechanical and heat energy. As more detailed descriptions of the various processes in nature became necessary, the approach was to seek rational theories or models for the processes that allow a quantitative measure of the energy change in the process and then to include it and its attendant energy balance within the system of interest, subject to the overall need for the conservation of energy. This approach has worked for the chemical energy in the molecules of fuel and oxidizer liberated by their burning in an engine to produce heat energy that subsequently is converted to mechanical energy to run a machine; it has also worked for the conversion of nuclear mass into energy in the nuclear fusion and nuclear fission processes.

The role of friction

ENERGY CONSERVATION AND TRANSFORMATION

Understanding of details of the conservation of energy requires both the quantitative evaluation of energy in its various forms and understanding of the mechanisms by which the transition of energy from one form to another may occur.

In order to be quantitative in the discussion of any physical process, a consistent set of units is necessary. The modern practice is to use the so-called metre–kilogram–second (mks) set. Thus, distance is measured in metres, mass in kilograms, and time in seconds. Fundamental quantities are expressed in terms of these units, and then any derived quantities are assigned units consistent with the mks system. Thus the derived unit for force is obtained from the relation force F equals the product of mass m and acceleration a:

$$F = ma.$$

When the mass expressed in kilograms is multiplied by the acceleration in metres per second^2, the force has the unit newtons. Energy has the units of force times distance; that is, newton-metres. This unit is also called the joule, which is also equal to a watt-second. In the discussions that follow, the energy units will usually be given in joules. The units of other derivable quantities will be given when introduced.

An ideal system. A simple example of a system in which energy is being converted from one form to another is provided in the tossing of a ball with mass m into the air. When the ball is thrown vertically from the ground, its speed and thus its kinetic energy decreases steadily until it comes to rest momentarily at its highest point. It then reverses itself, and its speed and kinetic energy increase steadily as it returns to the ground. The kinetic energy E_k of the ball at the instant it left the ground (point 1) was half the product of the mass and the square of the velocity, or $\frac{1}{2}mv_1^2$, and decreased steadily to zero at the highest point (point 2). As the ball rose in the air, it gained gravitational potential energy E_p. Potential in this sense does not mean that the energy is not real but rather that it is stored in some latent form and can be drawn upon to do work. Gravitational potential energy is energy that is stored in a body by virtue of its position in the gravitational field. Gravitational potential energy of a mass m is observed to be given by the product of the mass, the height h attained relative to some reference height, and the acceleration g of a body resulting from the Earth's gravity pulling on it, or mgh. At the instant the ball left the ground at height h_1 its potential energy E_{p1} is mgh_1. At its highest point, its potential energy E_{p2} is mgh_2. Applying the law of conservation of energy and assuming no friction in the air, these add up to form the following equations:

Potential energy

$$E_{k1} + E_{p1} = E_{k2} + E_{p2}$$

or

$$\tfrac{1}{2}mv_1^2 + mgh_1 = 0 + mgh_2.$$

In this idealized example, the kinetic energy of the ball at ground level is converted into work in raising the ball to h_2 where its gravitational potential energy has been increased by $mg(h_2 - h_1)$. As the ball falls back to the ground level h_1, this gravitational potential energy is converted back into kinetic energy and its total energy at h_1 again is $\frac{1}{2}mv_1^2 + mgh_1$. In this chain of events, the kinetic energy of the ball is unchanged at h_1; thus the work done on the ball by the force of gravity acting on it in this cycle of events is zero. This system is said to be a conservative one.

If there were a change in kinetic energy at the end of the complete cycle, however, the work done by the forces acting on the ball would not be zero. Thus the force of air friction acting on the ball as it is rising and falling does negative work on the ball; *i.e.*, the ball does work on the air, thereby reducing its kinetic energy. In this case, the net work done by the force of gravity through the round trip is not zero. This system is said to be nonconservative.

Perpetual motion. At this point, it is useful to consider the possibility of constructing a perpetual-motion machine. It is necessary first to define the meaning of the term. A perpetual-motion machine is one that can go through a series of energy transfers and, on returning to its original energy condition, has had the effect of doing work on some external system while undergoing these energy transfers. The law of the conservation of energy directly implies that there is no perpetual motion since that law limits the energy in a system to a finite value, and, hence, if the original energy conditions are regained, no energy was available for a perpetual-motion machine to do external work. To the extent that external work was done, the system failed to regain its original energy condition. The difference in the initial and final energy states (*i.e.*, the change in energy content divided by the initial energy) is a measure of the efficiency of the system in going through the sequence of energy transfers that comprise the cycle.

Gravitational potential energy. *The bouncing ball.* When a ball is moved vertically upward, the kinetic energy removed from the ball is converted into gravitational potential energy by virtue of overcoming the force of gravity over the distance that the ball is being lifted. On its downward path the loss in gravitational potential energy results in an increase in the kinetic energy of the ball given by the magnitude of the gravitational force times the distance that the ball had fallen. In this case the potential energy is directly proportional to the vertical displacement of the object—*i.e.*, $E_p = mgh$. The magnitude of the force that must be overcome in lifting the body against gravity is mg, and it is in the direction opposite to the increase in gravitational potential energy. More generally, the mode of coupling between the loss or gain of kinetic energy and the attendant gain or loss of potential energy, whatever its form, is via the force acting on the object in question. In the one-dimensional case, the force F is equal to the derivative of the potential with respect to distance; *i.e.*, the rate of change (dE_p) of the potential with respect to a change in distance (dr), evaluated at a particular point, and the equation can be written as

Deriving the potential energy

$$F = -\frac{dE_p}{dr}.$$

In the general three-dimensional case, the relation between the potential and the force can be written using vector notation. A vector value is indicated by bold-faced symbol, and so the force is written $\mathbf{F}$. The gradient of the potential is written ∇E_p, and thus: $\mathbf{F} = -\nabla E_p$. But ∇E_p is the shorthand way of writing the sum of the compo-

nents; *i.e.*, the partial derivative of E_p with respect to x, $\frac{\partial E_p}{\partial x}$, the change in E_p along the x-coordinate and $\boldsymbol{i}$ the unit vector along the x-coordinate, and the other symbols that are similar, thus:

$$\nabla E_p = \boldsymbol{i}\frac{\partial E_p}{\partial x} + \boldsymbol{j}\frac{\partial E_p}{\partial y} + \boldsymbol{k}\frac{\partial E_p}{\partial z}.$$

The result is a vector quantity. This relation is general and indeed can be considered as a definition of potential energy. It is important to realize that the storage of potential energy is a result of the system and not of its individual components. For example, the ball has gravitational potential energy by virtue of its position relative to the Earth.

Attraction between two masses. Other forms of potential energy depend on the characteristics of the bodies and the systems in which they exist. Thus, as one considers the gravitational attraction between two masses, m and m', wherever large variations in their separation distance, r, can occur, the observed variation in their mutual attractive force is given by the product of the masses and a constant G divided by the distance squared:

$$F = \frac{Gmm'}{r^2}.$$

The potential energy, then, is

$$E_p = -\frac{Gmm'}{r}.$$

The constant of universal gravitation

G is called the constant of universal gravitation. When the force is expressed in newtons, the masses in kilograms, and the distance in metres, then $G = 6.673 \times 10^{-11}$ newton-metre^2 per kilogram^2. The units of E_p, the potential energy, are newton-metres, or joules. Writing the gravitational potential as mgh, as was done in the Earth–ball example, is valid only when there are small changes in height, h, relative to the separation between the two masses. Since the attraction of the Earth on masses external to it acts as if all the Earth's mass were concentrated at its centre, when only moderate changes in height relative to the Earth's radius occur near its surface, the acceleration, g, may be assumed to be almost constant. For example, in moving from the Earth's surface (which is about 4,000 miles [6,000 kilometres] from its centre) to 10 miles (16 kilometres) above the Earth's surface, the acceleration decreases from about 9.80 metres per sec^2 to 9.75 metres per sec^2.

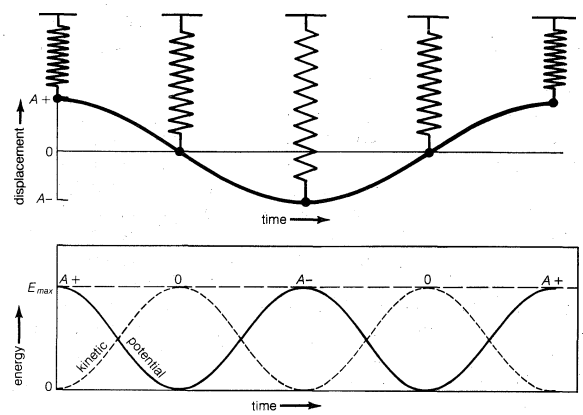
Energy of elastic spring. Energy may be stored in an elastic spring by either extending or compressing the spring. The force F, acting to restore the spring to its equilibrium position in an elastic spring, is proportional to the displacement from the equilibrium position and is opposite in direction to the displacement. When the displacement x is measured in metres, the restoring force F in newtons, and the spring constant k in newtons per metre, then the equation is simply $F = -kx$, and the elastic potential energy E_p, in joules, is given by

$$E_p = \tfrac{1}{2}kx^2.$$

The elastic potential energy is independent of the direction of displacement from equilibrium; *i.e.*, displacement that either stretches or compresses the spring will store energy.

Understanding the distribution of energy in its various forms in a physical problem can often be facilitated by constructing an appropriate graphical energy diagram. If a mass is suspended from a spring in a gravity-free space, displacement of the spring from its equilibrium position stores potential energy in the spring. When the spring is released, the restoring force accelerates the mass, giving it kinetic energy. At the moment it is released, the potential energy stored in the spring is at its maximum value, while the kinetic energy is zero. As the system moves toward the equilibrium position, the stored energy decreases while the force tending to restore the system to equilibrium imparts to the mass a kinetic energy until, at the equilibrium displacement, the potential energy is zero, but the kinetic energy has reached a value that is equal in magnitude to the potential energy at the point of maximum displacement. As the mass moves past the equilibrium position, the spring exerts a retarding force on it, which tends to slow the mass down and thus decrease its kinetic energy. Concurrently, the displacement of the spring from equilibrium again results in energy being stored in the spring. Thus, in an ideal system the total kinetic plus potential energy remains unchanged but is being transformed continually from one to the other. This transfer of energy from one form to another is shown graphically in the Figure. The total energy in the

Transfer of energy in a spring



(Top) Schematic representation of the sequence of positions taken by the moving end of an oscillating spring; at position 0 the potential energy is zero, and the kinetic energy is maximum. At A+ and A— the potential energy is maximum and the kinetic energy is zero. (Bottom) Graph of the energy changes from maximum potential energy to zero, with time.

oscillating system remains constant, but at specific times is completely in the form of kinetic energy, while at other times it is completely in the form of potential energy. In such oscillating systems the characteristic frequency of oscillation can be determined from the basic physical parameters for the spring–mass system. The frequency f is given by the product of $\frac{1}{2}\pi$ times the square root of the ratio of spring constant to mass:

$$f = \tfrac{1}{2}\pi\sqrt{k/m}.$$

Thus, the frequency of oscillation is independent of the energy stored in the system; since k is a constant, any increase in mass will reduce the frequency.

Energy of interaction between charged particles. In considering the interaction between charged particles in electrostatics, the force is given by Coulomb's law, which states that between two charges at rest there is a force that is directly proportional to the product of the charges and inversely proportional to the square of the distance between them. The force acts along the straight line between the two charges. Thus, if q_1 and q_2 are the respective charges measured in coulombs, r is the separation distance between them, F is the force in newtons, and $\frac{1}{4\pi\varepsilon_0}$ is the constant needed to make the two sides of the equation equal, then

Coulomb's law of electrostatics

$$F = \frac{1}{4\pi\varepsilon_0}\frac{q_1q_2}{r^2}.$$

The value of $\frac{1}{4\pi\varepsilon_0}$ is determined experimentally to be 9×10^9 newton-metre^2 per coulomb^2. If the charges are of the same sign, the force tends to push the charges apart; if they are of opposite sign, the force tends to pull them together. In applying Coulomb's law, it is useful to introduce the concept of an electrical field E, the force per unit charge acting on a charge at a point as a result of all the other charges. Thus, the electric field at point 1 due to the charge at point 2 would be, using the equation above:

$$E(1) = \frac{1}{4\pi\varepsilon_0}\frac{q_2}{r^2},$$

so that the force on a charge q_1 located at point 1 is (by

substituting for $\frac{q_2}{r^2}$) the product of charge and field:

$$F(1) = q_1 E(1).$$

The real advantage to the concept of the electric field is in investigations into the effect of a complicated array of charges. Evaluation of the energy in an electric field yields an energy density u_E at a point with a field of magnitude E in a vacuum, in a relationship of the same type as the first equations above ($E_p = \frac{1}{2}kx^2$):

$$u_E = \tfrac{1}{2}\varepsilon_0 E^2.$$

In this equation, if E has the units volts per metre (or joules per coulomb-metre), then ε_0 has the units coulomb^2 per newton-metre^2, and u_E has the units of joules per metre^3.

A particle with charge q moving through a magnetic field B will interact with the field in such a way that a force will be exerted on the particle. The magnitude of the force acting on the charged particle depends on its velocity v relative to the orientation of the magnetic field and on the size of the charge. The force is a maximum when the magnetic field vector $\boldsymbol{B}$ and the velocity vector $\boldsymbol{v}$ are mutually perpendicular and the force acts in a direction perpendicular to the plane defined by the direction of the $\boldsymbol{B}$ and $\boldsymbol{v}$ vectors. More specifically, the magnitude of the force in vector notation can be written as a product of these values:

$$\boldsymbol{F} = q\boldsymbol{v}\boldsymbol{B}.$$

Yukawa potential

The potential term thought to describe the nuclear force between two particles in the nucleus is called the Yukawa potential, after its discoverer, Hideki Yukawa, and has a form that makes the Yukawa potential decrease much more rapidly with distance than the potential does in the relation that holds for gravitational and electrostatic potential by an exponential factor, so that the potential and thus the force falls to zero very rapidly. Experimentally, nuclear forces do not extend beyond about 10^{-13} cm.

Energy of rotational system. Energy can also be stored in a rotational system. Examples of systems of this type include the solar system at one extreme and individual atoms at the other. In either case, an attracting force is balanced by the centrifugal force produced by the rotation of one body about the other. In the solar system, the attracting force is gravitational, while within an atom the attracting force is electrostatic. The kinetic energy of translation (movement) in the orbit and the consequent centrifugal force is necessary to prevent the attractive force from pulling the objects together. For the electron rotating around the nucleus other constraints are necessary; these will be discussed briefly later.

Potential energy can also be stored by the distortion of a stretched thin sheet or membrane from its equilibrium shape. Such a sheet could be the head of a drum or the film of a soap bubble.

Energy in electrical network. It is possible to store energy in an electrical network. For example, a system consisting of a capacitor (a device in which large charges can be accumulated at low voltage) and a pure inductance (see below) can sustain an oscillating electrical current indefinitely, assuming that the circuit's electrical resistance is zero. An inductor is a device for converting an electrical current into a magnetic field with the consequent storage of energy in that field. The energy u_B in joules stored in the magnetic field of an inductance of L henrys (unit of inductance) by passing a current of i amperes through it is given by the type of equation used several times above:

$$u_B = \tfrac{1}{2}L i^2.$$

A capacitor can similarly store energy in its electric field when a charge is placed in it. The energy u_E in joules stored in the electric field of a capacitor of C farads (unit of capacitance) by placing a charge of q coulombs in it is determined by the formula (of the type already used above):

$$u_E = \tfrac{1}{2}\frac{q^2}{C}.$$

Assuming a capacitor containing an initial charge is connected to an inductance, a current will be caused to flow through the inductance as the capacitor discharges. The discharge of the capacitor results in a decrease in the energy stored in its electric field, while the flow of the current through the inductance results in the building of a magnetic field. The total energy at any time is given by the instantaneous value of the energy in the electric field and the magnetic field. Assuming no resistance, the total energy in the system will remain constant while the energy stored in each component will oscillate at a frequency independent of the amount of energy stored.

ENERGY AND STATISTICAL AND QUANTUM MECHANICS

Statistical mechanics. Earlier, it was mentioned that the classical physics of particles interacting with each other could be derived in principle without resorting to the concept of energy. In practice, however, the number of variables necessary to describe the position and velocity of many particle systems presents a computing problem that rapidly gets out of hand. In fact, unique solutions are not possible even for three-body problems. In systems with many individual particles, useful generalizations about the properties of the system can be made by analyzing the energy distribution within the group of particles. This field is called statistical mechanics.

Stellar energies

In a star cluster, the potential energy depends upon the resultant sum of all the gravitational forces that act toward the centre of the cluster. As a star moves about within the cluster, its relative kinetic and potential energies will vary according to its distance from the cluster centre. The detailed path of the star can be extremely complicated, as it may pass in the vicinity of other individual stars and be deflected in its path by the gravitational attraction of the nearby star. After a cluster has existed for a very long time, enough gravitational interactions will have taken place so that the energy of the cluster is divided approximately equally among the individual stars. The energy will never be equal for all stars since the specific results of an individual encounter is to change the energy of the two interacting stars: presumably one leaves the encounter with more energy than the average and the other with less. On the average, however, each star should have its share of the total energy. Under these circumstances the cluster is said to be in a state of statistical equilibrium. It has been shown that under such conditions the total potential energy of the cluster is exactly twice as great as the combined kinetic energy of all the cluster stars. This relation is known as the virial theorem. Thus, by measuring the average speeds of the stars in the cluster, the total kinetic energy and hence the total potential energy can be estimated, and an estimate of the mass of the cluster can be derived. Consideration of the energy balance within the cluster allows an estimate of the total mass of the cluster.

Statistical mechanics considerably beyond the scope of this article is also applied to the problem of a large number of molecules of gas in a system in which the total energy remains fixed. Very briefly, the Maxwell–Boltzmann law of energy distribution states that the particles in the system tend to distribute themselves according to a factor in which the Boltzmann constant, k, corresponds to the energy required to raise the temperature of the molecules 1° K, and T is the absolute temperature. The distribution of energy among the particles (with a mass m) is such that the number of particles n having a speed between v and $(v + dv)$ is $n(v)$ and these values are related in a somewhat more complex equation:

The Maxwell-Boltzmann law of energy distribution

$$n(v) = \sqrt{\frac{m}{2\pi kT}}\, e^{-mv^2/2kT} dv.$$

(The value of Boltzmann's constant is 1.3806×10^{-23} joule per degree Kelvin.)

Concepts of quantum mechanics. Detailed descriptions of the energy associated with atomic phenomena necessitate the use of quantum mechanics. The essence of quantum mechanics lies in the realization that only certain specific values of the momentum of certain particles are possible and that these allowed values differ by h, Planck's constant (6.6262×10^{-34} joule-second). As a re-

sult, an electron rotating around the nucleus of an atom has fixed allowed values for its transitional kinetic energy and hence its total energy.

Transition of an electron from a high energy level to a lower one will make energy available to be radiated as a photon of electromagnetic radiation. The photon will have a characteristic frequency ν given by the ratio of E, the energy available from the transition, to Planck's constant: $\nu = E/h$. Indeed, all electromagnetic radiation consists of photons whose energy and frequency are related in this way. The photon also has a momentum given by the ratio of the energy to c, the speed of light: E/c. The value of the concept of the momentum of a photon is demonstrated in the consideration of the elastic collision between a photon and an electron, a phenomenon called Compton scattering. The problem is readily solved using classical methods of two-body collisions, assuming conservation of energy and momentum and assigning the photon the energy and momentum as indicated.

The use of quantum mechanics is necessary in dealing with atomic phenomena because the absolute amounts of energy involved are so small that the changes associated with going from one allowed quantum level to the next are significant. Energy quanta exist in all physical problems, but they are usually so small that they are indistinguishable from the energy continuum normally assumed. The nucleus–electron problem can be solved under the restrictions imposed by quantum mechanics using position and velocity models that are quite similar to the models used in ordinary mechanics to solve two-body problems. When one considers the large number of particles that can be in the interior of the nucleus itself, however, using statistical methods becomes desirable. Thus, the field of statistical quantum mechanics has become a powerful tool in this area.

The propagation of energy. The transfer of energy can be accomplished either by mechanical interaction between particles or by the radiation of energy in the form of an electromagnetic wave. When a mechanical collision occurs between particles, the computation of the energy transfer is straightforward, assuming the laws of conservation of energy and conservation of momentum.

The energy E associated with an electromagnetic wave has already been identified as being the product of its frequency, ν, and Planck's constant, h: $E = h\nu$. Closer examination of electromagnetic radiation reveals that, as the name suggests, electric and magnetic fields are associated with the propagation of the radiation. The mutually perpendicular fields are coupled in such a way that the energy of the wave oscillates between the two (considered as vector fields) in the plane perpendicular to the direction of propagation of the radiation. The total energy content of the wave is also given by a vector quantity called Poynting's vector. It is a measure of the rate at which the field energy moves through space. The electromagnetic wave also has a momentum density that is given by $1/c^2$ times the Poynting vector. The implications of this flow of energy and momentum are that, among other things, when radiation is absorbed or emitted, a transfer of energy and momentum occurs. Thus the direction of a beam of electromagnetic radiation in one direction with its momentum flow results in the imparting of an oppositely directed momentum to the light source.

Poynting's vector as a measure of wave energy

Equivalence of mass and energy. Expressing the kinetic energy of a moving mass as $\frac{1}{2}mv^2$ becomes invalid as the velocity gets very large and, in that instance, Einstein's theory of relativity must be used to calculate the effect upon energy of this large velocity. The kinetic energy, E_k, is expressed in terms of the product of the mass at low velocity, m_0, and the square of the speed of light, c, multiplied by a ratio relating the square of the velocity, v^2, to c^2:

$$E_k = m_0c^2\,[(1 - v^2/c^2)^{-\frac{1}{2}} - 1].$$

This equation reduces to $\frac{1}{2}m_0v^2$ for small v. If a variable relativistic mass, m, of the body is defined as being equal to the mass at low or no speed, m_0, multiplied by the ratio relating v^2 to c^2—*i.e.*, $m = m_0\,[1 - v^2/c^2]^{-\frac{1}{2}}$ (showing that as the velocity approaches the speed of light, the relativistic mass, m, approaches infinity)—and this is substituted in the equation, then the kinetic energy can also be written as

$$E_k = (m - m_0)\,c^2.$$

Rearranging this equation,

$$m = m_0 + E_k/c^2.$$

It is apparent from the equation that the relativistic mass of a body increases directly with its kinetic energy. Also, the equivalence of mass and energy is evident when it is expressed as energy = mass × (velocity of light)^2. This relationship is valid not only for kinetic but for all other forms of energy and thus it is possible to achieve a significant compression in the classical laws of physics. In classical physics the law of the conservation of mass and the law of conservation of energy were considered independent and equally vital fundamental statements. With relativity, the conservation of mass is a special case of a more general law of the conservation of energy. Direct experimental verification of this equivalence can be obtained from the study of nuclear reactions in which particles are added to existing nuclei with a consequent liberation of energy. Comparison of the mass of the ingredients going into the reaction with the mass of the products created in the reaction shows a mass discrepancy that agrees with the measured amount of energy liberated. Using these data, the equivalence of mass and energy can be directly tested. The equivalence also holds for chemical reactions in which the combining of ingredients results in either the liberation or the absorption of energy. For full details concerning nuclear energy, see NUCLEUS, ATOMIC; NUCLEAR FISSION; NUCLEAR FUSION.

Hamiltonian and Lagrangian formulations. In the sections above, the study of energy, generally started with a definition of the position and velocity of particles and then an understanding for their interaction developed. One of the products of this analysis is the identification of the quantity called energy and the discovery of the usefulness and the generality of the concept of energy. It is quite reasonable to question the possibility of beginning with the concept of energy and building a formalism that will allow the determination of a suitable set of coordinates for the particles in the system and the computation of their values. This approach has proven to be successful, and two quite complementary methods have evolved. One is the method of Lagrange, in which a quantity L, called the Lagrangian, is defined and which is equal to the difference between kinetic energy E_k and potential energy E_p:

Lagrange method

$$L = E_k - E_p.$$

In mechanical problems, the kinetic energy E_k is in the form of a quadratic in velocity, while the potential energy E_p is a function of position. Use of the Lagrangian method affords a technique for analyzing not only conservative systems but also nonconservative ones. The method results in an evaluation of the external forces imposed on a system along each of its coordinate axes. For a conservative system, this external force is zero by definition; for nonconservative ones, it is not. Solving the Lagrangian equations including the externally imposed forces allows the evaluation of nonconservative systems. Also, the method can be used to test a system to determine if it is indeed conservative or not. Of even greater value is the fact that the method is not restricted to mechanical systems. It can also be applied to systems in which the motions are not directly observable, including the phenomena in optics, acoustics, and electricity. The problem is to determine an appropriate set of coordinates such that they can be used to describe the potential energy, while a quadratic in their time derivative (*i.e.*, the "velocity") describes the kinetic energy.

Another approach is through use of the special function called the Hamiltonian. The Hamiltonian H is, in fact,

Hamiltonian method

$$H = E_k + E_p,$$

and thus has a numerical value equal to the total value of the energy of the system. The Hamiltonian and Lagrange

techniques each have greater value when applied to certain kinds of specific problems. A significant operational difference is that a solution of a given problem using the Lagrange method requires twice as many boundary conditions as the solution using the Hamiltonian.

SOME SPECULATIONS ON ENERGY

The concept of energy has demonstrated a remarkable survivability in the rapidly expanding understanding of the nature of things. To a large extent this survivability is no doubt a result of its adaptability to cover an increasingly complex set of phenomena that are revealed as ability to observe the physical world increases. Thus, in attempting to understand the structure of the nucleus within the context of the conservation of energy and with the hint afforded by the discovery of the positron, a complete set of antimatter particles was postulated and their existence verified experimentally. Indeed, it would appear that the concept of energy has value not only in explaining the readily observable physical world but, when data accumulate that cannot be explained with the incomplete facts, the use of the concept of the conservation of energy can be very helpful in indicating where to look for missing answers. On the other hand, the concept of the conservation of energy might suggest a "steady-state universe." It is difficult, however, to reconcile all of the observed data to a steady-state universe. Perhaps the data that have been collected have been obtained by observing quantities on only one-half of the universe.

BIBLIOGRAPHY. ROBERT B. LINDSAY and HENRY MARGENAU, *Foundations of Physics* (1936, reprinted 1963), carefully presented development of many of the concepts relative to energy, particularly useful since in many cases both the historical and the logical development of the ideas are presented; RICHARD P. FEYNMAN, ROBERT B. LEIGHTON, and MATTHEW SANDS, *The Feynman Lectures on Physics*, 3 vol. (1963–65), a detailed and careful treatment of all the concepts considered in this article, for the relatively informed reader; ROBERT RESNICK and DAVID HALLIDAY, *Physics for Students of Science and Engineering*, 2 vol. (1960), a widely adopted text used for teaching physics at the university level.

(R.L.Se.)

Energy Sources

Energy sources as a term covers all fuels, flowing water, sunlight, and the winds with which man, by the use of appropriate conversion devices, provides himself with necessary energy. This article is a survey of highlights in the discovery and exploration of primary, renewable, and expendable sources of energy, and of the conversion devices used to produce that energy. For information on specific engines and electric power see DIESEL ENGINE; GASOLINE ENGINE; TURBINE; ELECTRIC POWER.

HISTORY

Man first made use of energy from a source outside his own body when he discovered how to use fire, greatly enhancing his ability to survive.

Early sources of fuel

Wood was long the principal—in some regions the only—fuel. In addition to its use for heating and cooking, it became a source of charcoal for reducing ores to metals.

The first fossil fuels to be exploited were surface deposits of asphalt, peat, and coal; oil from surface seepage; and gas, venting from underground reservoirs. Asphalt is known to have been burned as early as 6000 BC in the Middle East, where huge surface deposits exist. The abundance of fossil fuels probably contributed to the development of burned brick, pottery, and other furnace products in this period. Asphalt played a notable role in the technological advances made under the Babylonian Empire, Mesopotamia (2500 to 538 BC), but its use diminished after 538 BC.

Natural gas escaping from underground reservoirs through vents sometimes became ignited and burned for years, even centuries. Even before 1000 BC natural gas was used in China for lighting, heating, and cooking. The Chinese, furthermore, are said to have drilled deep wells through hard formations to obtain natural gas, which they transported through bamboo pipelines.

Many parts of the world contain extensive coal deposits, but coal was difficult to mine with primitive tools and only the Chinese are known to have mined it to supplement natural gas and their dwindling supply of wood. Although coal was used to some extent in England in Roman times, finally becoming an article of commerce in the 13th century, fossil fuels were otherwise forgotten for nearly 2,000 years. Wood was the primary fuel for heating, cooking, and producing charcoal. With wood supplies dwindling rapidly, it was established in 1709 that coke from coal could be substituted for charcoal in the reduction of ores. The new coke industry, together with the production of coal gas for illumination, led to a rapid expansion of the coal-mining industry, soon to be accelerated even more by the introduction of steam engines.

Petroleum, known to the Babylonians though not extensively used, first came into prominence in the 17th century, when an oil well drilled in Modena, Italy, in 1640 provided fuel for street lighting. Commercial oil wells were developed in Romania in 1650. Oil had been discovered early in the history of the U.S., principally in New York and Kentucky, but the petroleum industry was not established in the U.S. until a major commercial oil well was brought into production in western Pennsylvania in 1859. Within ten years, kerosene for lamps was being shipped throughout the world from Pennsylvania. A generation later the automobile created a huge demand for gasoline. Since then the market for products made from this fossil fuel has grown rapidly and steadily.

The wind as an energy source for ship propulsion has been used by man for a very long time indeed. The earliest sailing ships, with both oars and sails, were replaced, as navigation techniques developed, by larger ships that depended on sails alone.

Windmills, probably invented in China, spread slowly into the Middle East. They were used in Persia (now Iran) before the 10th century and reached Europe, via the Arabs, by the 12th century. These early windmills were an important inanimate source of mechanical energy, especially for drainage and milling grain.

Waterwheels irrigated Babylonian fields 3000 years ago, and the Romans powered large flour mills and sawmills with them. In the Middle Ages this source of mechanical energy was much used in the Arab world and in Europe. Most early factories, depending upon water for power, were located beside dams. Although the amount of energy produced by waterwheels was small compared to modern power plants, water at one time provided most of the energy for industry. The waterwheel with its spillway was a part of many early American communities, and some rural areas still have mills operated by waterpower. The building of large hydraulic power plants followed the development of equipment for generating and transmitting electricity.

The development of steam power in the 18th century and its refinement and proliferation in the 19th began a revolution in energy exploitation that is still continuing. Its many facets can best be examined by starting with a definition of the related terms energy and power.

FORMS OF ENERGY

Energy is the *capacity* for doing work (this definition is generally accepted although thermal energy cannot be converted completely into work). Power is the *rate* of doing work and is thus the rate of energy flow.

Energy may be potential, kinetic, thermal, electrical, chemical, nuclear, or other forms. In addition, there are heat and work; *i.e.*, energy in the process of transfer from one body to another. After being transferred, energy is designated according to its nature. For example, heat transferred may become thermal energy, and work done may appear as mechanical energy.

Energy and motion

All forms of energy are associated with motion. For instance, a body has kinetic energy if it is in motion. A raised object, such as a trip-hammer, and a tensioned device, such as a spring or bow, though at rest have the potential for creating motion. They contain potential energy as a result of their position or configuration. Ther-

mal energy is a manifestation of motion of and within the molecules of a substance. For example, the temperature of a gas such as helium is related directly to the velocity of its molecules as they move in random motion. An electric current is the flow of electrons through a conductor. Chemical energy and nuclear energy are potential energies resulting from the configuration of atoms in a molecule or of subatomic particles in the nucleus of an atom.

Energy is measured by the work required to create the motion, the position, or the configuration. The minimum work required to raise a mass equals exactly the potential energy of the mass as a result of its being raised. Mechanical work is the product of force times the distance through which the force acts. Thus, one pound of force acting through one foot does one foot-pound of work. Or, to take an example from the new metric-based SI system, the work done by a force of one newton acting over a distance of one metre is one joule (or one watt-second).

A unit of thermal energy may be the foot-pound, although other units frequently are more convenient. The British thermal unit, BTU, is the quantity of heat required to raise the temperature of one pound of water one degree Fahrenheit near the point of maximum water density (39.1° F). In the SI (International System of Units) system, the preferred term for heat is the joule. One BTU is equal to 1,055 joules. An appropriate unit for expressing the quantities of energy in world fuel reserves is the *Q* unit, which is equal to 10^{18} BTU.

The nature of heat

The nature of energy and the relationship between thermal energy and mechanical energy were obscure and controversial for a long period of time. Greek philosophers believed that energy was a fluid permeating a substance. The nature of heat was clarified in 1798 by Sir Benjamin Thompson, count von Rumford. Engaged in superintending the boring of cannon in the military arsenal at Munich, he observed the amount of heat produced by the action of the boring bar upon the brass castings. Finally, James Prescott Joule showed conclusively in his experiments that the nature of heat and the nature of energy are the same, and that other forms of energy can be transformed into an equivalent amount of heat. In the best known of his experiments, Joule caused a paddle to rotate in a cylindrical vessel of water containing fixed paddles, acting as baffles, to prevent the water itself from rotating; the agitated water rose in temperature enough to be measured by a sensitive thermometer. Joule determined the mechanical equivalent of the energy required to raise the temperature of the water. Refinements in this measurement led to the value of 778 foot-pounds per BTU. This is the mechanical equivalent of heat.

Since power is the rate of doing work, such units as foot-pounds per second and BTU per hour are used. The horsepower unit was established by James Watt, who measured the maximum rate of work done by horses. He multiplied the result by a factor of 1.5 to rate his steam engines conservatively in terms of horsepower. One horsepower equals 550 foot-pounds per second.

Energy can be changed in form, but the total amount of energy in an isolated system cannot be changed. Thus, energy can neither be created nor destroyed. This statement summarizes many experimental observations of the conversion of energy, and this fact is useful in solving many problems involving energy. For example, the total energy contained in the fuel and air for an engine must equal the sum of the work done by the engine, the thermal energy from the engine, and the energy present in the combustion products.

Although the total amount of energy in an isolated system remains unchanged, there may be a great difference in the quality of different forms of energy. Many forms of energy, in theory, can be transformed completely into work or into other forms of energy. This is true for mechanical energy and electrical energy. The random motions of constituent parts of a material associated with thermal energy, however, represent energy that is not available completely for conversion into directed energy.

The Carnot cycle

In 1824 Sadi Carnot described a theoretical power cycle of maximum efficiency for converting thermal into mechanical energy. He demonstrated that this efficiency is determined by the magnitude of the temperatures at which heat energy is added and waste heat is given off during the cycle. A practical engine operating on the Carnot cycle has never been devised, but the Carnot cycle determines the maximum efficiency of thermal energy conversion into any form of directed energy.

According to the theory of relativity proposed by Albert Einstein in 1905, the apparent mass of a body increases in proportion to its kinetic energy. Mass and energy are thus said to be equivalent, and one can be converted into the other. This relationship between mass and kinetic energy may be shown to hold for all other forms of energy as well. The principle of conservation of mass proves to be merely a special case of the broader principle of conservation of energy. The equivalence of mass and energy has been confirmed by investigations of atomic nuclei whose energies affect mass in a measurable amount. In chemical reactions the change in mass with energy is not measurable. Even in nuclear fission the decrease is small for the large amount of energy released.

ENERGY CONVERSION

Energy can be changed from one form to another in many ways. Useful mechanical or electrical energy is produced by fuel-burning heat engines, by electric generators, batteries, photoelectric cells, and by many other devices, most of which have important applications.

The steam engine

Man's need for mechanical energy first became acute toward the end of the 17th century, when deposits of coal and metal ores near the surface were becoming exhausted, and the deep pits were flooded with groundwater. The problem of pumping water out of the useless mines attracted the attention of such men as Thomas Newcomen, who developed a steam pumping engine in England in 1705. By converting thermal energy into mechanical energy, this engine made available quantities of mechanical energy limited only by the capacity of the engines and the availability of fuels. The Newcomen engine converted less than 1 percent of thermal energy into work. But because it filled an important need, it continued to be used until it was replaced by Watt's engine, invented in 1763. Equipped with a separate condenser to lower the temperature of the exhaust steam, Watt's engine had an efficiency of more than 2 percent. By 1900 many improvements, including three separate expansion cycles and higher steam temperatures, had raised the efficiency of the steam engine to about 17 percent. This engine has been replaced by the more compact and efficient steam turbine. With improvements and the use of very high temperature steam, the efficiency of thermal energy conversion is approaching 40 percent.

The energy conversion processes of the steam power plant were described by William Rankine, a Scottish physicist whose "Rankine cycle" (1859) explained how a liquid in a boiler vaporized by the addition of heat converts part of this energy into mechanical energy when the vapour expands in an engine. As the exhaust vapour is condensed by a cooling medium such as water, heat is lost from the cycle. The condensed liquid is pumped back into the boiler. This concept of a power cycle is useful to engineers in developing equipment and designing power plants (see also STEAM POWER).

Poor materials and faulty methods of construction were the cause of many disastrous explosions in early steam boilers. The resultant casualties and property losses motivated the Reverend Robert Stirling to invent a power cycle that operated without a high-pressure boiler. In his engine, patented in 1816, air or some other gas, heated by external combustion through a heat exchanger, was displaced, compressed, and expanded by two pistons. At this time, when thermodynamics was still unknown as an engineering science, Stirling originated the idea of a regenerator to store thermal energy during part of the cycle and then return this energy to the working fluid; *i.e.*, regenerate. His engine in theory is equal in efficiency to

that of the Carnot engine. Although a successful engine was built for a factory in 1843, general use has been restricted by the engine's high cost. Nevertheless, until about 1920, small engines of this general type pumped water on farms and generated electricity for small communities, burning wood or coal in the firebox. The high efficiency and freedom from air-polluting emissions have created renewed interest in the Stirling engine.

The gas turbine. The operation of a gas turbine power plant, such as a turbojet engine, is described by the Brayton cycle, originally devised for reciprocating engines—engines in which combustion took place at constant pressure in the cylinder. The urgent need for improved aircraft propulsion during World War II led to the development of efficient, lightweight air compressors and gas turbines that made possible the turbojet engine operating on the Brayton cycle. Compressed air, heated by the combustion of petroleum, is used to turn a turbine to drive the compressor, while excess energy accelerates the exhaust gas to high velocity for producing thrust (see also JET ENGINE).

The Brayton cycle

The internal-combustion engine. Engines in which combustion takes place within the working cylinder originated early in the 19th century. Not until 1876, however, did Nikolaus August Otto of Germany demonstrate his practical gas engine. Its many applications, which included motor vehicles, led to the rapid development of the internal-combustion engine, and with that came a great demand for gasoline.

The internal-combustion engine patented by Rudolf Diesel in 1892 was developed to burn powdered coal, at that time an unwanted waste product of coal mining. In the diesel engine, air is compressed to a temperature high enough to ignite fuel injected into the cylinder. Changing the fuel to oil resulted in an engine suited to many industrial and marine applications, including the propulsion of submarines (see also DIESEL ENGINE).

Direct energy-conversion devices. Requirements for energy for space vehicles led to an exhaustive investigation, starting in 1955, of all possible energy sources (much of the literature on this subject has come from these programs). Direct energy-conversion devices are of interest for providing electrical energy in space vehicles because of their reliability and their lack of moving parts. Sometimes called static energy-conversion devices, they can be considered to use electrons as the working fluid in place of the vapour or gas in a dynamic heat engine.

In a thermoelectric generator, for example, electrons driven by thermal energy across a potential difference at the junction of two conductors made of dissimilar materials create an electric current. This effect was discovered in 1821 by Thomas Johann Seebeck of Germany when he observed that a compass needle near a circuit made of different conducting materials was deflected when one of the junctions was heated. He investigated many materials, including some now classified as semiconductors, that produce electrical energy with an efficiency of 3 percent. This efficiency was comparable to that of the steam engines of that time. Nevertheless, the usefulness of the discovery of the thermoelectric effect went unrecognized as a means for producing electricity because of Seebeck's interpretation of this as a magnetic effect caused by a difference in temperatures. Only recently have thermoelectric generators come into their own, producing electricity in special situations in which static thermal energy conversion devices with high reliability but a modest efficiency of some 5 percent are preferred (see also THERMOELECTRIC DEVICES).

Another static energy conversion device that has been investigated extensively for the space program is the thermionic generator. Electrons are driven in a vacuum by thermal energy from a high-temperature electrode, across a potential difference, to a cooler electrode. Thomas Alva Edison was the first to observe an electric current between an incandescent filament and a cold electrode in an evacuated tube, but not until 1958 did active development of thermionic generators begin. Although their performance has since been improved greatly, the lack of suitable materials for use at very high temperatures and problems in construction have limited their reliability. When these problems are solved, thermionic generators will be properly suited for use with nuclear reactors or radioisotope heat sources (see also THERMIONIC DEVICES).

Thermionic generators

Radiant energy also can drive electrons across a potential difference at a semiconductor junction in which the concentrations of impurities are different on the two sides of the junction. The silicon solar cell, announced in 1952, is a reliable power source with no moving parts and is particularly suited for power generation in space where the sun presents an inexhaustible source of energy. Efficiencies of the order of 14 percent have been achieved in converting solar energy directly into electrical energy. The cost of solar cells limits their commercial use.

Chemical energy can be converted directly into electrical energy by electrochemical reactions in a battery or fuel cell. The efficiency of the process is not limited by the maximum theoretical efficiency of energy conversion when chemical energy is converted to thermal energy for operation of a heat engine. Batteries and fuel cells with efficiencies of over 80 percent (at very low current) have been demonstrated, whereas the best efficiencies obtained in heat engines are about half this value.

In 1800 Alessandro Volta constructed an electrochemical cell for producing electricity from two metals in contact with a salt solution. A battery for producing higher potential, called Volta's pile, was made of alternate plates of silver and zinc separated by cloth or paper soaked in salt solution. Many types of batteries were developed during the 19th century, and this source of electricity made possible early electrochemical discoveries. The lead–acid secondary battery for storing electrical energy was developed by Gaston Planté of France in 1859. Among its many applications was the electric automobile.

In a fuel cell, another electrochemical producer of electricity, continuous operation is achieved by feeding fuel and oxidizer to the cell and removing the reaction products. Sir William Grove constructed such a cell in 1839, with hydrogen and oxygen reacting on platinum electrodes to produce electricity. One of the main objectives of electrochemical research was to find a process for the direct electrochemical oxidation of fuels, such as coal, to supply the growing requirements for electricity. When large generators driven by steam power plants and waterpower were developed to meet these requirements, interest in fuel cells dwindled. The need for silent portable electric generators for electronic equipment during World War II, and, more recently, for efficient energy conversion devices in the space program, has created renewed interest in fuel cell research. Interesting developments include the electrochemical oxidation of alcohols and hydrocarbon fuels to produce electricity and the hydrogen–oxygen fuel cell for manned space travel (see also BATTERIES AND FUEL CELLS).

The fuel cell

Many other devices produce electrical energy. Some of these produce a potential difference by heating or applying a mechanical force on certain crystals (piezoelectricity). The magnetohydrodynamic generator utilizes a conducting fluid flowing through a magnetic field to generate electricity. In reverse, electrical energy can accelerate a fluid to produce thrust in an electric rocket motor (see also MAGNETOHYDRODYNAMIC DEVICES).

The processes by which muscles convert chemical energy into mechanical energy are the subject of intensive research and are now partly understood. Blood carries food material and oxygen to the muscle where they react through the mediation of a system of proteinaceous substances called enzymes, producing heat as well as mechanical energy. A muscle at rest produces very little heat; but very soon after stimulation, resistance to stretch develops rapidly and heat production increases. Additional heat is produced as the muscle shortens. The amount is proportional to the distance of shortening and independent of load. A rapidly shortening muscle exerts less tension than when shortening slowly, and less thermal energy is released per unit distance. Perhaps some

day these remarkable processes can be applied to inanimate energy conversion (see also MUSCLE CONTRACTION).

ENERGY SOURCES AND THEIR DEPLETION

Primary energy sources. The primary sources of energy include nuclear fusion, nuclear fission, radioactivity, and the motion of the earth and the moon. All forms of energy can be traced to these origins.

Nuclear fusion. Energy is released in the sun by thermonuclear reactions that fuse hydrogen and other light elements into heavier elements, releasing very large amounts of energy. This is the source of the radiant energy that illuminates the earth and provides a practically inexhaustible supply of energy on earth and in space. Winds in the earth's atmosphere are created by the thermal effects of solar energy, and even the energy in flowing streams is produced indirectly from solar radiation.

Nuclear power generation

The time will undoubtedly come when, after other fuels are consumed, man will have to depend upon thermonuclear reactions for energy. While large amounts of energy will be produced from nuclear fission, the total will be insignificant compared with the potential of nuclear fusion. The attainment of practical power generation from controlled fusion is perhaps the most formidable technical challenge that man has faced. If the problem can be solved, deuterium (heavy hydrogen) is available from water in quantity sufficient to provide virtually unlimited energy (see also NUCLEAR FUSION).

Nuclear fission. Fissionable nuclear fuels are remarkable sources of energy. The energy released by the fissioning of one atom of uranium, for example, is several million times greater than that produced by the combustion of one molecule of gasoline. The knowledge of means for releasing atomic energy came about through a series of discoveries with the final development being motivated by the effort to create superior bombs for military use.

Radioactivity. The energy released in nuclear fission is in the form of radiations and kinetic energy of the fission products, which are released at high velocity. Most of these become thermal energy in the fuel elements as the fission products come to rest and as some of the radiations are absorbed. The thermal energy can be converted in part into mechanical energy, which can then be converted into electrical energy by a steam turbine driving a generator.

Only 0.72 percent of natural uranium is composed of the fissionable isotope, uranium-235. This can be separated from uranium-238 by diffusion processes to construct bombs and compact nuclear reactors. Since the energy from nuclear fuels would be quite limited if only the uranium-235 were utilized, reactors have been developed to produce fissionable plutonium-239 from uranium-238, in addition to producing thermal energy for the generation of electricity. Similarly, fissionable uranium-233 can be produced from nonfissionable thorium to increase greatly the energy available from nuclear fuels, which far exceeds the energy available from fossil fuels (see also NUCLEAR FISSION and NUCLEAR REACTOR).

Many radioactive isotopes, particularly of the heavy elements, are unstable, releasing energy as they decay. Not only do they occur naturally in thorium and uranium and in some other minerals but they can be created through fission and other nuclear processes. Thus, there are a large number of both natural and artificial radioisotopes whose properties vary in decay rate and energy production. Natural radioactive materials make up only a minute part of the earth, but they release sufficient energy to maintain the elevated temperature of the earth's interior. Radioactivity is therefore the origin of energy obtained from the earth in geothermal steam wells.

Radioisotopes can produce electrical energy in small amounts for special purposes. Demonstrated as early as 1913 was a type of nuclear battery for generating electricity directly from radiations released in radioactive decay. In 1953 a toy steam engine was operated with the radioisotope polonium-210 as its heat source to publicize power generation from nuclear energy. This led to the development of a radioisotope-heated thermoelectric generator demonstrated in 1954. Similar generators have been incorporated in space satellites to provide part of the power for operating the electronic equipment on board. Many other radioisotope power sources have since been developed, and some of these have been left to generate energy for experimental apparatus set up on the moon (see also RADIOISOTOPES, APPLICATIONS OF).

Recurring energy sources. Because some kinds of energy are continuously being created from primary sources, they are said to be recurring. These include solar energy, the energy in wind and flowing streams, geothermal energy, the energy in tides, fuel from trees and other plants, and a possible source of energy in thermal gradients in the ocean (oceanothermic energy). Although the total amount of energy in these recurring sources is large in relation to world energy needs, only a very small amount can be collected at reasonable cost. Thus, these sources provide only a small part of the world's total energy requirements.

Solar energy

Solar radiation outside the earth's atmosphere is 1.4 kilowatts per square metre. Expressed in terms of Q units ($Q = 10^{18}$ BTU) the amount of solar radiation striking the earth's atmosphere in one year is 5,000 Q. More than half of this reaches the surface of the earth, with 900 Q per year being absorbed over land and the remainder over water. The amount of solar energy absorbed each year by the process of photosynthesis in land vegetation has been estimated to be 0.15 Q. The energy in streams is about 0.17 Q per year, of which only about 2 percent is utilized by man. The energy in the winds is about 90 Q annually, but the amount that drives windmills is relatively insignificant.

Many attempts have been made to use solar energy to operate power plants, to cook, and to heat. Solar concentrators such as large parabolic mirrors, accurately oriented toward the sun, are necessary to produce the required high temperatures for generating power and cooking. Their cost makes power generation uneconomical if other energy sources are available. The practicability of heating homes with solar radiation depends upon their location and the cost of other fuels.

The wind contains tremendous amounts of energy, but it is intermittent and diffuse. Windmills usually are expensive relative to the energy delivered. Nevertheless, windmills are being reconsidered as energy sources, even in industrial nations, with a unit for generating 24,000 kilowatts of electric power being studied. Interest in energy from the winds will increase as fossil-fuel costs rise and as methods of energy transmission and storage are improved. This source of energy is predicted to provide at least a small percentage of the world's requirements.

The energy that could be recovered from undammed streams to generate power economically has been estimated at some 0.02 Q per year. Not included in this estimate is the energy in large rivers in remote areas of Asia or from the proposed closing of the Straits of Gibraltar. If the Straits are closed, evaporation would lower the Mediterranean Sea and power could be produced at the Straits and at river mouths. These sources of energy and many other proposed hydroelectric projects also will become more attractive as fossil-fuel reserves diminish. At present, hydroelectric power is estimated to provide 2 percent of the world's energy requirements.

Geothermal energy flows from the hot interior of the earth to the surface, where it is lost by radiation into space. Studies have been made of methods to extract energy directly from the earth through pipes buried at great depths, but the amount of energy obtained would be small because the earth is a poor thermal conductor. No economically feasible method seems possible, except in a few fortunate areas that have wells of natural steam or hot water. The most famous of these is at Larderello, Italy, where wells drilled as deep as 2,000 feet produce superheated steam at great pressure. The steam not only drives turbines to generate electricity but in addition contains chemical impurities that are recovered as valuable by-products. Another is near Reykjavík, Iceland, where the hot water from geothermal wells is piped to

heat buildings. A large natural vent of superheated steam in a remote region of New Zealand was once considered as a source of energy for the production of heavy water (deuterium oxide) for the atomic energy program. In California, however, geothermal steam was generating electricity in the 1960s and the facilities to take greater advantage of this energy from the earth were being expanded in the early 1970s.

Experiments by Georges Claude in 1929 demonstrated that energy can be obtained from the ocean in tropical regions where the difference in temperature between the warm water at the surface and cold water at great depths can operate a Rankine power cycle. The equipment for extracting oceanothermic energy is expensive and subject to storm damage, but there is a continuing interest in the process.

Tidal energy

There is a great amount of energy in the tides, but the ocean has been a difficult energy source to harness. This energy comes from the energy of rotation of the earth. As a result, the rate of rotation of the earth is slowing slightly. Spring tides sometimes rise as high as 52 feet in some parts of the world and many areas contain bays that favour the construction of dams to create storage basins, but the energy source is intermittent and varies with the seasons. At Rance, France, a successful tidal power project was developed in the 1960s. Several other sites, including the Bay of Fundy, in North America, have been studied. Many small waterwheels have been operated by tidewater. Even though only a small fraction of the energy in tides can be captured, this source of energy is expected to be put to greater use in the future.

Until about AD 1700, wood and farm wastes (stalks, cobs, and dried dung) provided most of the energy requirements of the world. In many countries where other fuels are not available, these sources still provide energy for heating and cooking. In forested areas wood is used extensively for fuel, supplying about 4 percent of the world's energy requirements. A secondary fuel, ethyl alcohol, is produced in quantity by fermentation of farm products. Tropical countries have expressed some interest in alcohol as a fuel for internal-combustion engines.

Energy storage. Most energy systems need a provision for storing energy to meet peak power requirements or emergencies or to drive systems that operate entirely on stored energy. Most machines have provisions for storing either potential or kinetic energy: a counterweight stores potential energy when it is lifted; springs of all types store energy; water pumped into an elevated reservoir stores energy; capacitors in circuits store small amounts of electricity; batteries store large amounts of energy and for long periods. Thermal energy can be stored by heating a material or by a phase change, such as evaporating water into steam. Fuels release stored chemical energy by combustion with natural oxygen in the atmosphere or with liquid oxygen. The performance of many products is limited by their energy storage capacity.

Expendable energy sources. The fossil fuels (coal, lignite, peat, petroleum, and natural gas) are natural resources that are being consumed at an appreciable rate. Although deposits continue to be discovered, the reserves of these fuels remaining in the earth are finite. The amounts that can be recovered economically are difficult to estimate, depending as they do upon the future value of the fuels and their mining cost. For example, coal occurs in seams of various thicknesses and at various depths. If coal seams no less than 24 inches thick are to be mined, much less coal is available than if recovery extends to seams as narrow as 14 inches. Estimating oil reserves presents similar difficulties. A 1953 estimate put minimum recoverable resources at 38 *Q* (32 *Q* of coal, 5 *Q* of oil and gas, and 1 *Q* of oil shales and tar), whereas other estimates put recoverable coal reserves at 200 *Q*. Nevertheless, the quantities of fossil fuels available in the earth's crust are not large compared to increasing rates of consumption. Although the amount of recoverable fuel will increase as more efficient mining methods are developed, certainly at some time in the future uranium and thorium will supply a major part of world energy requirements. Ultimately, if the technical problems of controlling nuclear fusion are solved, man will have an essentially unlimited source of energy.

BIBLIOGRAPHY. P.C. PUTNAM, *Energy in the Future* (1953), a study of maximum plausible demands for energy for the next century; E. AYRES and C.A. SCARLOTT, *Energy Sources: The Wealth of the World* (1952), a history of the use of fossil fuels with estimates of reserves; H. THIRRING, *Power Production* (1956; U.S. title, *Energy for Man*, 1958), a compilation of information on energy sources as they relate to future requirements; *Proceedings of the United Nations Conference on New Sources of Energy* (1964), papers on many energy sources including geothermal energy; C.R. RUSSELL, *Elements of Energy Conversion* (1967), descriptions of energy conversion and storage devices and their principles of operation; A. PARKER, *Survey of Energy Resources: World Power Conference* (1968), data on fossil and nuclear fuel reserves and hydroelectric capabilities of the nations.

(C.R.R.)

Engels, Friedrich

The German socialist philosopher Friedrich Engels was, with Karl Marx, one of the two founders of modern Communism. As the junior member of the "partnership" that worked out the theories of what has come to be known as Marxism or Marxist Communism, Engels played an indispensable role. Marx once noted, with pleasure, that their enemies used the singular verb when speaking of "Marx-Engels."

Born on November 28, 1820, in Barmen, an industrial city in the Prussian Rhine Province, now in West Germany, Engels grew up in the environment of a family marked by moderately liberal political views, a steadfast loyalty to Prussia, and a pronounced Protestant faith. His father was the owner of a textile factory in Barmen and also a partner in the Ermen & Engels cotton plant in Manchester, England. Even after Engels openly pursued the revolutionary goals that threatened the traditional values of the family, he usually could count on financial aid from home. The influence of his mother, to whom he was devoted, may have been a factor in preserving the tie between father and son.

Aside from such disciplinary actions as the father considered necessary in rearing a gifted but somewhat rebellious son, the only instance in which his father forced his will on Engels was in deciding upon a career for him. Engels did attend a *Gymnasium* (secondary school), but he dropped out a year before graduation, probably because his father felt that his plans for the future were too undefined. Engels showed some skill in writing poetry, but his father insisted that he go to work in the expanding business. Engels, accordingly, spent the next three years (1838–41) in Bremen acquiring practical business experience in the offices of an export firm.

In Bremen, Engels began to show the capacity for living the double life that characterized his middle years. During regular hours, he operated effectively as a busi-

Novosti Press Agency

Engels, print by H. Schey.

ness apprentice. An outgoing person, he joined a choral society, frequented the famed *Ratskeller*, became an expert swimmer, and practiced fencing and riding (he outrode most Englishmen in the fox hunts). Engels also cultivated his capacity for learning languages; he boasted to his sister that he knew 24. In private, however, he developed an interest in liberal and revolutionary works, notably the banned writings of "Young German" authors such as Ludwig Börne, Karl Gutzkow, and Heinrich Heine. But he soon rejected them as undisciplined and inconclusive in favour of the more systematic and all-embracing philosophy of Hegel as expounded by the "Young Hegelians," a group of leftist intellectuals, including the theologian and historian Bruno Bauer and the anarchist Max Stirner. They accepted the Hegelian dialectic, basically that rational progress and historical change result from the conflict of opposing views, ending in a new synthesis. The Young Hegelians were bent on accelerating the process by criticizing all that they considered irrational, outmoded, and repressive. As their first assault was directed against the foundations of Christianity, they helped convert an agnostic Engels into a militant atheist, a relatively easy task since by this time Engels' revolutionary convictions made him ready to strike out in almost any direction.

Hegel and the dialectic process

In Bremen Engels also demonstrated his talent for journalism by publishing articles under the pseudonym of Friedrich Oswald, perhaps to spare the feelings of his family. He possessed pungent critical abilities and a clear style, qualities that were utilized later by Marx in articulating their revolutionary goals.

After returning to Barmen in 1841, the question of a future career was shelved temporarily when Engels enlisted as a one-year volunteer in an artillery regiment in Berlin. No antimilitarist disposition prevented him from serving commendably as a recruit; in fact, military matters later became one of his specialties. In the future, friends would often address him as "the general." Military service allowed Engels time for more compelling interests in Berlin. Though he lacked the formal requirements, he attended lectures at the university. His Friedrich Oswald articles gained him entrée into the Young Hegelian circle of "The Free," formerly the "Doctors Club" frequented by Karl Marx. There he gained recognition as a formidable protagonist in the philosophical battles, mainly directed against religion.

After his discharge in 1842, Engels met Moses Hess, the man who converted him to communism. Hess, the son of wealthy Jewish parents and a promoter of radical causes and publications, demonstrated to Engels that the logical consequence of the Hegelian philosophy and dialectic was communism. Hess also stressed the role that England, with its advanced industry, burgeoning proletariat, and portents of class conflict, was destined to play in future upheavals. Engels eagerly seized the opportunity to go to England, ostensibly to continue his business training in the family firm in Manchester.

Conversion to communism

In England (1842–44), Engels again functioned successfully in business. After hours, however, he pursued his real interests: writing articles on communism for continental and English journals, reading books and parliamentary reports on economic and political conditions in England, mingling with workers, meeting radical leaders, and gathering materials for a projected history of England that would stress the rise of industry and the wretched position of the workers.

In Manchester, Engels established an enduring attachment to Mary Burns, an uneducated Irish working girl, and though he rejected the institution of marriage, they lived together as man and wife. In fact, the one serious strain in the Marx-Engels friendship occurred when Mary died in 1863 and Engels thought that Marx responded a little too casually to the news of her death. In the future, however, Marx made a point of being more considerate, and, when Engels later lived with Mary's sister Lizzy, on similar terms, Marx always carefully closed his letters with greetings to "Mrs. Lizzy" or "Mrs. Burns." Engels finally married Lizzy, but only as a deathbed concession to her.

In 1844 Engels contributed two articles to the *Deutsche-französische Jahrbücher* ("German-French Yearbooks"), which were edited by Marx in Paris. In them Engels put forth an early version of the principles of scientific socialism. He revealed what he regarded as the contradictions in liberal economic doctrine and set out to prove that the existing system based on private property was leading to a world made up of "millionaires and paupers." The revolution that would follow would lead to the elimination of private property and to a "reconciliation of humanity with nature and itself." On his way to Barmen, Engels went to Paris for a ten-day visit with Marx, whom he had earlier met in Cologne. This visit resulted in a permanent partnership to promote the socialist movement.

The partnership with Marx

Back in Barmen, Engels published *Die Lage der arbeitenden Klasse in England* (1845; *The Condition of the Working Class in England in 1844*, 1887), a classic in a field that later became Marx's specialty. Their first major joint work was *Die deutsche Ideologie* (1845; *The German Ideology*, 1938), which, however, was not published until more than 80 years later. It was a highly polemical critique that denounced and ridiculed certain of their earlier Young Hegelian associates and then proceeded to attack various German socialists who rejected the need for revolution. Marx's and Engels' own constructive ideas were inserted here and there, always in a fragmentary manner and only as corrective responses to the views they were condemning.

Upon rejoining Marx in Brussels in 1845, Engels endorsed his newly formulated economic, or materialistic, interpretation of history, which assumed an eventual communist triumph. That summer he escorted Marx on a tour of England. Thereafter he spent much time in Paris, where his social engagements did not interfere significantly with his major purpose, that of attempting to convert various émigré German worker groups—among them a socialist secret society, the League of the Just—as well as leading French socialists to his and Marx's views. When the league held its first congress in London in June 1847, Engels helped bring about its transformation into the Communist League.

Marx and he together persuaded a second Communist Congress in London to adopt their views. The two men were authorized to draft a statement of Communist principles and policies, which appeared in 1848 as the *Manifest der kommunistischen Partei* (commonly known as *Communist Manifesto*). It included much of the preliminary definition of views prepared earlier by Engels in the *Grundsätze des Kommunismus* (1847; *Principles of Communism*, 1925) but was primarily the work of Marx.

Communist Manifesto

The Revolution of 1848, which was precipitated by the attempt of the German states to throw off an authoritarian, almost feudal, political system and replace it with a constitutional, representative form of government, was a momentous event in the lives of Marx and Engels. It was their only opportunity to participate directly in a revolution and to demonstrate their flexibility as revolutionary tacticians with the aim of turning the revolution into a Communist victory. Their major tools were the *Neue Rheinische Zeitung, Organ der Demokratie*, which Marx edited in Cologne with the able assistance of Engels. Such a party organ, then appearing in a democratic guise, was of prime importance for their purposes; with it they could furnish daily guidelines and incitement in the face of shifting events, together with a sustained criticism of governments, parties, policies, and politicians.

After the failure of the revolution, Engels and Marx were reunited in London, where they reorganized the Communist League and drafted tactical directives for the Communists in the belief that another revolution would soon take place. But how to replace his depleted income soon became Engels' main problem. To support both himself and Marx, he accepted a subordinate position in the offices of Ermen & Engels in Manchester, eventually becoming a full-fledged partner in the concern. He again functioned successfully as a businessman, never allowing his Communist principles and criticism of capitalist ways to interfere with the profitable operations of his firm. Hence he was able to send money to Marx

constantly, often in the form of £5 notes, but later in far higher figures. When Engels sold his partnership in the business in 1869, he received enough for it to live comfortably until his death in 1895 and to provide Marx with an annual grant of £350, with the promise of more to cover all contingencies.

Engels, who was forced to live in Manchester, corresponded constantly with Marx in London and frequently wrote newspaper articles for him; he wrote the articles that appeared in the *New York Tribune* (1851–52) under Marx's name and that were later published under Engels' name as *Revolution and Counter-Revolution in Germany in 1848*. In the informal division of labour that the two protagonists of communism had established, Engels was the specialist in nationality questions, military matters, to some extent in international affairs, and in the sciences. Marx also turned to him repeatedly for clarification of economic questions, notably for information on business practices and industrial operations.

Marx's *Kapital*, his most important work, bears in part a made-in-Manchester stamp. Marx similarly called on Engels' writing facility to help "popularize" their joint views. While Marx was the brilliant theoretician of the pair, it was Engels, as the apt salesman of Marxism directing attention to *Das Kapital* through his reviews of the book, who implanted the thought that it was their "bible." Engels almost alone wrote *Herrn Eugen Dührings Umwälzung der Wissenschaft* (1878; *Herr Eugen Dühring's Revolution in Science* [*Anti-Dühring*], 1934), the book that probably did most to promote Marxian thought. It destroyed the influence of Karl Eugen Dühring, a Berlin professor who threatened to supplant Marx's position among German Social Democrats.

After Marx's death (1883), Engels served as the foremost authority on Marx and Marxism. Aside from occasional writings on a variety of subjects and introductions to new editions of Marx's works, Engels completed volumes 2 and 3 of *Das Kapital* (1885 and 1894; *Capital*, 1907 and 1908) on the basis of Marx's uncompleted manuscripts and rough notes. All the while he corresponded extensively with German Social Democrats and followers everywhere, so as to perpetuate the image of Marx and to foster some degree of conformity among the "faithful." His work was interrupted when he was stricken with cancer; he died in London on August 5, 1895.

Assessment

During his lifetime, Engels experienced, in a milder form, the same attacks and veneration that fell upon Marx. An urbane individual with the demeanour of an English gentleman, Engels customarily was a gay and witty associate with a great zest for living. He had a code of honour that responded quickly to an insult, even to the point of violence. As hatchetman of the "partnership," he could be most offensive and ruthless, so much so that in 1848 various friends attempted unsuccessfully to persuade Marx to disavow him.

Except for the Communist countries, where Engels has received due recognition, posterity has generally lumped him together with Marx without adequately clarifying Engels' significant role. The attention Engels does receive is likely to be in the form of a close scrutiny of his works to discover what differences existed between him and Marx. As a result, some scholars have concluded that Engels' writings and influence are responsible for certain deviations from, or distortions of, "true Marxism" as they see it. Yet scholars in general acknowledge that Marx himself apparently was unaware of any essential divergence of ideas and opinions. The Marx-Engels correspondence, which reveals a close cooperation in formulating Marxist policies, bears out that view.

MAJOR WORKS

Die Lage der arbeitenden Klasse in England (1845; trans. as *The Condition of the Working Class in England in 1844*, with an appendix and preface by Engels, 1887); (with Karl Marx) *Manifest der kommunistischen Partei* (1848; "The Communist Manifesto," trans. in *The Red Republican*, 1850); "Der deutsche Bauernkrieg," in *Neue Rheinische Zeitung Politisch-ökonomische Revue* (1850, republished by Engels as a book in 1870; *The Peasant War in Germany*, 1926); *Herrn Eugen Dührings Umwälzung der Wissenschaft* (1877–78; *Herr Eugen Dühring's Revolution in Science* [*Anti-Dühring*], 1934); *Die Entwicklung des Sozialismus von der Utopie zur Wissenschaft* (1882), three chapters from *Anti-Dühring* with an appendix on "The Mark," trans. by E. Aveling as *Socialism, Utopian and Scientific* (1892); *Der Ursprung der Familie, des Privateigenthums und des Staats* (1884; *The Origin of the Family, Private Property and the State*, 1902); *Zur Wohnunsfrage*, six articles in the *Volksstaat* (1872, reprinted as a pamphlet in 1887; *The Housing Question*, 1935); *Ludwig Feuerbach und der Ausgang der klassischen deutschen Philosophie* (1888; *Ludwig Feuerbach and the Outcome of Classical German Philosophy*, 1934); *Revolution and Counter-Revolution in Germany in 1848*, ed. by Eleanor Marx Aveling (1896), 20 articles from the *New York Tribune*, 1851–52, written by Engels, revised and signed by Marx (U.S. edition, *Germany: Revolution and Counter Revolution*, 1933).

BIBLIOGRAPHY. GUSTAV MAYER, *Friedrich Engels: A Biography* (1936), an abridged translation of the two-volume comprehensive and authoritative German work; OSCAR J. HAMMEN, *The Red '48ers: Karl Marx and Friedrich Engels* (1969), a study of Marx and Engels to 1850, notably as revolutionary practitioners in 1848–49; Z.K. COATES, *The Life and Teaching of Friedrich Engels* (1945; first pub. as *The Life and Work of Friedrich Engels*, 1920), still useful; W.O. HENDERSON (ed.), *Engels: Selected Writings* (1967), excellent excerpts from Engels' writings and letters, with a chronology of his life and principal works.

(O.J.H.)

Engineering

Engineering is the professional art of applying science to the optimum conversion of the resources of nature to benefit man. The words engine and ingenious are derived from the same Latin root, *ingenerare*, meaning "to create." The early English verb *engine* meant "to contrive." Thus the engines of war were devices such as catapults, floating bridges, and assault towers; their designer was the "engine-er," or military engineer. His counterpart was the civil engineer, who applied the same knowledge and skills to devising structures, streets, water supplies, sewage systems, and other projects of benefit to civilians.

Associated with engineering is a great body of special knowledge; preparation for professional practice involves extensive training in the application of that knowledge. Standards of engineering practice are maintained through the efforts of professional societies, usually organized on a national or regional basis, with each member acknowledging a responsibility to the public over and above responsibilities to his employer or to other members of his society.

Engineering is an art requiring the judgment necessary to adapt knowledge to practical purposes, the imagination to conceive original solutions to problems, and the ability to predict performance and cost of new devices or processes.

The function of the scientist is to know, while that of the engineer is to do. The scientist adds to the store of verified, systematized knowledge of the physical world; the engineer brings this knowledge to bear on practical problems. Engineering is based principally on physics, chemistry, and mathematics and their extensions into materials science, solid and fluid mechanics, thermodynamics, transfer and rate processes, and systems analysis.

Unlike the scientist, the engineer is not free to select the problem that interests him; he must solve problems as they arise; his solution must satisfy conflicting requirements. Usually efficiency costs money; safety adds to complexity; improved performance increases weight. The engineering solution is the optimum solution, the end result that, taking many factors into account, is most desirable. It may be the cheapest for a specified level of performance, the most reliable within a given weight limit, the simplest that will satisfy certain safety requirements, or the most efficient for a given cost. In many engineering problems, the social costs are significant.

Resources employed by engineers

Engineers employ two types of natural resources—materials and energy. Materials are useful because of their properties: their strength, ease of fabrication, lightness, or durability; their ability to insulate or conduct; their chemical, electrical, or acoustical properties. Important sources of energy include fossil fuels (coal, petroleum,

gas), wind, sunlight, falling water, and nuclear fission. Since most resources are limited, the engineer must concern himself with the continual development of new resources as well as the efficient utilization of existing ones.

The results of engineering activities contribute to the welfare of man by furnishing food, shelter, and comfort; by making work, transportation, and communication easier and safer; and by prolonging life and making it pleasant and satisfying.

HISTORY OF THE ENGINEERING PROFESSION

The first engineer known by name and achievement is Imhotep, builder of the famous stepped pyramid at Ṣaqqārah, near Memphis, probably *c.* 2550 BC. Imhotep's successors—Egyptian, Persian, Greek, and Roman—carried civil engineering to remarkable heights on the basis of empirical methods aided by arithmetic, geometry, and a smattering of physical science. The Pharos (lighthouse) of Alexandria, Solomon's Temple in Jerusalem, the Colosseum in Rome, the Persian and Roman road systems, the Pont du Gard aqueduct in France, and many other large structures, some of which endure to this day, testify to their skill, imagination, and daring. Of many treatises written by them, one in particular survives to provide a picture of engineering education and practice in classical times: Vitruvius' *De architectura*, published in Rome in the first century after Christ, a ten-volume work covering building materials, construction methods, hydraulics, measurement, and town planning.

Medieval European engineers, like the engineers of the classic world, combined military and civil skills and in the realm of construction carried technique, in the form of the Gothic arch and flying buttress, to a height unknown to the Romans. The sketchbook of Villard de Honnecourt, one of the best known of the Gothic engineers, who lived during the first half of the 13th century, reveals the wide knowledge of the professional engineer in mathematics, geometry, natural and physical science, and draftsmanship.

In the Far East, in India, China, Japan and other regions, engineering had a separate but very similar development, with more and more sophisticated techniques of construction, hydraulics, and metallurgy helping to create advanced civilizations such as the Mongol Empire, whose large, beautiful cities impressed Marco Polo in the 13th century.

Civil engineering

The term civil engineer first came into general use in the 18th century, which also saw the first engineering school founded (1747). This was the École Nationale des Ponts et Chaussées, or National School of Bridges and Highways, which grew out of the Corps des Ponts et Chaussées, or National Road and Highway Commission, of France. Its graduates were imbued with the new taste for the rationalization of knowledge, a spirit that led to research and formulation of theories on such subjects as fluid pressure. John Smeaton (1724–1792), the British designer of Eddystone Lighthouse and other notable structures, was apparently the first to call himself a civil engineer and thereby sever his profession from that of the military engineer. Thomas Telford (1757–1834), Scottish builder of the Menai Straits Bridge, accepted the presidency of the Institution of Civil Engineers, obtaining a royal charter in 1828 that made it the world's first engineering society.

Civil engineers of the 19th-century world built structures of all kinds, designed water-supply and sanitation systems, laid out railroad and highway networks, and planned cities. England and Scotland were the birthplace of the second branch of the modern profession, mechanical engineering (*q.v.*), as a derivation of the inventions of the Scottish engineer James Watt and the textile machinists of the Industrial Revolution. The development of the British machine-tool industry gave tremendous impetus to the study of mechanical engineering both in Britain and abroad.

The slow growth of knowledge of electricity, from Alessandro Volta's original electric cell of 1800 through the experiments of Michael Faraday and others, culminating in the Gramme dynamo and electric motor in 1872 (named after the Belgian Z.T. Gramme), created a branch of the profession of enormous influence: electrical and electronics engineering (*q.v.*). The electronics aspect came into prominence through the researches of such scientists as James Clerk Maxwell of Britain and Heinrich Hertz of Germany in the latter 19th century, the development of the vacuum tube by Lee De Forest of the U.S. in the early 20th century, and the invention of the transistor in the mid-20th century. In the early 1970s, electrical and electronics engineers outnumbered all others in the world.

Chemical engineering

Another major branch of the profession, chemical engineering, grew out of the 19th-century proliferation of industrial processes involving chemical reactions in metallurgy, food, textiles, and many other areas. By 1880 the use of chemicals in manufacturing had created a new industry whose function was the mass production of chemicals. The design and operation of the plants of this industry became a function of the new chemical engineer.

Aerospace engineering, an offshoot of mechanical engineering and the development of powered flight and rocketry, and computer engineering, which traces its origins to the 19th-century inventions of Charles Babbage of Great Britain and J.M. Jacquard of France, are among the most important frontier areas of the engineering profession in the 20th century, with many spectacular achievements already recorded and an incalculable promise for the future.

ENGINEERING FUNCTIONS

The branches indicate what the engineer works with; the functions describe what he does. In order of decreasing emphasis on science, the major functions of all engineering branches are the following:

Research. The research engineer seeks new principles and processes by employing mathematical and scientific concepts, experimental techniques, and inductive reasoning.

Development. The development engineer applies the results of research to useful purposes. Ingenious and creative application of new knowledge may result in a working model of a new electrical circuit, a chemical process, or an industrial machine.

Design. In designing a structure or a product, the engineer selects methods, specifies materials, and determines shapes to satisfy technical requirements and to meet performance specifications.

Construction. The construction engineer is responsible for preparing the site, determining procedures that will economically and safely yield the desired quality, directing the placement of materials, and organizing the personnel and equipment.

Production. Plant layout and equipment selection, with consideration of human and economic factors, is the responsibility of the production engineer. He chooses processes and tools, integrates the flow of materials and components, and provides for testing and inspection.

Operation. The operating engineer controls machines, plants, and organizations providing power, transportation, and communication. He determines procedures and supervises personnel to obtain reliable and economic operation of complex equipment.

Management and other functions. In some countries and industries, engineers analyze customers' requirements, recommend units to satisfy needs economically, and resolve related problems. In some industries, too, engineers decide how assets are to be used.

ENGINEERING EDUCATION

The future engineer must study sciences and their application, resources and their conversion, and man and his needs. The first stage of professional preparation is usually a college degree, but as in other modern professions where rapid change is the rule, learning must be continuous.

Undergraduate curriculum. The integrated undergraduate engineering program, developed primarily in Europe and the U.S. over the past century, provides the basic education for entrance into the profession. It provides

Development of skills

opportunities for students to master important concepts; to become adept in powerful techniques; to develop creative approaches to problem solving; and to become skillful in oral, written, graphic, and mathematical communication. By integrating the study of humanities, social sciences, mathematics, physical sciences, and technology and by providing experience in analysis, synthesis, and experimentation, the undergraduate engineering program offers a modern liberal education.

The typical curriculum leading to the bachelor's degree is defined by a series of courses. General education is provided by a pattern of courses selected from the humanities and social sciences. Basic science education includes mathematics, physics, computer programming, and often statistics, chemistry, and biology. Engineering sciences include materials, mechanics, thermodynamics, transfer and rate processes, electrical science, and information processing.

The combination of general education with basic and engineering science may provide the common preparations for all engineering students at a given school. Specialization to meet the needs and interests of individuals is offered through majors or options—combinations of courses to provide depth in a specific branch or function.

Graduate study. More and more engineers in the U.S., U.S.S.R., and other countries hold advanced degrees. In engineering, graduate study is characterized by highly sophisticated concepts, more advanced mathematics, greater depth in special topics, more opportunity for individual work, and less dependence on formal courses.

Modern trends. Early engineers were trained by apprenticeship to a skilled practitioner. As the body of knowledge gained by observation, testing, and research increased, organized programs of study were initiated in established schools. Following World War II, new trends became apparent, the result of the rapidly growing body of theoretical knowledge. Greater emphasis was placed on mathematics, physical science, and the engineering sciences. The current trend is toward more basic and less applied courses, greater use of computers and probability theory, and, at least in some countries, additional humanities and social sciences.

ENGINEERING AS A PROFESSION

The organization of the engineering profession in most countries reflects the pattern of branching development that resulted in diverse technical societies. This fragmentation is evidenced by the fact that no national or international organization represents all the members of the profession.

Technical societies. The first engineering society, the British Institution of Civil Engineers, reflected the new importance of the civil engineer in the early 19th century. As new fields opened up and new discoveries created the need for specialists, organizations were formed to facilitate the sharing of ideas. In the developed countries, there are hundreds of technical societies devoted to serving such special interests. Nearly every one publishes a journal containing accounts of the work of its members and conducts meetings to keep members informed of new developments. In addition, the major technical societies

Professional status of the engineer

are concerned with the professional status of the engineer and his education, often defining the first and assisting in formulating the second. One international engineering society is the Institute of Electrical and Electronics Engineers (IEEE), New York. Its 160,000 members belong to geographical units in the U.S., Canada, Europe, the Near East, North Africa, and Latin America.

Registration. In some countries, society membership includes the license to practice in that branch. In others registration by a government body ensures that those "professing" a knowledge of science and skill in the art of engineering are competent to practice. The requirements usually include being of good moral character; having practiced in a responsible capacity for a stipulated period; and passing an examination of qualifications to plan, design, or supervise the construction of engineering projects.

Recognition of the increasing role of engineering in government decision making led to the creation (1964) of the National Academy of Engineering (NAE), Washington, D.C., a private organization chartered by Congress. A major objective of the NAE is to identify the changing needs of society, assess the technical resources available, and sponsor programs to meet these needs.

ENGINEERING RESPONSIBILITIES

One activity common to all engineering work is problem solution. The problem may involve quantitative or qualitative factors; it may be physical or economic; it may require abstract mathematics or common sense. Of great importance is the process of creative synthesis or design, putting ideas together to create a new and optimum solution of the problem. Since the engineer functions at the sociotechnological "interface" (with science and technology on one side and individuals and communities on the other), he bears a unique responsibility to decide on priorities, establish performance criteria, select materials and processes, and specify evaluation procedures.

Problem solution. Although engineering problems vary greatly in scope and complexity, the same general approach is applicable. First comes an analysis of the overall situation and a preliminary decision on a plan of attack. In line with this plan, the usually broad and vague problem is reduced to a more categorical question that can be clearly stated. The stated question is then answered by deductive reasoning from known principles or by creative synthesis, as in a new design. The answer or design is always checked for accuracy and adequacy. Finally, the results for the simplified problem as stated are interpreted in terms of the original problem and reported in an appropriate form.

In his search for solutions to problems, especially new problems, the engineer is in conflict with a rather intractable environment and often in competition with predecessors who tried and failed and with contemporary rivals who are trying to solve the same problem. His success depends on his ability to create a new idea, a new device, a new process, or a new material.

Decision making. The engineer not only makes technological advances available to man but also may be charged with the responsibility to see that such advances do indeed enhance the welfare of man. By virtue of his knowledge, his skill, and his unique role in society, the engineer must concern himself with the results of technological progress such as its effect on the physical environment and its social and economic impact. The successful engineer of the future will be prepared for decision making on complex problems in broad areas. He will be skilled in the use of sophisticated tools and creative in the development of new techniques. He will have the vision to conceive vast projects, the talent to analyze them as integrated man–machine–environment systems, and the ability to predict their technical performance and their human impact.

Social implications

BIBLIOGRAPHY. Works on the history of the engineering profession, with discussions of the evolution of its functions, include: A.P.M. FLEMING and H.J.S. BROCKLEHURST, *A History of Engineering* (1925); R.S. KIRBY *et al.*, *Engineering in History* (1956); J.K. FINCH, *The Story of Engineering* (1960); B.V. BOWDEN (ed.), *Faster than Thought: A Symposium on Digital Computing Machines* (1953); and S. THOMAS, *Computers: Their History, Present Applications, and Future* (1965). Contemporary descriptions may be found in R.J. SMITH, *Engineering as a Career*, 3rd ed. (1969); and T.J. HOOVER and J.C.L. FISH, *The Engineering Profession*, 2nd ed. (1950).

(R.J.Sm.)

Engineering, Military

Military engineering is the oldest of the engineering skills, and military engineers were the first scientific soldiers, out of whose experience the profession of civil engineering came into being. Since then, the arts of military and of civil engineering have developed side by side.

Today the wartime function of military engineers is to apply at all stages engineering knowledge and resources to the furtherance of their commander's plan. Their role can conveniently be divided into three main tasks: (1)

combat engineering, or tactical engineer support on the battlefield; (2) strategic support by the execution of works services required in the communications zones, such as the construction of airfields, depots, the improvement of ports and road and rail communications, and the storage and distribution of fuels; and (3) ancillary support such as the provision and distribution of maps and the disposal of unexploded bombs and warheads.

The military engineers also carry out in peacetime a wide variety of civil works programs. Some are employed on vast projects dealing with coastal engineering and flood control, improving navigational capabilities of major rivers, development of hydroelectric power, water supply, provision of recreational amenities, and pollution control. Military engineers are also employed in providing aid to the civil community, particularly in disaster relief following earthquakes, hurricanes, and severe flooding.

The earliest military engineers. Evidence of the work of the earliest military engineers can be found in many parts of the world; for example, in Europe the hill forts constructed by the men of the late Iron Age, in the Middle East the massive fortresses incorporating the remarkably sophisticated military ideas of the Medes and Persians, and in Africa the high dressed-stone wall fortifications protecting the holy city of Zimbabwe in Rhodesia, built by the military engineers of a lost civilization. The greatest ancient defensive work ever built by human hands is probably the Great Wall of China, started in the 3rd century BC to defend the 1,000-mile frontier of the Celestial Empire from its barbarian neighbours. Another epic feat of ancient military engineering is known only through the written record of Herodotus, the Greek historian of the Persian wars of the 5th century BC. According to Herodotus, the Persian king Xerxes ordered his engineers to construct a pontoon bridge across the Hellespont (modern Dardanelles), which was accomplished by a mile-long chain of boats, 676 in all, arranged in two parallel rows.

Protection of garrison towns

The Romans. Examples of Roman military engineering works can still be seen throughout Europe and the Near East. Their *castra*, or military garrison towns, were protected by ramparts and ditches and interconnected by straight military roads along which their legions could speedily march to wherever they were needed. Like the Chinese, the Roman military engineers also built walls to protect the limits of their empire, though not of the same size and extent. A famous example is Hadrian's Wall in Britain, 73 miles (117 kilometres) long, from the Solway Firth to the North Sea just north of Newcastle-upon-Tyne, which protected the northern frontier of Roman Britain from the turbulent Picts and Scots. Throughout the long Punic Wars (264–146 BC), when Rome and Carthage struggled for supremacy at sea, harbours and defended naval bases were built, along with a system of watchtowers and forts, to guard against sea raiders. Roman military engineers thus became skilled in the construction of seaward defenses, harbours, and lighthouses and other navigational aids, tasks that remained the military engineer's province for many centuries. Roman engineers also developed a battering ram, slung from a beam mounted on a wheeled platform and provided with overhead cover, to be used against the wall or gate of an enemy fort. They also developed such engines as the giant catapult and the ballista, a large crossbow whose bowstring was drawn back by a windlass. These devices were used to hurl heavy rocks, showers of stones, or firebrands at defended places or at troops in the open. Still another impressive area of achievement by Roman military engineers was bridging; Caesar gives a detailed account of how his soldiers constructed a wooden lashed-trestle bridge over the Rhine, a feat that would tax the ingenuity of a modern military engineer.

The Middle Ages. The arts of fortifying and of attacking fortified places reached new heights in the centuries following the fall of the Roman Empire, partly as a result of the growth of cities in the later Middle Ages. Medieval engineers, both Arab and European, became proficient at mining operations, by which tunnels were driven under walls of castles or cities and their timbering set fire, causing the masonry overhead to collapse.

Prefabricated forts in 1066

William the Conqueror brought a novel wooden fort, or donjon, from Normandy to England in prefabricated sections to erect on the Hastings beachhead in 1066. As he advanced through the country William built a great number of such temporary forts. The method used was to raise a high mound of earth, called a motte, the top of which was flattened and surrounded by a wooden palisade. Within this palisade a tall wooden tower, or donjon, was built. Timber was used because it lent itself to rapid construction and was readily available locally. Furthermore, the unconsolidated made-up earth of the motte would not have supported the weight of a stone tower. Around the motte an outer defense was provided, called the bailey, consisting of a ditch the soil from which was thrown inward to form a bank that was surmounted by a wooden palisade. These temporary forts were abandoned as castles of permanent construction were built to replace them; the motte and bailey design was retained, however. The donjon, later known as the keep, was often a great tower made of flint, bound together by cement, with walls 20 feet (six metres) thick at the base. The outer defenses of the bailey were equally massive. These castles were an emblem of conquest and became the most effective instruments for enforcing law and order and the most important military and administrative centres in the land. A system of military roads, following the Roman pattern, interconnected them.

Early in the 12th century the crusaders, having captured Jerusalem and retaken the Holy Places, were faced with the problem of defending their conquests with minimum garrisons. Their solution was to build great fortresses, the most impregnable of their day, in strategic locations chosen mostly by French military engineers. As the increasing power of the Muslims gradually forced the knights out of the Holy Land, they withdrew first to Cyprus, then to Rhodes, and finally to Malta. Employing Italian and Spanish engineers, they constructed a series of fortifications on Malta and defenses covering dockyard creeks essential for the maintenance of their ships. In 1565 Süleiman the Magnificent sent a vast naval and military force against Malta. The Turkish forces landed and overran some of the outer defenses of the island, but the remaining works withstood all attacks, and Süleiman's expedition ignominiously withdrew.

The age of firearms. The Renaissance produced a succession of outstanding military engineers and men of remarkable genius, the most famous being Leonardo da Vinci. The age also witnessed rapid development of the cannon and all forms of firearms, a development that brought about a reappraisal of the design of fortifications and the techniques of siege warfare. Leonardo brought his fertile and inventive brain to bear on all facets of military engineering. He carefully analyzed the elementary military engineering tasks of using pick and shovel and basket and barrow for removing excavated soil. He developed the art of casting cannons and was expert in the use of gunpowder, both as a propellant and as a demolition explosive. He designed cranes, earth-moving equipment, and other devices of military value. Leonardo also invented types of assault boats and bridges and several siege devices. A gifted hydraulic engineer, he was a forerunner of the British military engineers who later built vast irrigation projects in India and the 20th-century U.S. Army Corps of Engineers, which is responsible for flood control and works connected with the rivers, harbours, and other waterways of the United States.

The frequent campaigns fought in the Low Countries during the 17th century largely involved besieging and defending the great barrier fortresses guarding the frontier between France and the Netherlands. The twin techniques of fortification and siege warfare were exclusively a matter for the military engineer. By midcentury these techniques reached such a state of perfection that they did not change until the latter half of the 19th century, when breach-loading artillery and the high explosive shell called for drastic alterations in the design and construction of defenses.

The outstanding military engineer of this era was Sébastien Le Prestre de Vauban of France (1633–1707). Although he is credited with having said "You do not fortify by systems but by good sense and experience," the design of his fortresses was often rigidly stylized and elaborate. The lines, or traces, of his fortifications, moreover, designed so that direct or flanking fire from both gun and musket could be brought to bear on all approaches, were slavishly copied by succeeding generations of military engineers. Indeed, Vauban's bastion trace became the accepted trace for the fire trench of World War I (see FORTIFICATIONS).

So lasting was Vauban's influence that the French school of military engineering was held to be the leading authority on the profession for many generations, and when the Army Corps of Engineers was constituted in 1779 during the U.S. War of Independence, George Washington turned to France to provide him with military engineer officers.

French Revolutionary and Napoleonic Wars, 1793–1815. In these wars, armies grew to unprecedented sizes. Fighting spread from Portugal to Moscow and further afield to the Americas, Egypt, South Africa, and India. In the many campaigns, siege operations were carried out by the traditional Vauban methods. When the French threatened England, vulnerable points on the southern coast of England were defended by hastily erected Martello towers, built on the Norman motte and bailey principle. The outer walls, however, consisted of masonry bastions and the thick-walled stone tower was armed with cannon. The Martello tower construction was later used for coast defense in both Canada and the U.S.

The Martello tower

An entirely new system of field defenses was adopted during the construction of the Lines of Torres Vedras by the Duke of Wellington during the Peninsular War in Portugal. Artillery fire could be brought down from a series of mutually supporting redoubts of low silhouette sited in great depth on dominant features and covering all approaches. Fields of fire for the guns were cleared of all obstructions, and streams were dammed to produce inundations to channel enemy lines of approach. Each redoubt was defended by a small close-support garrison of Portuguese militia. Wellington's field army was left available for manoeuvre and counterattack. Though Wellington had less than half as many men as the French, the latter could not risk assaulting these defenses and had to withdraw.

19th-century military engineering. The long period of peace that followed the Napoleonic Wars witnessed another great step forward in the fields of science and engineering and a rapid opening of new territories. The military engineers were quick to study how these technological advances could be put to a military use. In the new territories, they often found themselves the only professional engineers in the land and took personal responsibility for their exploration, mapping, defense, and development.

The U.S. Army Corps of Engineers led the way in exploring much of the American West, breaking trail for the immigrants who followed. They explored, surveyed, and mapped the land, built forts and roads, and later assisted in building the transcontinental railways. They also improved harbours and inland waterways and constructed dams and levees. They were responsible for maintaining coastal defenses. Similarly in Canada, India, Australia, New Zealand and other countries of the British Empire, the surveys and many of the public buildings and utilities, roads, bridges, railways, telegraphic communication systems, irrigation projects, harbours, and maritime defenses were the work of the British military engineers.

British and French military engineers first used the electric telegraph in the Crimean War (1853–56). Forward headquarters were connected by land line to their respective commanders in chief, who in turn were connected by means of a 340-mile (550-kilometre) submarine cable, laid by the Royal Engineers, to an existing transcontinental line to their governments in London and Paris. A new era in the control of war had dawned; among other political effects, the public at home received prompt news reports from correspondents at the front.

Submarine mining and searchlights. During the U.S. Civil War (1861–65), some 40 ships were sunk or disabled by mines or torpedoes; and interest in this form of warfare was intensified. Engineer submarine mining units were formed to lay electrically controlled mines in time of war to protect naval bases and important commercial ports. The Brennan torpedo, an underwater wire-controlled guided missile with a range of 1,000 yards (900 metres), was also developed. Submarine mining later became a naval responsibility, and the military engineers then developed defense electric lights to illuminate targets by night for coastal batteries. Subsequently they developed anti-aircraft searchlights and sound locators for air defense. Though infrared light and radar made searchlights obsolete in their traditional role, they are still used for battlefield illumination by military engineers.

The Brennan torpedo

For many hundred years military engineers used gunpowder and firing accessories obtained from the artillery when they had to carry out demolitions. In the 1860s, however, guncotton with detonating fuse was introduced as a demolition explosive. World War I produced trinitrotoluene (TNT), which was used to fill high-explosive shells and, later, shaped demolition charges that produced a molten jet of metal to penetrate the target. The search for new high explosives during the interwar years produced cyclonite (RDX), which probably represents the limit of what is chemically practicable for a service explosive. In 1942 plastic explosive (PE) was introduced. It made explosives quicker and easier to place and was more resistant to weather and battle hazards. Ammonal, common in World War I, was retained in World War II for mined charges. Pellets of RDX–wax–aluminum and ammonium nitrate and fuel oil slurries have since been used in mined charges with success (see EXPLOSIVES).

Early military aviation. Early in the 1860s French, British, and U.S. military engineers adapted balloons for a variety of military purposes, mostly intelligence. Subsequently, man-carrying kites were employed for artillery spotting when the wind was too strong for captive balloons. Later, powered airships and airplanes were brought into service. Aerial navigational aids, ground-to-air wireless communication, and aerial photography followed. The pioneering days of military flying were over, and shortly before World War I military engineers of most armies handed over their responsibility for aeronautics to other organizations.

Transportation and movement control. Since the introduction of railways, military engineers have been responsible in theatres of war for the construction and maintenance of the permanent way and bridges of railway systems, the control of the rail movement of troops and military matériel, and the repair of locomotives and rolling stock. They have operated armoured trains and have been responsible for the rail mounting of superheavy guns. Military-operated railways played a constantly growing role in wars of the 19th and 20th centuries as logistical requirements grew.

Road mechanical transport and power tools. To increase mobility and work capacity, steam traction engines, capable of towing a train of trailers, were introduced into the service by military engineers in the 1860s. These were the forerunners of military road mechanical transport. Employed as static machines these traction engines could, by means of certain attachments, be used as cranes or as prime movers to operate a variety of power tools. From this beginning the vehicle-mounted air compressor and its associated pneumatic tools later developed. Many of these pneumatic devices in their turn were replaced by electric power tools driven by lightweight generators. Thus over the years the source of power has become smaller, while the output from individual tools for use in the field has simultaneously increased.

Developments during World War I. Protracted trench warfare in World War I called upon all of the traditional siegecraft skills of the military engineers. Mining

and countermining were carried out on a scale never before attempted. The greatest achievement was the firing in June 1917 by British sappers of over 1,000,000 pounds (450,000 kilograms) of explosive, placed in 16 chambers 100 feet (30 metres) deep, which completely obliterated Messines Ridge in Belgium and inflicted 20,000 German casualties.

Trench tramways and light railways were built for the maintenance of forward troops. Large camouflage projects were carried out to screen gun positions, storage dumps, and troop movements from enemy observation. Forestry and quarrying units were raised and production workshops established, for provision of trench-fortification matériel and pit props and to meet the requirements of the vast construction projects carried out in rear areas.

The scope of military signalling increased enormously and reached such a size and complexity that, when World War I ended, military telecommunication engineers became a separate corps in all armies. New techniques were developed for fixing enemy gun positions. Map making by use of aerial photographs (photogrammetry) developed. Field printing presses were set up to provide vast quantities of maps of the fighting areas, and a grid system was introduced for maps covering the whole theatre of operations (for a full discussion, see MAPS AND MAPPING). German, British, and French military engineers also introduced means of using toxic gas and flame throwers, and tanks as weapons of war were developed by the British and French.

Photogrammetry in map making

In the Middle East, British and Indian engineers carried out several major river-crossing operations in Mesopotamia (Iraq) and operated craft on the Tigris and Euphrates rivers, which were used as the main supply routes for the campaign. In the Sinai Peninsula of Egypt and in Palestine the provision of water for the horse cavalry divisions was a major engineering task. The Royal Engineers piped Nile water from Egypt across the Sinai Desert to Jerusalem.

Interwar years. Defense was the major concern of most European governments during the interwar years. French military engineers designed and constructed the Maginot Line, a most sophisticated and assumed impregnable defensive system protecting France's common frontier with Germany and Luxembourg. The Germans constructed their Siegfried Line covering the same frontiers, and the Czechs built a defensive system to protect their northern frontier. The Russians constructed a looser system covering their long western frontier. For the air defense of Great Britain anti-aircraft batteries and searchlight positions were built to defend towns and service installations, and for passive air defense deep, gasproof air raid shelters were constructed in large numbers, and attempts were made to camouflage vulnerable points from aerial observation.

Developments during World War II. Military engineers of World War II faced and mastered problems on a scale and often of a character previously not experienced. Because of the importance of air power, airfields and airstrips had to be built, usually in great haste, frequently under fire. Amphibious operations, involving the landing of troops on a hostile shore, created a host of engineering problems, from underwater demolition of obstacles to rapid construction of open-beach dock facilities, such as the prefabricated mulberry harbour to maintain the Normandy landings in 1944. To meet the logistical needs of huge armies operating in remote regions, enormous base facilities were constructed.

The air battle. The construction and maintenance of forward airfields was a major problem for the engineers of all armies, especially the British and U.S. Allied airfields could never have been built without large quantities of earthmoving machinery mostly of American origin. For the construction of forward airfields it was often necessary to airlift light items of plant and later to lift sections of heavier items that could be reassembled on the ground. Great advances were made in the science of soil stabilization, and a variety of rapid airfield surfacing materials were developed, the American pierced steel plank being the most successful. For combined operations in the Pacific, floating airfields, consisting of prefabricated sections bolted together, were developed. Methods of piping aviation and other fuel from seagoing tankers, and its storage ashore in tank farms, were brought to a high state of perfection. After the Allied landing in Normandy in 1944, a pipeline was laid on the bed of the English Channel to supply fuel to the Allied forces in northwestern Europe.

Development of floating airfields

Armoured warfare. The most common devices used by World War II military engineers to stop enemy armour were the antitank ditch and the antitank mine. The ditch was only effective in locations where it could not be bypassed and where it could be kept under constant observation and fire. The antitank mine was first widely used during the fighting in the Western Desert of Egypt, where the minefield was the most easily provided obstacle to the movement of armoured forces. Special mine-warfare drills, techniques, and equipment were developed by the British and German military engineers for minelaying, mine location, and breaching minefields. As an adjunct to the antitank mine, antilifting devices, antipersonnel mines, and booby traps were introduced.

Bridging. Most armies at the outbreak of World War II possessed bridging equipment that would carry a tank weighing up to 30 tons. By 1940, however, tanks of 40-ton weight were being built, and it was known that this would soon not be the upper limit. To carry these increased weights, the Bailey Bridge was developed by the Royal Engineers in Britain. Its pin-connected rectangular truss panels made it easy to mass-produce. It also provided great flexibility on span and load capabilities. It could be used as a fixed span bridge or, with pontoons, as a raft or floating bridge. It was simple to construct quickly by day or at night. The equipment was produced in large quantities in Great Britain, the United States, and in Italy and was a vital factor in preserving the mobility of Allied armies in all theatres of operations, thus making a substantial contribution to ultimate victory. On the eastern front Soviet engineers developed a type of bridge with a deck just below the surface of the water, making it difficult to spot from the air.

Military roads. One of the most extraordinary feats of military engineering during World War II was the building in 1944 by Allied forces of a supply road from Ledo, India, to the Burma road at a point where the road was still in Chinese hands. This Stilwell, or Ledo road, as it was called, opened in January 1945, was 478 miles (770 kilometres) long, twisting through mountains, swamps, and jungles.

American amphibious operations were the main feature of the war in the Pacific. Amphibious operations in the European theatre culminated in the Normandy landings, the greatest combined operation in military history.

Although the engineering tasks varied in size in these operations, they were basically similar. They included the removal of beach obstacles and the neutralization of beach mines, the construction of beach roadways, the clearance of beach exits and the liquidation of fixed defenses, the construction of roads and administrative areas on the beachhead, and the construction of forward landing grounds and air-dropping zones. Special equipment, including armoured engineer equipment that had to be capable of wading ashore from landing craft, was developed for these tasks.

Great advances were made in air survey due largely to the use of American equipment and calculating machines to speed up survey computations. Methods of map and air chart reproduction also became more sophisticated, and the requirement for military maps for all theatres of operations reached astronomical proportions. All of the camouflage techniques developed during World War I were employed, and, in addition, inflatable dummy tanks, vehicles, and landing craft were produced, mostly in the U.S., and used extensively in deception plans.

Despite the many novel tasks imposed by World War II, traditional engineering jobs played a substantial role. Armies on the defensive constructed fortifications and carried out defensive demolitions on a scale never before seen. Outstanding were the German fortifications along

the Atlantic coast ("Fortress Europe") as well as the Siegfried Line in Germany. Concrete pillboxes played a major role, in addition to a variety of obstacles, many underwater; mines, both land and underwater, were laid by the millions. These vast works were constructed by the German Todt organization, which also specialized in demolition as German armies slowly retreated through the Soviet Union, Italy, and France.

The Manhattan Project. But the largest task carried out by World War II engineers, and indeed by any military engineering establishment in history, was undoubtedly the Manhattan Project, which produced the atomic bombs dropped on Hiroshima and Nagasaki. Civilian scientists as well as engineers were recruited in large numbers for this mammoth project. Its success made it a model for later large-scale engineering efforts involving many scientists and engineers from different disciplines. In a sense, it brought military engineering full circle; whereas early military engineers were the first all-round scientific soldiers, the modern military engineering project must enlist a team bringing together experts in all branches of engineering and scientific knowledge.

Engineers since World War II. In the late 1950s the U.S. Corps of Engineers constructed a network of radar warning systems across Canada and a series of Nike batteries to defend vulnerable points against the possibility of an attack across the polar ice cap by jet aircraft carrying nuclear bombs. In the 1960s they constructed a series of intercontinental ballistic missile bases. Introduction of these intercontinental missiles caused military surveyors to take a closer interest in the shape and size of the world. They developed ways of linking missile bases with far distant targets using geodetic satellites. Modern weapon systems, using advanced tracking, navigation, and locating devices, require mapping information direct from a computer.

Construction of air bases

The military confrontation of the Cold War and the conflicts in Korea and Vietnam resulted in construction of strategically placed air bases in Europe, the countries bordering the Arctic, Africa, and the Far East. In the varying emergencies military engineers performed not only their traditional combat and support roles but also, in addition, often carried out important peaceful works, such as dams and power stations to assist underdeveloped countries.

In the foreseeable future military engineers will probably continue to carry out their present civil-engineering tasks, such as river control. Their skills will be required in disaster areas stricken by earthquake, flood, or hurricane. In confrontations and insurgencies the military engineers will continue to combine traditional combat engineering with civil tasks. They will, of course, be equipped and trained for war in which nuclear weapons may be used. They will also undoubtedly play a part in the exploitation of the seabed and the continued exploration of the moon and deep space.

Education. Military engineering is not itself a discipline but rather the application of civil and other engineering disciplines to military problems. The military engineer must strive in his profession as a soldier and study the engineering lessons of past campaigns; and he must also keep abreast of developments in the engineering disciplines in civil life and, where possible, apply them to military purposes. The British Royal Military Academy was founded in 1741 at Woolwich to train cadets destined for commissions in the engineers and artillery, but it was not until 1812 that a school was set up at Chatham for the instruction of junior officers and soldiers of the engineers in fieldworks and siegecraft. It is now the Royal School of Military Engineering (RSME) and instructs in all branches of military and civil engineering undertaken by the Royal Engineers. The first school founded specifically to train military engineers was established at Mézières in France in 1748. In 1795 it moved to Metz where it was renamed the École Polytechnique. After World War II, the École l'Application du Génie was established at Angers and in 1946 the École Supérieure Technique du Génie was opened for the higher education of engineer officers. Claude Crozet, a graduate of the École Polytechnique, joined the staff of the United States Military Academy at West Point in 1817, becoming the first professor of civil engineering in the United States. Engineering continued to be stressed at West Point, and for many years every commandant was an engineer. Schools of military engineering, organized on the lines of the British RSME, exist in Canada, Australia, India, and Pakistan.

BIBLIOGRAPHY. B.D. COLL, J.E. KEITH, and H.H. ROSENTHAL, *The Corps of Engineers: Troops and Equipment* (1958), a description of the organization and deployment of the U.S. Corps of Engineers and of engineer equipment in service; U.S. ARMY, ARMY FORCES, PACIFIC, OFFICE OF THE CHIEF ENGINEER, *Engineers of the Southwest Pacific*, 8 vol. (1947–51), vol. 1–4 deal with the period 1947–49, and vol. 5–8 with the period 1949–51; INSTITUTION OF ROYAL ENGINEERS, *The History of the Corps of Royal Engineers*, 9 vol. (1889–1958), a history of the British military engineers from Norman times to 1948; E.W.C. SANDES, *The Royal Engineers in Egypt and the Sudan* (1937), covering the period from 1800 to 1936; A.J. KERRY and W.A. MCDILL, *History of the Corps of Royal Canadian Engineers*, 2 vol. (1962–66), a history of the Royal Corps of Engineers from 1749 to 1946; R.J. UNSTEAD, *Castles* (U.S. title, *British Castles;* 1970), a description of the development of castles and permanent fortifications from the earliest times to the 16th century; P. WARNER, *Sieges of the Middle Ages* (1968), an account of the major sieges in Europe and Asia Minor between the years 1000 and 1400; VEZIO MELEGARI, *Great Sieges* (1970), a description of historically decisive sieges from the legendary siege of Troy to the siege of Stalingrad during World War II; B. GILLE, *The Renaissance Engineers* (1966), a study of the inventions and techniques of Leonardo da Vinci and other military engineers of the Renaissance.

(J.H.S.L.)

England

England is the predominant part of the United Kingdom, yet constitutionally it does not exist. It is not mentioned in the title of its sovereign. Elizabeth II is queen of "the United Kingdom of Great Britain and Northern Ireland and Her other Realms and Territories." Scotland and Wales have their own departments of state and Cabinet ministers, while Northern Ireland has its own Parliament and government for domestic affairs. But England has no such special rights, not even a separate set of official statistics, which are conventionally compiled either for the United Kingdom, in which the administration embraces the whole country—as in overseas trade, taxation, and defense—or for England and Wales, as in the administration of the law, education, and various social services (where Scotland exercises an independent authority). It is rare for institutions to operate for England alone. A notable exception is the Church of England, the Anglican Communion having separate churches in Wales and Scotland. Otherwise, England retains its separate identity only in such activities as sport. Cricket is strictly a matter for England and English counties (apart from the Glamorgan County Cricket Club in Cardiff, South Glamorgan, Wales), and England fields its own teams in rugby and soccer football. But in its institutions, England gives the appearance of having been swallowed up in the larger mass of the United Kingdom since the Act of Union of 1707. It sometimes seems that Scotland and Wales have been more successful in retaining or securing their own special institutions. Thus, the British Broadcasting Corporation (BBC) has set up national broadcasting councils for Scotland and Wales but none for England, and there are other such examples of apparently preferential treatment.

England's dominance in the United Kingdom, nevertheless, is unquestionable. Occupying a little more than half of the landmass, England has more than four-fifths of the population, estimated in 1974 to be 46,436,000 out of a total of 55,968,000, for the U.K. as a whole. The bulk of the most fertile lowlands are in England, and it enjoys a more agreeable climate, warmer and drier than the rest of the kingdom. Six of the U.K.'s seven major metropolitan areas, or conurbations, are in England, which also has a higher proportion of wealth and natural resources than the rest of the U.K. Population and wealth increase

with proximity to London, the national capital. London, in turn, with its ring of densely populated home counties in an area up to 50 or 60 miles in radius, dominates England, as England dominates Great Britain (England, Scotland, and Wales). There are good historical grounds for this being so. London was the biggest town in Roman Britain and has been the capital of a unified England since the Norman Conquest of 1066. Roman control of England fanned out with a network of roads based on London—a radial pattern that persists today. The Romans extended their rule approximately to the Welsh and Scottish borders of today, driving the Brythons back into the far southwest and the mountains of Wales and holding off the Scottish clans in the north. The frontiers of Roman rule in Britain were close to those of England today. In the subsequent 1,500 years, England has retained its frontiers short of most of upland Britain. The hill and mountain people of Wales and Scotland retain their sense of identity, and the English theirs, though they share a common language (although Welsh is still the common tongue in many rural areas of the principality) and a single sovereignty.

Extent of Roman rule

The "Englishness" of England is a quality hard to define, yet there is no doubt that it exists. Most institutions are British, but poets have very rarely saluted Britain (Britannia who rules the waves in the patriotic song with words by a Scottish poet being the exception). But poets by the score have celebrated England and the English, from William Langland in the medieval English of the 14th century onward. Shakespeare is self-consciously English and for England. The art historian Nikolaus Pevsner has written of "The Englishness of English Art." Yet this is not an insular character. It has been continually enriched from abroad. The English of Anglo-Saxon England were able to absorb the Danish and Norse invaders and their Norman French conquerors. And when the English went on to conquer first the remainder of the British Isles and then a worldwide empire, they did so without the dilution or dispersal of their essential character. The dominant role of the English is recognized implicitly in the numerous histories of Britain and the British Empire that are inaccurately called "A History of England." A remarkable capacity to absorb and transmute alien elements can be seen in English literature, which in the 20th century has been able to take into the mainstream of its tradition poets as Irish as W.B. Yeats, as Welsh as Dylan Thomas, or as securely in the classic line as the U.S. expatriate T.S. Eliot. (For related information, see UNITED KINGDOM; LONDON; MANCHESTER; LIVERPOOL; BIRMINGHAM; THAMES RIVER; ENGLISH CHANNEL; NORTH SEA; BRITAIN AND IRELAND, HISTORY OF.)

THE LAND AND ITS DIVERSITY

Extent and boundaries. One fundamental English characteristic is diversity within a small compass. England occupies about one-thousandth of the world's land area but contains about one-eightieth of the world's population and is, therefore, one of the most densely populated countries in the world. Even the farthest points of England are no more than a day's journey from London. Some approximate road distances from London are: to Dover, for the Channel crossing to France, 72 miles; to Southampton, the port for Atlantic routes, 77 miles; to Land's End, the farthest southwesterly point, 289 miles; to the southwestern port of Bristol, 116 miles; to the northwestern port of Liverpool, 197 miles; to the Scottish border at Carlisle, 300 miles; and to Berwick-upon-Tweed, 337 miles. No place in England is more than 75 miles from the sea.

Geological history. Within this small area, the diversity of England is based on a geological structure of remarkable complexity, spanning 600,000,000 years. Almost every phase of geological history is illustrated in the intricate patterns of the geological map of England, from the oldest igneous rocks, to be found at the western extremities in Cornwall and Cumberland, to the most recent alluvial soils of reclaimed fenland in East Anglia. Between lie bands of sandstones and limestones of different geological periods, witness to primeval times when large parts of central and southern England were submerged below warm seas. In the period of Alpine mountain building, these sedimentary rocks were lifted and folded, so that they now run from southwest to northeast, to yield chains of low hills rising to about 800 feet and fanning out from the Channel coast into the North and South Downs, the Wiltshire and Berkshire downs, and the Cotswolds and the Chilterns, extending through Lincolnshire and Yorkshire to the sea. In more recent geological times these hills were rounded into characteristic plateaus with west-facing escarpments by the great glaciers of three successive Ice Ages, the glaciers extending as far south as the northern outskirts of London. When the ice melted, the landscape was further altered by deep deposits of sand, gravel, and glacial mud.

Hills and downs

Subsequently, the processes of erosion by rain and river and the tides and subsidence in parts of the east of England have further shaped the hills and the coastline. A more substantial chain of hills, the Pennines, with moorland tops rising to between 2,000 and 3,000 feet, split the north of England into northwest and northeast. These hills of limestone, grit stone, and carboniferous strata are associated with major coalfields, some outcropping to the surface, others deeply buried. This geological history underlies the characteristic variety of the English landscape, for in journeys of only a few miles it is possible to pass through a succession of different soil structures: from chalk down to alluvial river valley, from limestone to sandstone and acid heath, from clay to sand, and so on, each type of soil bearing its own special class of vegetation. The geological complexity of England is strikingly illustrated in the cliff structure of its shoreline. Along the south coast from the ancient granite cliffs of Land's End in the extreme southwest are found a succession of sandstones of different colours and of limestones of different ages, culminating in the white chalk from the Isle of Wight to Dover.

These main groups are cut into by coves, bays, and beaches of particular local identity. A constantly changing panorama of cliffs, bays, and river estuaries distinguishes the whole of the English coastline, which, with its many indentations, is some 2,000 miles long.

Man's influence on the landscape. Though the significance of the ever-changing geological structure can hardly be exaggerated, the landscape as it is seen today has been significantly changed by man. Only the remotest moorland and mountain tops have been untouched; even the bleak Pennine moors of the north are crisscrossed by dry stone walls, and their vegetation modified by the cropping of mountain sheep. There is virtually no genuine wilderness left in England. The face of the country has been constantly worked over for centuries, acre by acre, and the marks of its exploitation and use dominate the contemporary landscape. The oldest traces are the antiquarian survivals, such as the Bronze Age forts studding the chalk downs of the southwest, the little rectangular Celtic fields of Cornwall, and the corrugations left by the strip farming of medieval open fields.

Towns and villages

More significant is the structure of town and village, which was established in Roman-British and Anglo-Saxon times and which has persisted as the basic pattern. The English live in scattered groups of high-density occupation, whether in village, or town, or, in modern times, in cities. The effect is to make England even in the 20th century look still a rural country, though 90 percent of its people are urban dwellers and less than 3 percent are engaged in agriculture. Despite a large amount of badly managed urban sprawl, as cities spilled out into conurbations, England, as seen from the air or from any number of hilltop viewpoints, is notable for extensive tracts of farming countryside that lie between its towns, its smaller villages often engulfed in the vegetation of trees, copses, hedgerows and fields: in a phrase of the poet Gerard Manley Hopkins, "the sweet especial rural scene," which occupies so large a place in English literature and English art.

The quality of the English scene is deeply rooted in its geology. Local building materials account for the traditional look of the Cotswold village, built of silvery lime-

stone with stone roofs. The colour of the stone also often identifies the locality. A honey-coloured stone (which weathers badly) was much used in Oxford, and a rusty ironstone is typical in North Oxfordshire and Northamptonshire, along the line of an ironstone belt. The Pennine gritstone adds a special severity to the moorland villages of the north. Timber frame and thatch are the building materials of the river valleys, and excellent clay provides the warm red brick of southern England. In recent years the easy transport of cheap but alien materials is to be blamed for many jarring intrusions into the harmonious towns and villages originally built mainly of local materials.

The visual impact of a mostly green and pleasant land can, however, be seriously misleading. England is primarily an industrial country, built up in the second half of the 18th century, during the first few generations of the Industrial Revolution, by ruthless exploitation of the coalfields and equally ruthless exploitation of cheap labour, especially in the cotton-textile areas of Lancashire, in the woollen-textiles areas of Yorkshire, and in the engineering centres of the Midlands and the Northeast. England has large tracts of dismally ugly derelict areas, scarred by the spoil heaps of the coal mines, by quarries and clay pits, by abandoned industrial plants, and run-down slums. About 100,000 acres have been officially listed as derelict, and the figure is probably an underestimate.

Climate. The diversity of the English landscape is accentuated by its weather. In a temperate maritime zone, it does not run to extremes but is certainly one of the most erratic climates in the world. The averages are moderate enough, ranging in the Thames Valley from 39° to 40° F (about 4° C) for the coldest month and 62° to 64° F (17°–18° C) for the warmest; but the extremes recorded in England go below 0° F (−18° C) and above 90° F (32° C). The Roman historian Tacitus recorded that the climate was "objectionable, with frequent rains and mists, but no extreme cold." Yet the winter of 1962–63 saw most of England heavily snowbound for nearly ten weeks, and snow cover may be expected in the higher parts of England on about 50 days a year. Though it is thought of as a wet country, lowland England has less than 40 inches (1,000 millimetres) of rainfall annually, and it frequently suffers from drought, the rising demand on water supplies becoming an increasing cause of official concern. In parts of Southeast England, the annual rainfall averages as little as 20 inches. Charles II expressed a judicious view when he said that the English climate was the best in the world—"a man can enjoy outdoor exercise in all but five days of the year." But no one would dispute that it is unpredictable: hence Dr. Samuel Johnson's observation that "when two Englishmen meet their first talk is of the weather." This changeability of the weather, not only season by season but day by day and even hour by hour, has had a profound effect on English art and literature.

Rainfall

The English painters John Constable and J.M.W. Turner were precursors of Impressionism in their exploration of the constantly changing light of English skies, and Constable painted cloudscape for its own sake. English poets down the centuries can be found constantly referring to the weather, from William Langland in the 14th century opening his poem *Piers Plowman* with the line, "In a somer seson whan soft was the sonne," to Shakespeare in the 16th century commenting on "the uncertain glory of an April day," and W.H. Auden in the 20th century with his "traffic of magnificent cloud."

Plant and animal life. England shares with the rest of Britain a below-average range of vegetation and living creatures, partly because the island was separated from the mainland of Europe soon after much of it had been swept bare by the last Ice Age (about 10,000 years ago) and partly because the land has been so industriously worked by its dense population. Predatory wild animals, such as the wolf, disappeared centuries ago. More recently birds of prey have suffered at the hands of farmers protecting their stock and their game birds. Under new protective laws, including a law restricting the collecting of birds' eggs, some of the less common birds have been re-establishing themselves. The bird life of England is unusually varied mainly because the country lies along the line of bird migrations and provides winter feeding grounds for some species and summer breeding habitats for others. Some birds have found town gardens, in which they are likely to be fed, a favourable environment, and in London about 100 different species are recorded annually. London has also been found a favourable habitat by foxes, which in small numbers have colonized woods and heaths within a few miles of the city centre. Though the species of plants may not be so great in number as those of the European mainland, they span a wide range and include some rarities. Certain Mediterranean species are to be found in the sheltered and almost subtropical valleys of the Southwest, while vestiges of the Ice Age tundra survive in parts of the moorland of the Northeast. England has been noted for a profusion of summer wildflowers in its fields, lanes, and hedgerows. In some parts these have been severely reduced by the use of herbicides on farms and roadside verges. Cultivated gardens account for much of the varied vegetation of the country, being enriched by the importation of many species of trees, shrubs, and flowering plants from all over the world.

Conservation efforts. The landscape for which England is famed has been seriously threatened by rising population, increasing urbanization, and increased wealth that has led to greater personal mobility. More and more townspeople seek recreation in the countryside and, by sheer numbers, are destroying what they look for. This dilemma has come to be known as the "leisure explosion." Not only England is affected, of course, but it was in England that these pressures first led to active intervention, to start with by individual action and later by action by public authorities, to set up a system of institutions and law to preserve the heritage of the past. One of the earliest of such initiatives was the establishment of the National Trust, a private organization dedicated to the permanent preservation of beautiful countryside and fine buildings. With a membership of over 250,000, the National Trust has become the largest single private landowner in the country with holdings covering more than 600 square miles. Its properties, which include some notable beauty spots, are visited by 2,000,000 to 3,000,000 people annually, and it has launched a special campaign to save 900 miles of coastline from the threat of damaging development. Although the National Trust extends to Wales and Northern Ireland, there is a separate National Trust for Scotland. In a similar spirit, the Civic Trust was established in 1957 with the object of preserving and enhancing the best "townscape." Many hundreds of local societies dedicated to the protection of the urban environment have been set up, and there are many other voluntary organizations working to protect and improve the English scene, among them the Council for the Protection of Rural England; the Commons, Open Spaces and Footpaths Preservation Society; and the Historic Churches Preservation Trust. Here the law has been extended and strengthened. The Town and Country Planning Acts of 1947 and 1968 require local government planning authorities to prepare development plans and to control development. Buildings of architectural and historic interest are listed and may not be demolished or altered without authority. The Civic Amenities Act of 1967 provided for areas of distinguished urban architecture to be designated "conservation areas." Local authorities may issue tree preservation orders to prevent indiscriminate felling. "Green belts" of countryside protected from building development have been mapped out in the plans for London and other conurbations. The quality of town life has been improved by smoke control and checks on river pollution, so effectively that the recorded sunshine in London and other major urban centres has been greatly increased and the "pea soup" fogs that occurred in London in the early 1950s have become memories of the past. Fish are returning to reaches of rivers such as the Thames, the Tyne, and the Tees from which they had been driven by industrial pollution.

The "leisure explosion"

Among a number of governmental institutions concerned with the environment is the Countryside Commission, which exists to promote recreation in the countryside, to protect designated areas of outstanding natural beauty, and to manage the national parks, seven of which are in England. It also manages long-distance footpaths, among them the Pennine Way, which runs for 250 miles over the summits of the moorland spine of northern England; some 80 miles of the South Downs Way; and a North Downs equivalent that runs, in part, along the route that Chaucer and his pilgrims took to Canterbury. Government recognition of the importance of these policies was indicated by the appointment of a cabinet minister responsible for the environment by the Labour government of 1969 and the founding of a Department of the Environment headed by a cabinet minister by the Conservative government in 1970. A Royal Commission was set up in 1969 to keep a continuous watch on pollution of the environment. While these measures applied to Great Britain as a whole, they were of particular relevance to England with its high-density urban development.

THE PEOPLE AND THEIR LANGUAGE

By ethnic origin the English are a mongrel breed. Their language is polyglot, drawn from a variety of sources, and its vocabulary has been augmented by importations from all over the world. The English language does not serve to identify the English, for it is the main language of Wales, Scotland, Ireland, many Commonwealth countries, and of the United States. The primary source of the language, however, is the main ethnic stem of the English. The Anglo-Saxons who invaded and colonized England in the 5th and 6th centuries have a record of continuous occupation since. It was their language that has prevailed and provides about half the words in modern English vocabulary.

Roman and Anglo-Saxon occupation

In the millennia following the last Ice Age, the British Isles were peopled by migrant tribes from the continent of Europe and, later, by traders from the countries of the Mediterranean area. In the time of the Roman occupation, England was inhabited by Celtic Brythons, but the Celts withdrew before the Teutonic Angles, Saxons, and Jutes (from northwest Germany) into the mountain areas of the west and north. The Anglo-Saxons neither preserved nor absorbed the Roman-British culture they found in the 5th century. There are remarkably few traces of Celtic or Roman Latin in the early English of the Anglo-Saxons, though some words survive in place-names, such as the Latin *castra* for camp providing the suffix "cester" and *combe* and *tor*, Celtic words for "valley" and "hill." Old Norse, the language of the Danes and Norsemen, has left more extensive traces, partly because it was a language with closer affinities to Anglo-Saxon and because the Danish occupation of large tracts of eastern and northern England was for a time quite deeply rooted, as some place-names show.

The history of the centuries before the Norman Conquest is poorly documented, but what stands out is the tenacity of the Anglo-Saxons in surviving a succession of invasions. They united most of what is now England at different times in the 9th, 10th, and 11th centuries, only to be overthrown by the Normans in 1066. For two centuries French became the language of the court and the ruling nobility; yet English prevailed and, by the middle of the 13th century, had re-established itself as the predominant tongue, to become an official language again in 1362. The characteristic English way of life prevailed too. From what evidence comes down from the Dark Ages, it seems that the Anglo-Saxons were stolid, industrious, and persistent colonizers, clearing the forests and opening up new land for farming. They were neither migrants nor mere piratical marauders but settlers.

Both before and after the Norman Conquest, the English showed a notable capacity for absorbing and transmuting outside influences. Church Latin and Norman French were incorporated into the language during the Middle Ages. It was then enriched by the Latin and Greek of the educated scholars of the Renaissance. The seafarers, explorers, and empire builders of the modern history have imported words from all corners of the globe, most copiously from Europe but also from Asia. These foreign words have been so completely absorbed into the language that they pass unselfconsciously as English. The English, it might be said, are great anglicizers.

The English have also absorbed and anglicized people of alien race, from Scandinavian pillagers and Norman conquerors to Latin churchmen. In the royal line, a Welsh dynasty of monarchs, the Tudors, was succeeded by the Scottish Stuarts, to be followed by the Dutch William of Orange and the German Hanoverians. For the Scots, the Welsh, and the Irish, English became their main language, and many moved into England to seek their fortunes and to settle down as English by adoption. England provided a haven for refugees from the time of the Huguenots in the 17th century to the totalitarian persecutions of the 20th century. Jews, too, have settled in England in large numbers. In recent decades there has been large-scale immigration of non-white peoples from India and Pakistan, Africa, and the Caribbean, posing seemingly more difficult problems of assimilation, and restrictive immigration regulations have been successively imposed that are out of key with the open-door policy that had been accepted as an English tradition for many generations. To be counted English, it has never been necessary to be of purebred English stock, and indeed, there can be few English who are.

Immigration of non-white peoples

THE REGIONS OF ENGLAND

Traditional regions. Though a small and homogeneous country bound together by law, administration, and a comprehensive transport system, distinctive regional differences persist in England. These arise from geography and history. It was natural for different groups of the population to establish themselves in recognizable physical areas. In the north, for example, the east and west are separated by the Pennines, which rise to more than 2,000 feet, and the estuaries of the Humber, the Thames, and the Severn form natural barriers. The regions that have been defined for administrative purposes have some arbitrary boundaries but nevertheless reflect true regional feeling. There are eight of them: the Southwest, the Southeast, the West Midlands, the East Midlands, East Anglia, the Northwest, Yorkshire and Humberside, and Northern England.

The Southwest. The Southwest contains the last stronghold of Celtic Britain, Cornwall, where a Celtic language was spoken until the 18th century and where the countryside is dotted with some 300 Celtic stone crosses. There is even a small nationalist movement, *Mebyon Kernow* (Sons of Cornwall), seeking to revive the old language. Though it has, as yet, no political significance, it reflects the disenchantment of a declining area, with the exhaustion of mineral deposits toward the end of the 19th century. The neighbouring county of Devon also shares this economic decline. On the other hand, these southwestern counties have a large share of the rapidly expanding tourist industry, with a splendid coastline and the Dartmoor and Exmoor national parks. Farther east and closer to the economic magnet of the London area, the counties of Dorset, Hampshire, Wiltshire, and Gloucestershire have a greater variety of new industry.

The Southeast. The Southeast, centred upon London, is of a size and wealth to match many nation states, with more than 17,000,000 people, 7,400,000 of them in Greater London. This is the dominant area of England and the most rapidly growing one despite planning controls directed toward restricting its growth. London itself is a victim of urban sprawl that has been checked too late. It is the administrative headquarters not only of government but of many of Britain's industrial and commercial undertakings. The area as a whole has an extensive range of manufacturing industry, though not in great concentrations apart from east London and the Thames Estuary. There is an above-average concentration of science-based industry and research around London. An effectively enforced green-belt policy has slowed

The expansion of London

down the expansion of London, and the region as a whole still has one-third of its area devoted to farming or horticulture. London is the focus of the nation's transport system, with the mainline railway network based on the London termini, two international airports, at Heathrow and Gatwick, and a third proposed at Foulness at the mouth of the Thames. The Port of London is the largest and commercially most important in Britain. Eight new towns have been established in the Southeast, and five existing towns have been expanded to take the overflow of London population, the scarcity of housing in inner London being particularly acute. With its theatres, concert halls, museums, and art galleries, London is emphatically the cultural capital of the country. Whether the people of the Southeast feel a regional identity is questionable. The Sussex Downs and the Bedfordshire plain or Oxford and Canterbury have nothing much in common apart from being within the magnetic pull of London. Loyalties are more specifically to towns, such as Oxford, Cambridge, or Brighton, and within London there is a sense of belonging more to localities—such as Chelsea or Hampstead, which acquire something of the character of urban villages—than to the metropolis as a whole.

The West Midlands. Regional characteristics are stronger outside the Southeast. The West Midlands, centred upon Birmingham, includes the Black Country (an urban agglomeration whose name accurately reflects the coating of grime and soot afflicting the buildings of the region), with some 1,500 different industrial trades, mainly in the metal industries. The major industry is automobile manufacture, with a wide range of component suppliers. With a long history going back to the beginnings of the Industrial Revolution, the West Midland towns are ugly but prosperous, with some of the highest wages in the country. It is an area in which there is a remarkable specialization place by place: locks and keys in Willenhall; needles and hooks in Redditch; the bulk of the country's pottery and china in five towns grouped around Stoke-upon-Trent; chocolate at Bournville; and carpets at Kidderminster. The region is not exclusively industrial, however, with the Shakespeare country around Stratford-on-Avon, the fruit-growing orchards of the Vale of Evesham, and hill country on the Welsh border.

The East Midlands. The East Midlands are less coherent as a region, taking in the manufacturing towns of Northampton (footwear); Leicester (hosiery and knitwear); Nottingham (hosiery, lace, bicycles, pharmaceuticals, tobacco); and Derby (engineering, aero-engines). The region also includes some of the most productive coalfields, ironstone mining, and much of England's best farmland in broad swathes between the industrial towns.

East Anglia. This region retains an air of remoteness that belongs to its history. With the North Sea on its northern and eastern flanks, it was at one time almost cut off by fenland to the west (now drained) and forest (long ago cleared) to the south. In medieval times it was one of the richest of the wool regions and, in some parts, was depopulated to make way for sheep. The traces of deserted villages can still be seen. Now it is the centre of some of the most industrialized farming in England. Vast fields are put down to cereals and sugar beet, and there are five times the national percentage of people employed in farming. It is the most thinly populated of the English regions but is beginning to be industrialized, with North Sea gas being brought ashore on the Norfolk coast.

Industrialized agriculture

The Northwest. The qualities of the regions become more distinctive the farther they are from London. The Northwest is in essence Lancashire, with the conurbations of Manchester and Liverpool, traditionally wet and murky, home of the declining cotton-textiles industry—rapidly being replaced by diversified engineering—but still much ravaged by the Industrial Revolution. It is a region that expresses itself in an accent of its own, with strongly flavoured variety-hall humour (of which George Formby and Gracie Fields were classic examples). It gave birth to British "pop" with the Beatles and other groups in Liverpool and still celebrates an earthy community life in wakes weeks (the factory holidays) and trips to Blackpool, the epitome of the north country seaside. It is a region still painfully breaking into the new territories of modern industry, its old cotton towns symbolically overshadowed by the grim gritstone Pennine escarpments, which have been stripped of their trees by the pollution of generations of industrial smoke.

Yorkshire and Humberside. A similar character belongs to industrial Yorkshire, on the east side of the Pennines watershed, where the valley mills are engaged in the manufacture of woollen textiles and clothing. Here too there is a strong accent to the regional style of speech and a rugged independence of character expressed in a tough style of humour. The conurbations of Leeds and Bradford take in the woollen and commercial centres. Farther south, steel is concentrated at Sheffield, world-famous for its cutlery. West Yorkshire based its prosperity on coal, but, like other parts of the country, it has been diversifying into other industries, notably chemicals and engineering. Here, too, industrialism has left ugly scars. But Yorkshire is not only industry. There are extensive areas of farming, the deep-sea-fishing industry operating from Hull, and tourist country along a fine coast and in the beautiful valleys of the Yorkshire Dales, with an exceptional repertoire of church architecture and country mansions.

Northern England. The official northern region extends to the Scottish border, taking in both North Yorkshire and the Lake District, and is therefore more than usually anomalous. The Lake District is a highly distinctive region of igneous mountains, the highest in England, rising to just over 3,000 feet, and so shut off that it retains a strong sense of local identity quite unlike the Northeast, which, from the river Tees to the Scottish border, is by cultural climate a region of its own. Here there are more concentrations of declining 19th-century industry, notably in the Durham coalfields and some of the Tyneside shipyards. It is at the same time a region in which modernization is going on rapidly, particularly in the complex of chemical industries on Teesside, one of the largest concentrations of chemical industries in Europe. The local flavour of life can be found in the folk songs of Tyneside and the coal-mining villages and in the dialect known as Geordie. The smallest of the designated conurbations, of some 800,000 population, is centred at Newcastle upon Tyne. The region also contains some of the most deserted land in England, in the Cheviot Hills along the Scottish border.

The Lake District

Regional disparities. The regional differences of England invite comparisons that re-emphasize the predominance of the south and particularly of the London area. The regions of the North and the Southwest may be more distinctive in character, but they are significantly poorer in wealth and social resources. The Southeast and the West Midland regions, on the other hand, are above the national average for incomes and the proportion of high-income earners. The worst parts of the North have a much higher mortality rate than the best parts of the Southeast, and this can be explained both by the bad housing, the higher level of unemployment, and the worse poverty of the North. There is a greater concentration of social services in the Southeast, especially in the London area: a wider choice of schools, hospitals, doctors, dentists, and medical specialists. The best professional and managerial jobs are to be found in London and the Southeast; there are more school pupils staying on after the minimum school-leaving age, more university students, and more universities. The attractions of this area and also of the southern parts of the Midlands are much greater for the setting up of new industries than places in the North that have a greater need of them. Hence the drift away from declining areas far from London into the already overcrowded Southeast has continued regardless of planning controls and incentives to industrialists to go to designated development areas in less favoured parts of the country.

ADMINISTRATION AND SOCIAL CONDITIONS

Government and politics. It is hard to identify a specifically English role in contemporary government and

politics, for these operate on a nationwide British basis. Historically, the English may be credited with the evolution of Parliament, which, in its medieval form, was related to the Anglo-Saxon practice of regular gatherings of notables, and the glory of the 1688 Revolution, which affirmed the rule of law, parliamentary control of taxation and of the army, freedom of speech, and religious toleration. Freedom of speech and opinion with proper opportunities for reasonable debate form part of the continuing English tradition, but the development of party and parliamentary government in its modern forms took place after the Act of Union of 1707, since when, in politics, the history of England has been the history of Britain. There have been prime ministers as robustly English as Robert Walpole, William Pitt, or Winston Churchill, but they are in a line that includes the Jew Benjamin Disraeli, the Welshman David Lloyd George, and the Scots Harold Macmillan and Alec Douglas-Home.

Local government

In government the English legacy remains conspicuous in local affairs, which are still largely administered on a county base that can be traced back to the Anglo-Saxon shires. The separation of county and town in local government was the principle underlying the late Victorian reorganization of local government, but the reforms of the 1970s, while retaining an exclusively urban base for local government in the conurbations, reunite county and town elsewhere with a minimum of disturbance to county boundaries that in many places date back 900 years.

The law. The English have given the world, notably North America and much of the Commonwealth, the system of English law that has acquired a status and universality to match Roman law. This, too, had its origins in Anglo-Saxon times, and two of its hallmarks are its preference for customary law (the common law) rather than statute law and its system of application by locally appointed part-time magistrates or justices of the peace, locally chosen juries, and by the travelling judges going from one county town to another on circuit. Under the Normans, the Anglo-Saxon system was retained but formalized; for example, by the recording of case law to provide uniform precedents. This began toward the end of the 13th century. In modern times there has been a greater reliance upon the statute law contained in about 3,000 acts of Parliament, but there are over 300,000 recorded cases to turn to for precedent. Other aspects of the English law are the fundamental assumption that an accused person is deemed innocent until proved guilty and the independence of the judiciary from intervention by crown or government in the judicial process.

Religion. Increasingly, England is a secular country. Though more than one-half the population are baptized in an Anglican church and for form-filling purposes would say they belonged to the Church of England, fewer than one in ten of the baptized are communicant churchgoers. The Church of England still has 18,000 churches, but it has been in financial difficulties, and it has closed down underused churches and sold off redundant sites. The nonconformist Free Churches have nominally fewer members, but there is probably greater dedication among them, as with the Roman Catholic Church, with about 4,000,000 adherents in England. Apart perhaps from some isolated centres of Irish settlement in the Northwest, there is complete religious tolerance in England and no overt prejudice against Catholics; it is quite usual for 30 or 40 members of Parliament to be Catholic, with a few of them government ministers. The decline in churchgoing has been thought to be an indicator of decline in religious belief, but opinion polls substantiate the view that belief in God and the central tenets of Christianity survives the flagging fortunes of the churches.

CULTURAL LIFE

England's contribution to the culture of the United Kingdom and of the world is too vast for anything but a random sampling. There is the complicating factor that in the contemporary cultural scene, England is not always distinguishable from Wales and Scotland or even Northern Ireland. Furthermore, what may count as mainly English achievements have often been shared with people from other countries.

Science. In the sciences, for example, the splitting of the atom was first accomplished in the Cavendish Laboratory in Cambridge by Ernest Rutherford, a New Zealander. The deoxyribonucleic acid (DNA) molecule was unravelled in Cambridge by two English geneticists, Francis Crick and Maurice Wilkins, and an American, James Watson. Radar detection of aircraft was perfected at a research station in Suffolk by a team headed by a Scot, Robert Watson-Watt. Penicillin was discovered in a London hospital by Alexander Fleming, a Scot, and developed as a medical drug by Howard Florey, an Australian, in Oxford. In the sciences the English are sometimes said to be better at fundamental theory than at its application. Certainly Isaac Newton on gravitation and Charles Darwin on evolution have changed the history of scientific thought. But William Harvey on the circulation of the blood and Joseph Lister on antiseptic surgery were working as practical surgeons. The innovating engineers of the early Industrial Revolution were highly practical men. The choice of the Greenwich meridian as the base line for measuring longitude reflects the practicality of English astronomers, navigators, and geographers. The practical application of highly sophisticated new techniques to fundamental research is seen in the radio telescopes at Jodrell Bank and at Cambridge. In the design of nuclear power plants, research institutions in England have been abreast or ahead of the rest of the world. England has also been ahead with the design of advanced aircraft. One of the first jet engines was designed by Frank Whittle and flown in England in 1941, and the Anglo-French supersonic airliner Concorde is built in England at Bristol and powered by Rolls-Royce engines made in Derby.

Eminent Englishmen in science

Architecture. England, however, is still most famous for its architecture and its literature, both reflecting the regional diversity of the country. Architecture has been much influenced from abroad, but foreign styles take on an English aspect. So the Gothic architecture of France was transformed into a characteristically English style by the delicate use of stone to provide a framework for walls that were almost all glass, culminating in triumphs of the perpendicular style, such as King's College Chapel at Cambridge. European Renaissance influences can be seen in the buildings of Christopher Wren, yet his many London churches seem essentially English. The magnificent country houses of the 18th century are likewise not mere importations of a foreign fashion but fit their own landscape; and many such landscapes were designed by the great English garden and park designers, William Kent, Lancelot ("Capability") Brown, and Humphry Repton. Small-scale domestic architecture is often just as commendable for its fitness to its situation, and an intimate, friendly urbanity is to be found in the best of England's town centres and villages, this being partly the result of using local materials, whether stone, brick, tile, or thatch. The bulk of the building of the last 100 years excites much less enthusiasm. Many industrial slums have been condemned for demolition, but much contemporary building that is adequate for habitation or work is drearily uninspired. The reconstruction of the war-damaged city areas of England provided opportunities for notable new architecture, and some original design and construction was undertaken; for example, the Barbican scheme in a large bombed area north of St. Paul's, London, and the group of concert halls with an art gallery and the new National Theatre on the south bank of the Thames. Outside London, notable projects include the Coventry precinct and cathedral by Sir Basil Spence; a batch of new universities founded during the 1960s, among them those near Brighton, Canterbury, Colchester, Norwich, and York; and a number of local-authority housing developments, including new towns.

Modern building trends

Literature. In literature, regional influences have often strongly affected English writers—such as Wordsworth in the Lake District, the Brontë sisters on the Yorkshire moors, Thomas Hardy in Dorset, D.H. Lawrence in the Midlands coalfields, and Charles Dickens in London. It

is a characteristic that persists, in J.B. Priestley, a Yorkshireman, in Evelyn Waugh, a west-country squire, and John Betjeman, a Londoner, whose poetry is a record of his passion for English towns and landscape.

Visual arts. Regional attachments are also to be found in the visual arts. The sculptors Henry Moore and Barbara Hepworth both come from Yorkshire, and something of the quality of moorland stone can be seen in their work. Hepworth's later work has echoes of the granite coastline of Cornwall, where she settled, and Cornwall has become the chosen working place of a number of English artists. Even in music the landscape intrudes, as in Benjamin Britten's opera *Peter Grimes*, based as it is on a story of the fishing village in Suffolk where Britten lives.

Institutionally, the arts in Britain have grown in both scope and vigour since World War II. Publishers break records for the numbers of books produced; the sale of paperbacks has spread from established bookshops to newsagents and tobacconists. London theatres (about 30 in the West End) play to full houses, and many new repertory theatres have been established in the provinces. Music flourishes as never before in England, and visiting conductors have said that they find the English concert audience one of the most appreciative in the world. Long-playing records are sold by the tens of millions annually. Attendance at art galleries and museums doubled in the 1960s, but it was feared that the introduction of admission charges might reduce attendance. Subsidies to support the arts and literature in Britain are administered by the Arts Council, a considerable proportion of them being devoted to maintaining and extending the arts in regional centres outside London. Arts Council subsidies have also been used to develop the international fame of companies such as the Royal Opera and the Royal Ballet (both of Covent Garden) and of leading London orchestras, which make regular foreign tours.

The Arts Council

Leisure. Though England has a lively intellectual and artistic life, its characteristic pursuits are of a more popular kind. The exploitation of leisure is increasingly the concern of commerce: holiday camps, foreign holiday package tours, gambling of many kinds from bingo to horse-race betting and football (soccer) pools, and the transformation of the traditional English pub by trendy interior decoration and the introduction of new kinds of beer allegedly more palatable to women. The English weekend, accorded to most of the population, is the occasion for trips into the countryside and for outdoor activities ranging from fishing to mountaineering. England gave to the world the sports of cricket, association football, and rugby football, but team and spectator sports tend to be giving way to more individualistic activities. In spite of persistent commercial tempting to do something else, the English remain a stay-at-home people. The Englishman's home is still his castle, and gardening the most frequent pastime, except for watching television, which absorbs nearly a quarter of the average person's leisure time. The English prefer living in houses to apartments, and about three out of four have gardens. They keep pets, they "do-it-yourself" in house maintenance and many other odd jobs, and their desire for privacy still mystifies foreign onlookers.

THE OUTLOOK

Will the Englishness of England survive? Its essential qualities are small scale and local, even insular and parochial, whether one is considering the landscape or the people, in town or country. The prevailing trends are almost all inimical, blatantly so when supergrid electric power lines dwarf the village scene or when the street corner village-in-the-city community is obliterated by an urban motorway or displaced by the supermarket. The life of provincial capitals is diminished when the fast new intercity rail services link them to London by a journey of two hours or so. England is most truly England when separate places retain their distinctive individuality, whether in the style of a town centre or in the speech and the occupations of the people. Currently a powerful and perhaps irresistible process of standardization averages out and levels down the differences between places. A standard housing estate in Southampton may be indistinguishable from its counterpart in Carlisle at the other end of the country. "Subtopia" is the word that has been coined for this spread of a faceless urban uniformity. A different kind of facelessness is spreading over the countryside as industrialized farming imposes its standardized requirements on the land, grubbing out hedgerows and hedgerow trees to create a prairie of 100-acre fields geared to the tempo of the tractor. Thousands of miles of hedgerow have been disappearing each year as England establishes itself as the country with more farm tractors to the acre than anywhere else in the world. The personality of regions and of individual towns is diluted by the move away from specialization in one industry. Often the declining industries—such as textiles, steel, shipbuilding, and coal—have been the foundation for a strong regional character. Diversification into other industries paradoxically has a sameness rather than variety about it, for much the same mix of new light industries settles into new trading estates wherever they may be. Ideas, manners, topics of conversation, styles of clothing, ways of speech, and tastes in food and drink increasingly tend to be standardized throughout the country by the pervasive influence of mass-circulation newspapers, network television and radio, and the chain stores that dominate every town centre. England is served by and serves the mass market. Physical mobility (from place to place) and social mobility (from one economic class to another) tend toward rootlessness. The English now more commonly move away from their birthplace than stay there, the population drift being away from established city and town centres toward suburbs. The 1971 census showed that all the conurbations except West Yorkshire had lost population since 1951. More than 600,000 moved out of London in the decade 1961 to 1971, and of the 35 next largest towns 22 had a declining population.

"Subtopia"

There are, however, vigorous counter-responses to these tendencies. Planning authorities, egged on by civic societies, attempt to remedy the visual clutter of subtopia, and the renewal of city centres is high in planning priorities. Conservationists are active in campaigning to protect the countryside, while the English as a whole have been growing more actively critical and concerned about their environment. The right to have a bigger say by public participation in local planning has been recognized in the 1968 Town and Country Planning Act.

In temperament the English have often been thought by outside observers to be reserved, conventional, deferential, snobbish; and (as the French Duc de Sully observed in the 17th century) to take their pleasures sadly. Much of that has changed since 1945. The English led the way in fashion with bright colours, miniskirts, and gay, casual dress for men. The English turned rock music into pop and spiced it with not-so-quiet social revolution. Social restlessness in England in some of its more conspicuous forms—student protest, the underground press, and women's liberation—seemed sometimes to be borrowing second-hand from abroad. England, like other countries, was not immune to global "trendiness." More important, in a society that had been noted for its elaborately stratified class structure, the revolt against deference saw the end of conventional class snobbery, though it was fast replaced by a snobbery of success. Outbreaks of teenage violence suggested that England might be becoming not only brassy and brazen but "bloodyminded" too. Yet foreign observers continued to repeat the well-established compliments to the English for their politeness, reasonableness, good nature, and the relaxed tolerance of their democracy.

BIBLIOGRAPHY. Many of the most useful sources deal with the United Kingdom rather than with England exclusively. An official handbook, *Britain* (HMSO, annual), surveys contemporary British life and institutions, and includes an extensive up-to-date bibliography. Other important official sources are the *Census 1971* reports and *Social Trends*, a statistical analysis periodically brought up to date.

Among recent histories are the *Oxford History of England*, 15 vol. (1936–65), whose last volume, A.J.P. TAYLOR, *English History, 1914–45*, has a discursive bibliography; and *The*

Pelican History of England, 8 vol. (1955–65), whose last volume, DAVID THOMSON, *England in the Twentieth Century, 1914–63*, also has a helpful bibliography. For social history, G.M. TREVELYAN, *English Social History: A Survey of Six Centuries, Chaucer to Queen Victoria* (1946), is classic but somewhat conventional; the interconnections of social and cultural history are examined by RAYMOND WILLIAMS, *Culture and Society, 1780–1950* (1958) and *The Long Revolution* (1961). RICHARD HOGGART, *The Uses of Literacy* (1957), deals with the impact of mass media on working class culture; the structure and role of newspapers is examined by COLIN SEYMOUR-URE in *The Press, Politics, and the Public* (1968). Regional differences are surveyed in B.E. COATES and E.M. RAWSTRON, *Regional Variations in Britain* (1971), and D. ELLISTON ALLEN, *British Tastes* (1968). Some other studies of 20th-century English life and attitudes are GEORGE ORWELL, *The English People* (1947); ROBERT GRAVES and ALAN HODGE, *The Long Weekend: A Social History of Great Britain, 1918–1939* (1963); MALCOLM MUGGERIDGE, *The Thirties*, new ed. (1967); R. LEWIS and A. MAUDE, *The English Middle Classes* (1949); GEOFFREY GORER, *Exploring English Character* (1955) and *Sex and Marriage in England Today* (1971); and E.J.B. ROSE, *Colour and Citizenship: A Report on British Race Relations* (1969).

England's place in contemporary British politics can be judged from D.E. BUTLER, with various associates, in *The British General Election* (separate volumes on the elections of 1950, 1951, 1955, 1959, 1964, 1966, and 1970); D.E. BUTLER and D. STOKES, *Political Change in Britain* (1969); D.E. BUTLER and J. FREEMAN, *British Political Facts 1900–1967* (1968); and D.E. BUTLER, *The Electoral System in Britain* (1953). *The Royal Commission Report on Reform of English Local Government* (1969) is the most valuable and thorough recent study even though its proposals were not adopted.

For geology, geography, and natural history, the "New Naturalist Series" provides authoritative studies for the general reader, among them DUDLEY STAMP, *British Structure and Scenery*, 2nd ed. (1947); GORDON MANLEY, *Climate and the British Scene* (1952); and J.A. STEERS, *The Sea Coast* (1953). Another work by J.A. STEERS, *The English Coast and the Coast of Wales* (1966), is a paperback associated with a campaign to preserve the coastline. A brief introduction to wildlife, with useful book lists, is R.S.R. FITTER, *Wildlife in Britain* (1963). For conservation of the landscape and of the human environment, recent history is summarized in DUDLEY STAMP, *The Land of Britain: Its Use and Misuse*, 3rd ed. (1962), and *Nature Conservation in Britain* (1969). W.G. HOSKINS, *The Making of the English Landscape* (1955), is based on original field research; COLIN BUCHANAN, *Traffic in Towns* (1964), is a seminal official study; NAN FAIRBROTHER, *New Lives, New Landscapes: Planning for the 21st Century* (1970), deals with agriculture in an urban society; ROBERT ARVILL, *Man and Environment* (1967); and JOHN BARR, *Derelict Britain* (1969), study the impact of industrialism. There are useful book lists in H.E. BRACEY, *People and the Countryside* (1970); and VICTOR BONHAM-CARTER, *The Survival of the English Countryside* (1971). Recent trends in urban development are examined in R.H. BEST and J.T. COPPOCK, *The Changing Use of Land in Britain* (1962); FREDERICK GIBBERD, *Town Design* (1967); F. OSBORN and A. WHITTICK, *The New Towns* (1969); J.B. CULLINGWORTH, *Housing and Local Government* (1966); and D.V. DONNISON, *The Government of Housing* (1967). England's historic architecture is cataloged with lively commentary in NIKOLAUS PEVSNER, *The Buildings of England*, 37 vol. (1951–70), county by county; and its local characteristics are the theme of A. CLIFTON-TAYLOR, *The Pattern of English Building* (1962). W. GAUNT, *A Concise History of English Painting* (1964), is a reliable guide; NIKOLAUS PEVSNER, *The Englishness of English Art* (1956), expands a series of broadcasts. A. SAMPSON, *Anatomy of Britain* (1971), is a breezy, informative work on the world of public institutions, politics, and business.

(W.H.Th.)

English Channel

A narrow, wedge-shaped sea separating Great Britain and France and tapering eastward to its junction with the North Sea at the Strait of Dover (French Pas de Calais), the English Channel (French La Manche), with a size of 34,700 square miles (89,900 square kilometres), is the smallest of the shallow seas covering the fringing continental shelf of Europe. From its mouth in the North Atlantic Ocean—an arbitrary limit marked by a line between the Scilly Isles and the Isle of Ushant—its width gradually narrows from 112 miles to a minimum of 21 miles, while its average depth decreases from 400 to 150 feet. Although the English Channel is a feature of notable scientific interest, especially in regard to tidal movements, its strategic location lent it immense significance, over the centuries, as both a route and a barrier during the peopling of Britain and the emergence of the nation states of modern Europe. By the 1970s, as the United Kingdom appeared to turn politically toward Europe, the long-discussed project for a Channel tunnel at the sea's eastern end—potentially one of the greatest engineering projects in the world—moved closer to realization.

Strategic importance

The current English name (in general use since the early 18th century) probably derives from the designation "canal" in late 16th-century Dutch sea atlases, although earlier names had included Oceanus Britannicus and the British See; the French have regularly used La Manche (a reference to the sleevelike coastal outline) since the early 17th century. Scholars adduced reasons for the Channel's existence as early as the 17th century, but detailed scientific study awaited the first official hydrographic surveys (French coast: 1829; English coast: 1847), and the geological map of the seabed based on borings made in 1866 was the world's first of its kind. Further studies were associated with early plans for a Channel tunnel, and modern surveys, using sophisticated techniques, have made the Channel seabed one of the most intensively studied sea floors in the world since the end of World War II.

Geological evolution. The contemporary Channel is probably the result of a complex structural downfolding dating from the mid-Tertiary (about 25,000,000 years ago), although signs of a downwarp tendency occur as early as 270,000,000 years ago. The direct ancestor of the Channel may well have been a sea occupying the downfold 1,000,000 or 2,000,000 years ago, with a sea level 600–700 feet above the present level. The withdrawal of water by the Ice Age glaciers produced, about 25,000 years ago, a sea level at least 300 feet lower than the present. Later, the melting of the ice raised the sea level to its present mark, and the ecologically important land bridge across the Dover Strait was finally submerged about 8,000 years ago.

Submarine topography and deposits. The sea floor dips fairly steeply near the coasts but is generally flat and remarkably shallow (especially in relation to nearby land elevations), with its greatest depth of 565 feet in the Hurd Deep, one of a group of anomalous deep, enclosed, troughs in the bed of the western Channel. It has been shaped by the effect upon its rock strata (with their varying degrees of hardness) of such forces as: weathering and erosion (when much of the area was dry land); sea-level changes; and contemporary erosion and deposition by marine currents.

The floor of the western Channel (200 to 400 feet deep) is relatively flat and featureless, reflecting fairly uniform rock types, mostly limestone. Harder igneous rocks cause shoals to emerge (as in the case of the Scilly Isles and Channel Islands), and submerged cliffs and narrow depressions provide some additional variety.

In the central Channel (150 to 200 feet deep) depths are fairly uniform over chalk outcrops, but alternations of clays and limestones give rise to an undulating terrain, with deeps reaching almost twice the average. A continuation of the Seine Valley system north of the Cotentin Peninsula complicates the relief forms. Farther east again, the sea floor is smoother and the geology simpler. Depths range from six to 160 feet, with such elongated banks as the Varne and the Ridge greatly constricting shipping lanes.

Because the Channel, unlike the Irish or North seas, lay south of the scouring action of Ice Age glaciers, superficial deposits are either very thin (three feet or less) or entirely absent. They represent a complex reworking of deposits of various ages, and their distribution reflects tidal currents: where these are strong, the seabed is bare except for pebbles; decreasing velocities give rise to sand and gravel ribbons and waves (the latter up to 40 feet thick) and to thick beds of fine-grained deposits in sheltered areas, notably Golfe de Saint-Malo.

Hydrography. Tides are generally strong, especially in the Dover Strait, and may be visualized as an oscillation

The "double-tide" phenomenon

(modified by Earth rotation and configuration) about a north–south line through the centre of the Channel—*i.e.*, with a rise to the west accompanying a fall to the east. The central portion experiences daily double tides (helpful to shipping movements at Southampton which has a double, or prolonged, high tide), and the Golfe de Saint-Malo experiences, at 28 feet or more, the greatest tidal range.

Surface temperatures range from 45° F (7° C) in February to 61° F (16° C) in September, although shallow coastal waters are warmer in summer. There is little temperature change with depth in the well-mixed eastern waters, but bottom-water temperatures fall to 41° F (5° C) in the west. Surface salinities decline eastward from just under the Atlantic level of 35.6 parts per thousand; coastal readings are further reduced by the influx of river water, especially from the larger French landmass. There is an overall water flow through the Channel to the North Sea, a complete replacement taking about 500 days.

Historical associations. From earliest times, depending on historical factors, the Channel served as a route for, and a barrier to, invaders of Britain from the continent. Early Stone Age man crossed the Strait of Dover; later invaders crossed the western end of the Channel, trading the copper, tin, and lead they found in England's Devon and Cornwall, and successive Bronze and Iron Age invaders followed the same route. Caesar's invasion of 55 BC again favoured the Dover route in the east, while William the Conqueror's famous invasion in 1066 crossed from Normandy to Hastings. With the later loss of Normandy to the British, the Channel again became a defensive line: it was the scene of major military actions from the sea battle of 1588, pitting the English against the Spanish Armada, to the Dunkirk evacuation in 1940 and the launching of the Normandy beachhead in 1944.

Resources and economic activity. Located at the junction of the Atlantic Ocean and the North Sea, the respective waters of which are rich in warm- and cold-water plankton, the Channel is favoured from the latter with cod, herring, and whiting and from the former with hake, pilchard, and mullet. The traditional fishing industry has declined in the present century with the development of deep-sea fishing, the exhaustion of resources, and the advent of pollution problems; but coastal fishing remains important in Bretagne (Brittany), and skin diving has stimulated shell fishing in Cornwall.

A good climate, sandy beaches, and an attractive coast have encouraged the growth of a major tourist industry on both sides of the Channel, starting with the fashionable resorts of the late 18th century. Portsmouth and Plymouth are now declining naval ports, with very minor commercial functions. Cherbourg is probably not changing much in character, but Southampton and Le Havre are losing passengers while gaining tremendous container and oil-refining capacity and also experiencing a general commercial growth. England and France both use Channel waters for cooling nuclear-powered generating stations, while the tidal-power generating station on the River Rance (in Bretagne), utilizing a tidal range of over 37 feet, is a unique feature. Another unique feature is the Channel tunnel project, conceived in 1802, initiated, and then halted, in the late 19th century and revived in 1957. By the 1970s the idea of a tunnel under the Strait of Dover seemed to be coming closer to realization: if finally implemented, it would have immense economic effects on the entire region bordering the Channel.

BIBLIOGRAPHY. The most concise and up-to-date accounts of the English Channel from the point of view of navigation, the ports, and natural conditions are contained in *The Channel Pilot* (1971) and *The Dover Strait Pilot* (1971), published by the Hydrographic Office of the British Navy. A useful summary of physical conditions, together with bibliography, may be found in R.W. FAIRBRIDGE (ed.), *Encyclopedia of Oceanography* (1966); and the wider geological setting is outlined in D.T. DONAVAN (ed.), *Geology of Shelf Seas* (1968). An important early account of geology and structure is W.B.R. KING, "The Geological History of the English Channel," *Q. Jl. Geol. Soc. Lond.*, 110:77–101 (1954); and morphology is discussed by A.H.W. ROBINSON in "The Floor of the British Seas," *Scott. Geog. Mag.*, 68:64–79 (1952). The most up-to-date review of Channel structure, geological outcrops, and superficial deposits, with comprehensive bibliographies, is found in F.M. DELANY (ed.), *The Geology of the East Atlantic Continental Margin*, 4 pt. (1971). Tidal information will be found in the British Admiralty Tide Tables and Pocket Tidal Atlases. A helpful summary is found in J.A. WILLIAMSON, *The English Channel: A History* (1959). The definitive publication on the Channel tunnel is *Proposals for a Fixed Channel Link* (HMSO Cmd. 2137; 1963); and a recent assessment is DONALD HUNT, "Channel Tunnel Enters Final Phase," *New Scient.*, 49:238–239 (1971). An excellent account of British ports and trade is JAMES BIRD, *The Major Seaports of the United Kingdom* (1963).

(C.E.E.)

English Language

English is widely spoken in all six continents and has had a strong effect in many regions in which it is not the principal language spoken. The portmanteau words Franglais, Russlish, and Japlish, for example, have been invented by resentful purists to describe the numerous expressions, in vogue among the young, resulting from the infiltration of English into French, Russian, and Japanese, respectively.

The widely diffused English-speaking community is fairly stable in the British Isles, North America, and Australasia. In Africa, the Indian subcontinent, and Southeast Asia its future remains uncertain and unpredictable. People who speak English fall into three groups: those who have inherited it as their native language; those who have acquired it as a second language within a society that is largely bilingual; and those who are driven by necessity to use it for some practical purpose—administrative, professional, or educational. One person in seven of the world's entire population now belongs to one of these three groups.

This article is organized as follows:

Origins and basic characteristics
Characteristics of Modern English
  Phonology
  Morphology
  Syntax
  Vocabulary
  Orthography
Historical background
  Old English
  Middle English
  Transition from Middle English to Early Modern English
  Restoration period
  Age of Johnson
  19th and 20th centuries
Varieties of English
  British English
  American and Canadian English
  Australian and New Zealand English
  The English of India–Pakistan
  African English
The future of English

ORIGINS AND BASIC CHARACTERISTICS

Parent tongue

English belongs to the Indo-European family of languages and is therefore related to most other languages spoken in Europe and western Asia from Iceland to India. The parent tongue, called Proto-Indo-European, was spoken about 5,000 years ago by nomads believed to have roamed the southeast European plains. Germanic, one of the language groups descended from this ancestral speech, is usually divided by scholars into three regional groups: East (Burgundian, Vandal, and Gothic, all extinct), North (Icelandic, Faeroese, Norwegian, Swedish, Danish), and West (German, Netherlandic [Dutch and Flemish], Frisian, English). Though closely related to English, German remains far more conservative than English in its retention of a fairly elaborate system of inflections. Frisian, spoken by the inhabitants of the New Dutch province of Friesland and the islands off the west coast of Schleswig, is the language most nearly related to Modern English. Icelandic, which has changed little over the last thousand years, is the living language most nearly resembling Old English in grammatical structure.

Modern English is analytic (*i.e.*, relatively uninflected),

whereas Proto-Indo-European, the ancestral tongue of most of the modern European languages (*e.g.*, German, French, Russian, Greek), was synthetic, or inflected. During the course of thousands of years, English words have been slowly simplified from the inflected variable forms found in Sanskrit, Greek, Latin, Russian, and German, toward invariable forms, as in Chinese and Vietnamese. The German and Chinese words for "man" are exemplary. German has five forms: *Mann, Mannes, Manne, Männer, Männern.* Chinese has one form: *ven.* English stands in between, with four forms: man, man's, men, men's. In English only nouns, pronouns, and verbs are inflected. Adjectives have no inflections aside from the determiners "this, these" and "that, those." (The endings *-er, -est,* denoting degrees of comparison, are better regarded as noninflectional suffixes.) English is the only European language to employ uninflected adjectives; *e.g.*, "the tall man," "the tall woman," compared to Spanish *el hombre alto* and *la mujer alta.* As for verbs, if the Modern English word ride is compared with the corresponding words in Old English and Modern German, it will be found that English now has only five forms (ride, rides, rode, riding, ridden), whereas Old English *ridan* had 13, and Modern German *reiten* has 16 forms.

In addition to this simplicity of inflections, English has two other basic characteristics: flexibility of function and openness of vocabulary.

Loss of inflection

Flexibility of function has grown over the last five centuries as a consequence of the loss of inflections. Words formerly distinguished as nouns or verbs by differences in their forms are now often used as both nouns and verbs. One can speak, for example, of "planning a table" or "tabling a plan," "booking a place" or "placing a book," "lifting a thumb" or "thumbing a lift." In the other Indo-European languages, apart from rare exceptions in Scandinavian, nouns and verbs are never identical because of the necessity of separate noun and verb endings. In English, forms for traditional pronouns, adjectives, and adverbs can also function as nouns; adjectives and adverbs as verbs; and nouns, pronouns, and adverbs as adjectives. One speaks in English of the Frankfurt Book Fair, but in German one must add the suffix *-er* to the place-name and put attributive and noun together as a compound, Frankfurter Buchmesse. In French one has no choice but to construct a phrase involving the use of two prepositions: Foire du Livre de Francfort. In English it is now possible to employ a plural noun as adjunct (modifier), as in "wages board" and "sports editor"; or even a conjunctional group, as in "prices and incomes policy" and "parks and gardens committee."

Openness of vocabulary implies both free admission of words from other languages and the ready creation of compounds and derivatives. English adopts (without change) or adapts (with slight change) any word really needed to name some new object or to denote some new process. Like French, Spanish, and Russian, English frequently forms scientific terms from Classical Greek word elements.

English possesses a system of orthography that does not always accurately reflect the pronunciation of words; this is discussed below in the section *Orthography.*

CHARACTERISTICS OF MODERN ENGLISH

Phonology. British Received Pronunciation (RP), by definition, the usual speech of educated people living in London and southeastern England, is one of the many forms of standard speech. Other pronunciations, although not standard, are entirely acceptable in their own right on conversational levels.

The chief differences between British Received Pronunciation, as defined above, and a variety of American English, such as Inland Northern (the speech form of western New England and its derivatives, often popularly referred to as General American), are in the pronunciation of certain individual vowels and diphthongs. Inland Northern American vowels sometimes have semiconsonantal final glides (*i.e.*, sounds resembling initial *w*, for example, or initial *y*). Aside from the final glides, this American dialect shows four divergences from British English: (1) the words cod, box, dock, hot, and not are pronounced with a short (or half-long) low front sound as in British "bard" shortened (the terms front, back, low, and high refer to the position of the tongue); (2) words such as bud, but, cut, and rung are pronounced with a central vowel as in the unstressed final syllable of "sofa"; (3) before the fricative sounds *s, f,* and *θ* (the last of these is the *th* sound in "thin") the long low back vowel *a,* as in British "bath," is pronounced as a short front vowel *a,* as in British "bad"; (4) high back vowels following the alveolar sounds *t* and *d* and the nasal sound *n* in words such as tulips, dew, and news are pronounced without a glide as in British English; indeed, the words sound like the British "two lips," "do," and "nooze" in "snooze." (In several American dialects, however, these glides do occur.)

The consonants

The 24 consonant sounds comprise six stops (plosives): *p, b, t, d, k, g;* the fricatives *f, v, θ* (as in "thin"), *ð* (as in "then"), *s, z,* ʃ (as in "ship"), ʒ (as in "pleasure"), and *h;* two affricatives: *t*ʃ (as in "church") and *d*ʒ (as the *j* in "jam"); the nasals *m, n,* ŋ (the sound that occurs at the end of words such as "young"); the lateral *l;* the vibrant or retroflex *r;* and the semivowels *j* (often spelled *y*) and *w*. These remain fairly stable, but Inland Northern American differs from British English in two respects: (1) *r* following vowels is preserved in words such as "door," "flower," and "harmony," whereas it is lost in British; (2) *t* between vowels is voiced, so that "metal" and "matter" sound very much like British "medal" and "madder," although the pronunciation of this *t* is softer and less aspirated, or breathy, than the *d* of British English.

Like Russian, English is a strongly stressed language. Four degrees of stress may be differentiated: primary, secondary, tertiary, and weak, which may be indicated, respectively, by acute (ˊ), circumflex (ˆ), and grave (ˋ) accent marks and by the breve (˘). Thus, "Têll mè thĕ trúth" (the whole truth, and nothing but the truth) may be contrasted with "Têll mé thĕ trûth" (whatever you may tell other people); "blâck bîrd" (any bird black in colour) may be contrasted with "blâckbìrd" (that particular bird *Turdus merula*). The verbs "permít" and "recórd" (henceforth only primary stresses are marked) may be contrasted with their corresponding nouns "pérmit" and "récord." A feeling for antepenultimate (third syllable from the end) primary stress, revealed in such five-syllable words as equanímity, longitúdinal, notoríety, opportúnity, parsimónious, pertinácity, and vegetárian, causes stress to shift when extra syllables are added, as in "histórícal," a derivative of "hístory" and "theatricálity," a derivative of "theátrical." Vowel qualities are also changed here and in such word groups as période, periódical, periodícity; phótograph, photógraphy, photográphical. French stress may be sustained in many borrowed words; *e.g.*, bizárre, critíque, duréss, hotél, prestíge, and techníque.

Pitch, or musical tone, determined by the rate of vibration of the vocal cords, may be level, falling, rising, or falling–rising. In counting "one," "two," "three," "four," one naturally gives level pitch to each of these cardinal numerals. But if a person says "I want two, not one," he naturally gives "two" falling pitch and "one" falling–rising. In the question "One?" rising pitch is used. Word tone is called pitch, and sentence tone is referred to as intonation. The end-of-sentence cadence is important for meaning, and it therefore varies least. Three main end-of-sentence intonations can be distinguished: falling, rising, and falling–rising. Falling intonation is used in completed statements, direct commands, and sometimes in general questions unanswerable by "yes" or "no"; *e.g.*, "I have nothing to add." "Keep to the right." "Who told you that?" Rising intonation is frequently used in open-ended statements made with some reservation, in polite requests, and in particular questions answerable by "yes" or "no": "I have nothing more to say at the moment." "Let me know how you get on." "Are you sure?" The third type of end-of-sentence intonation, first falling and then rising pitch, is used in

sentences that imply concessions or contrasts: "Some people do like them" (but others do not). "Don't say I didn't warn you" (because that is just what I'm now doing). Intonation is on the whole less singsong in American than in British English, and there is a narrower range of pitch. American speech may seem more monotonous but at the same time may sometimes be clearer and more readily intelligible. Everywhere English is spoken, regional dialects display distinctive patterns of intonation.

Intonation in American and British English

Morphology. *Inflection.* Modern English nouns, pronouns, and verbs are inflected. Adjectives, adverbs, prepositions, conjunctions, and interjections are invariable.

Most English nouns have plural inflection in (*-e*)*s*, but this form shows variations in pronunciation in the words cats (with a final *s* sound), dogs (with a final *z* sound), and horses (with a final *iz* sound), as also in the 3rd person singular present-tense forms of verbs: cuts (*s*), jogs (*z*), and forces (*iz*). Seven nouns have mutated (umlauted) plurals: man, men; woman, women; tooth, teeth; foot, feet; goose, geese; mouse, mice; louse, lice. Three have plurals in *-en*: ox, oxen; child, children; brother, brethren. Some remain unchanged; *e.g.*, deer, sheep, moose, grouse. Five of the seven personal pronouns have distinctive forms for subject and object.

The forms of verbs are not complex. Only the substantive verb ("to be") has eight forms: be, am, is, are, was, were, being, been. Strong verbs have five forms: ride, rides, rode, riding, ridden. Regular or weak verbs customarily have four: walk, walks, walked, walking. Some that end in a *t* or *d* have three forms only: cut, cuts, cutting. Of these three-form verbs, 16 are in frequent use.

In addition to the above inflections, English employs two other main morphological (structural) processes—affixation and composition—and two subsidiary ones—back-formation and blend.

Types of affixes

Affixation. Affixes, word elements attached to words, may either precede, as prefixes (do, undo; way, subway), or follow, as suffixes (do, doer; way, wayward). They may be native (*over*do, waywardness), Greek (*hyper*bole, the*sis*), or Latin (*super*sede, ped*iment*). Modern technologists greatly favour the neo-Hellenic prefixes *macro-* "long, large," *micro-* "small," *para-* "alongside," *poly-* "many," and the Latin *mini-*, with its antonym *maxi-*. Greek and Latin affixes have become so fully acclimatized that they can occur together in one and the same word, as, indeed, in "ac-climat-ize-d," just used, consisting of a Latin prefix plus a Greek stem plus a Greek suffix plus an English inflection. Suffixes are bound more closely than prefixes to the stems or root elements of words. Consider, for instance, the wide variety of agent suffixes in the nouns act*or*, artis*an*, dot*ard*, engin*eer*, financ*ier*, hire*ling*, magistr*ate*, merch*ant*, scient*ist*, secret*ary*, song*ster*, stud*ent*, and work*er*. Suffixes may come to be attached to stems quite fortuitously, but, once attached, they are likely to be permanent. At the same time, one suffix can perform many functions. The suffix *-er* denotes the doer of the action in the words worker, driver, and hunter; the instrument in chopper, harvester, and roller; and the dweller in Icelander, Londoner, and Trobriander. It refers to things or actions associated with the basic concept in the words breather, "pause to take breath"; diner, "dining car on a train"; and fiver, "five-pound note." In the terms disclaimer, misnomer, and rejoinder (all from French) the suffix denotes one single instance of the action expressed by the verb. Usage may prove capricious. Whereas a writer is a person, a typewriter is a machine. For some time a computer was both, but now, with the invention and extensive use of electronic apparatus, the word is no longer used of persons.

Composition. Composition, or compounding, is concerned with free forms. The primary compounds "already," "cloverleaf," and "gentleman" show the collocation of two free forms. They differ from word groups or phrases in phonology, stress, or juncture or by a combination of two or more of these. Thus, "already" differs from "all ready" in stress and juncture, "cloverleaf" from "clover leaf" in stress, and "gentleman" from "gentle man" in phonology, stress, and juncture. In describing the structure of compound words it is necessary to take into account the relation of components to each other and the relation of the whole compound to its components. These relations diverge widely in, for example, the words cloverleaf, icebreaker, breakwater, blackbird, peace-loving, and paperback. In "cloverleaf" the first component noun is attributive and modifies the second, as also in the terms aircraft, beehive, landmark, lifeline, network, and vineyard. "Icebreaker," however, is a compound made up of noun object plus agent noun, itself consisting of verb plus agent suffix, as also in the words bridgebuilder, landowner, metalworker, minelayer, and timekeeper. The next type consists of verb plus object. It is rare in English, Dutch, and German but frequent in French, Spanish, and Italian. The English "pastime" may be compared, for example, with French *passe-temps*, the Spanish *pasatiempo*, and the Italian *passatempo*. From French comes "passport," meaning "pass (*i.e.*, enter) harbour." From Italian comes "portfolio," meaning "carry leaf." Other words of this type are daredevil, scapegrace, and scarecrow. As for the "blackbird" type, consisting of attributive adjective plus noun, it occurs frequently, as in the terms bluebell, grandson, shorthand, and wildfire. The next type, composed of object noun and a present participle, as in the terms fact-finding, heart-rending (German *herzzerreissend*), life-giving (German *lebenspendend*), painstaking, and time-consuming, occurs rarely. The last type is seen in barefoot, bluebeard, hunchback, leatherneck, redbreast, and scatterbrain.

Back-formations and blends. Back-formations and blends are becoming increasingly popular. Back-formation is the reverse of affixation, being the analogical creation of a new word from an existing word falsely assumed to be its derivative. For example, the verb "to edit" has been formed from the noun "editor" on the reverse analogy of the noun "actor" from "to act," and similarly the verbs automate, bulldoze, commute, escalate, liaise, loaf, sightsee, and televise are backformed from the nouns automation, bulldozer, commuter, escalation, liaison, loafer, sightseer, and television. From the single noun "procession" are backformed two verbs with different stresses and meanings: procéss, "to walk in procession," and prócess, "to subject food (and other material) to a special operation."

Two types of blends

Blends fall into two groups: (1) coalescences, such as "bash" from "bang" and "smash"; and (2) telescoped forms, called portmanteau words, such as "motorcade" from "motor cavalcade." In the first group are the words clash, from clack and crash, and geep, offspring of goat and sheep. To the second group belong dormobiles, or dormitory automobiles, and slurbs, or slum suburbs. A travel monologue becomes a travelogue and a telegram sent by cable a cablegram. Aviation electronics becomes avionics; biology electronics, bionics; and nuclear electronics, nucleonics. In cablese a question mark is a quark; in computerese a binary unit is a bit. In astrophysics a quasistellar source of radio energy becomes a quasar, and a pulsating star becomes a pulsar.

Simple shortenings, such as "ad" for "advertisement," have risen in status. They are listed in dictionaries side by side with their full forms. Among such fashionable abbreviations are exam, gym, lab, lib, op, spec, sub, tech, veg, and vet. Compound shortenings, after the pattern of Russian *agitprop* for *agitatsiya propaganda*, are also becoming fashionable. Initial syllables are joined as in the words Fortran, for formula (computer) translation; mascon, for massive (lunar) concentration; and Tacomsat, for Tactical Communications Satellite.

Syntax. Sentences can be classified as (1) simple, containing one clause and predication: "John knows this country"; (2) multiple or compound, containing two or more coordinate clauses: "John has been here before, and he knows this country"; and (3) complex, containing one or more main clauses and one or more subordinate clauses: "John, who has been here before, knows this country" or "Because he has been here before, John knows this country." Simple, declarative, affirmative sentences have two main patterns with five subsidiary patterns within each. Verb and complement together

form the predicate. "Complement" is here used to cover both the complement and the object of traditional grammarians (see Table 1).

Table 1: Simple Sentences—First Pattern

subject	verb	complement
1. John	knows	this country
2. Science	is	organized knowledge
3. Elizabeth	becomes	queen
4. The captain	falls	sick
5. Nothing	passes	unobserved

In (1) the complement is the direct object of a transitive verb; in (2) it is a predicative nominal group forming the second component of an equation linked to the first part by the meaningless copula is; in (3) it is a predicative noun linked with the subject by the meaningful copula becomes; in (4) it is a predicative adjective; and in (5) it is a predicative past participle.

Table 2: Simple Sentences—Second Pattern

subject	verb	inner complement	outer complement
6. John	gives	Mary	a ring
7. The sailors	make	John	captain
8. You	have kept	your record	clean
9. The driver	finds	the road	flooded
10. We	want	you	to know

In Table 2 each sentence contains four components: subject, verb, and two complements, first and second, or inner and outer. In (6) inner and outer complements consist of indirect object (without preposition) followed by direct object; in (7) these complements are direct object and appositive noun; in (8) direct object and predicative adjective; in (9) direct object and predicative past participle; in (10) direct object and predicative infinitive.

One can seldom change the word order in these 10 sentences without doing something else—adding or subtracting a word, changing the meaning. There is no better way of appreciating the importance of word position than by scrutinizing the 10 frames illustrated. If, for instance, in (6) one reverses inner and outer complements, one adds "to" and says, "John gives a ring to Mary"; one does not say "John gives a ring Mary." Some verbs, such as "explain" and "say," never omit the preposition "to" before the indirect object: "John's father explained the details to his son." "He said many things to him." If, in (10), the inner and outer complements are reversed (*e.g.*, "We want to know you"), the meaning is changed as well as the structure.

Positions of adjectives

Apart from these fundamental rules of word order, the principles governing the positions of adjectives, adverbs, and prepositions call for brief comment. For attributive adjectives the rule is simple: single words regularly precede the noun, and word groups follow—*e.g.*, "an unforgettable experience" but "an experience never to be forgotten." There is a growing tendency, however, to abandon this principle, to switch groups to front position, and to say "a never to be forgotten experience." In the ordering of multiple epithets, on the other hand, some new principles are seen to be slowly emerging. Attributes denoting permanent qualities stand nearest their head nouns: "long, white beard," "six-lane elevated freeway." The order in multiple attribution tends to be as follows: determiner; quantifier; adjective of quality; adjective of size, shape, or texture; adjective of colour or material; noun adjunct (if any); head noun. Examples include: "that one solid, round, oak dining table," "these many fine, large, black race horses," "those countless memorable, long, bright summer evenings."

Adverbs are more mobile than adjectives. Nevertheless, some tentative principles seem to be at work. Adverbs of frequency tend to come immediately after the substantive verb ("You are often late"), before other verbs ("You never know"), and between auxiliaries and full verbs ("You can never tell"). In this last instance, however, American differs from British usage. Most Americans would place the adverb before the auxiliary and say "You never can tell." (In the title of his play of that name, first performed in 1899, George Bernard Shaw avowedly followed American usage.) Adverbs of time usually occur at the beginning or end of a sentence, seldom in the middle. Particular expressions normally precede more general ones: "Neil Armstrong set foot on the Moon at 4 o'clock in the morning on July 21, 1969." An adverb of place or direction follows a verb with which it is semantically bound: "We arrived home after dark." Other adverbs normally take end positions in the order of manner, place, and time: "Senator Smith summed it all up most adroitly [manner] in Congress [place] last night [time]."

In spite of its etymology (Latin *prae-positio* "before placing"), a preposition may sometimes follow the noun it governs, as in "all the world over," "the clock round," and "the whole place through." "This seems a good place to live in" seems more natural to most speakers than "This seems a good place in which to live." "Have you anything to open this can with?" is now more common than "Have you anything with which to open this can?"

The above are principles rather than rules, and in the end it must be agreed that English syntax lacks regimentation. Its structural laxity makes English an easy language to speak badly. It also makes English prone to ambiguity. "When walking snipe always approach up wind," a shooting manual directs. The writer intends the reader to understand, "When you are walking to flush snipe always approach them up against the wind." "John kept the car in the garage" can mean either (1) "John retained that car you see in the garage, and sold his other one" or (2) "John housed the car in the garage, and not elsewhere." "Flying planes can be dangerous" is ambiguous because it may mean either (1) "Planes that fly can be dangerous" or (2) "It is dangerous to fly planes."

Two ways in which "John gives Mary a ring" can be stated in the passive are: (1) "A ring is given to Mary by John" and (2) "Mary is given a ring by John." Concerning this same action, four types of question can be formulated: (1) "Who gives Mary a ring?" The information sought is the identity of the giver. (2) "Does John give Mary a ring?" The question may be answered by "yes" or "no." (3) "John gives Mary a ring, doesn't he?" Confirmation is sought of the questioner's belief that John does in fact give Mary a ring. (4) "John gives Mary a ring?" This form, differing from the declarative statement only by the question mark in writing, or by rising intonation in speech, calls, like sentences (2) and (3), for a "yes" or "no" answer but suggests doubt on the part of the questioner that the action is taking place.

Origins of English words

Vocabulary. The vocabulary of Modern English is approximately half Germanic (Old English and Scandinavian) and half Italic or Romance (French and Latin), with copious and increasing importations from Greek in science and technology and with considerable borrowings from Dutch, Low German, Italian, Spanish, German, Arabic, and many other languages. Names of basic concepts and things come from Old English or Anglo-Saxon: heaven and earth, love and hate, life and death, beginning and end, day and night, month and year, heat and cold, way and path, meadow and stream. Cardinal numerals come from Old English, as do all the ordinal numerals except "second" (Old English *other*, which still retains its older meaning in "every other day"). "Second" comes from Latin *secundus* "following," through French *second*, related to Latin *sequi* "to follow," as in English "sequence." From Old English come all the personal pronouns (except "they," "their," and "them," which are from Scandinavian), the auxiliary verbs (except the marginal "used," which is from French), most simple prepositions, and all conjunctions.

Numerous nouns would be identical whether they came from Old English or Scandinavian: father, mother, brother (but not sister); man, wife; ground, land, tree, grass; summer, winter; cliff, dale. Many verbs would also be identical, especially monosyllabic verbs—bring, come,

get, hear, meet, see, set, sit, spin, stand, think. The same is true of the adjectives full and wise; the colour names gray, green, and white; the disjunctive possessives mine and thine (but not ours and yours); the terms north and west (but not south and east); and the prepositions over and under. Just a few English and Scandinavian doublets coexist in current speech: no and nay, yea and ay, from and fro, rear (*i.e.*, to bring up) and raise, shirt and skirt (both related to the adjective short), less and loose. From Scandinavian, "law" was borrowed early, whence "by-law," meaning "village law," and "outlaw," meaning "man outside the law." "Husband" (*hus-bondi*) meant "householder," whether single or married, whereas "fellow" (*fe-lagi*) meant one who "lays fee" or shares property with another, and so "partner, shareholder." From Scandinavian come the common nouns axle (tree), band, birth, bloom, crook, dirt, egg, gait, gap, girth, knife, loan, race, rift, root, score, seat, skill, sky, snare, thrift, and window; the adjectives awkward, flat, happy, ill, loose, rotten, rugged, sly, tight, ugly, weak, and wrong; and many verbs, including call, cast, clasp, clip, crave, die, droop, drown, flit, gape, gasp, glitter, lift, rake, rid, scare, scowl, skulk, snub, sprint, thrive, thrust, and want.

Borrowing from French

The debt of the English language to France is large. The terms president, representative, legislature, congress, constitution, and parliament are all French. So, too, are duke, marquis, viscount, and baron; but king, queen, lord, lady, earl, and knight are English. City, village, court, palace, manor, mansion, residence, and domicile are French; but town, borough, hall, house, bower, room, and home are English. Comparison between English and French synonyms shows that the former are more human and concrete, the latter more intellectual and abstract; *e.g.*, the terms freedom and liberty, friendship and amity, hatred and enmity, love and affection, likelihood and probability, truth and veracity, lying and mendacity. The superiority of French cooking is duly recognized by the adoption of such culinary terms as boil, broil, fry, grill, roast, souse, and toast. "Breakfast" is English, but "dinner" and "supper" are French. "Hunt" is English, but "chase," "quarry," "scent," and "track" are French. Craftsmen bear names of English origin: baker, builder, fisher (man), hedger, miller, shepherd, shoemaker, wainwright, and weaver, or webber. Names of skilled artisans, however, are French: carpenter, draper, haberdasher, joiner, mason, painter, plumber, and tailor. Many terms relating to dress and fashion, cuisine and viniculture, politics and diplomacy, drama and literature, art and ballet come from France.

In the spheres of science and technology many terms come from Classical Greek through French or directly from Greek. Pioneers in research and development now regard Greek as a kind of inexhaustible quarry from which they can draw linguistic material at will. By prefixing the Greek adverb *tēle* "far away, distant" to the existing compound photography, "light writing," they create the precise term "telephotography" to denote the photographing of distant objects by means of a special lens. By inserting the prefix *micro-* "small" into this same compound, they make the new term "photomicrography," denoting the electronic photographing of bacteria and viruses. Such neo-Hellenic derivatives would probably have been unintelligible to Plato and Aristotle.

Greek borrowings

Many Greek compounds and derivatives have Latin equivalents with slight or considerable differentiations in meaning (see Table 3).

At first sight it might appear that some of these equivalents, such as "metamorphosis" and "transformation," are sufficiently synonymous to make one or the other redundant. In fact, however, "metamorphosis" is more technical and therefore more restricted than "transformation." In mythology it signifies a magical shape changing; in nature it denotes a postembryonic development such as that of a tadpole into a frog, a cocoon into a silkworm, or a chrysalis into a butterfly. Transformation, on the other hand, means any kind of change from one state to another.

Ever since the 12th century, when merchants from the Netherlands made homes in East Anglia, Dutch words have infiltrated into Midland speech. For centuries a form of Low German was used by seafaring men in North Sea ports. Old nautical terms still in use include buoy, deck, dock, freebooter, hoist, leak, pump, skipper, and yacht. The Dutch in New Amsterdam (later New York) and adjacent settlements gave the words boss, cookie, dope, snoop, and waffle to American speech. The Dutch in Cape Province gave the terms apartheid, commandeer, commando, spoor, and trek to South African speech.

Table 3: Equivalent Compounds and Derivatives*

from the Greek	from the Latin
Nouns	
dys-troph-y	mal-nutr-it-ion
hypo-sta-sis	sub-stan-ce
hypo-the-sis	sup-pos-it-ion
meta-morph-o-sis	trans-form-at-ion
meta-phor	trans-fer
meta-the-sis	trans-pos-it-ion
peri-pher-y	circum-fer-en-ce
peri-phra-sis	circum-loc-ut-ion
sym-path-y	com-pass-ion
syn-drom-e	con-curr-en-ce
syn-op-sis	con-spect-us
syn-the-sis	com-pos-it-ion
sy-zyg-y	con-junc-t-ion
Adjectives	
dia-phan-*ous*	trans-par-ent
hyper-aesth-et-ic	super-sens-it-ive
hyper-phys-ic-*al*	super-nat-ur-al
hypo-derm-ic	sub-cut-an-eous
hypo-ge-*al*	sub-terr-an-ean
melan-chol-ic	atra-bil-ious
mono-morph-ic	uni-form
oxy-phyll-*ous*	acut-i-fol-i-ate
peri-pat-et-ic	circum-amb-ul-at-ory
phos-phor-*escent*	lumin-i-fer-ous
poly-glott-*al*	multi-lingu-al
sphen-oid	cunei-form
syn-chron-ic	con-temp-or-ary

*The italicized suffixes *-al*, *-escent*, and *-ous*, attached to some of the Greek adjectives, are of Latin origin.

German contributions

The contribution of High German has been on a different level. In the 18th and 19th centuries it lay in technicalities of geology and mineralogy and in abstractions relating to literature, philosophy, and psychology. In the 20th century this contribution has sometimes been indirect. "Unclear" and "meaningful" echoed German *unklar* and *bedeutungsvoll*, or *sinnvoll*. "Ring road" (a British term applied to roads encircling cities or parts of cities) translated *Ringstrasse;* "round trip," *Rundfahrt;* and "the turn of the century," *die Jahrhundertwende.* The terms "classless society," "inferiority complex," and "wishful thinking" echoed *die klassenlöse Gesellschaft, der Minderwertigkeitskomplex*, and *das Wunschdenken.*

Along with the rest of the Western world, English has accepted Italian as the language of music. The names of voices, parts, performers, instruments, forms of composition, and technical directions are all Italian. Many of the latter—allegro, andante, cantabile, crescendo, diminuendo, legato, maestoso, obbligato, pizzicato, staccato, and vibrato—are also used metaphorically. In architecture, the terms belvedere, corridor, cupola, grotto, pedestal, pergola, piazza, pilaster, and rotunda are accepted; in literature, burlesque, canto, extravaganza, stanza, and many more are used.

From Spanish, English has acquired the words armada, cannibal, cigar, galleon, guerrilla, matador, mosquito, quadroon, tornado, and vanilla, some of these loanwords going back to the 16th century, when sea dogs encountered hidalgos on the high seas. Many names of animals and plants have entered English from indigenous languages through Spanish: "potato" through Spanish *patata* from Taino *batata*, and "tomato" through Spanish *tomate* from Nahuatl *tomatl*. Other words have entered from Latin America by way of Texas, New Mexico, Arizona, and California; *e.g.*, such words as canyon, cigar, estancia, lasso, mustang, pueblo, and rodeo. Some have gathered new connotations: bonanza, originally de-

noting "goodness," came through miners' slang to mean "spectacular windfall, prosperity"; mañana, "tomorrow," acquired an undertone of mysterious unpredictability.

From Arabic through European Spanish, through French from Spanish, through Latin, or occasionally through Greek, English has obtained the terms alchemy, alcohol, alembic, algebra, alkali, almanac, arsenal, assassin, attar, azimuth, cipher, elixir, mosque, nadir, naphtha, sugar, syrup, zenith, and zero. From Egyptian Arabic, English has recently borrowed the term loofah (also spelled luffa). From Hebrew, directly or by way of Vulgate Latin, come the terms amen, cherub, hallelujah, manna, messiah, pharisee, rabbi, sabbath, and seraph; jubilee, leviathan, and shibboleth; and, more recently, kosher, and kibbutz.

English has freely adopted and adapted words from many other languages, acquiring them sometimes directly and sometimes by devious routes. Each word has its own history. The following lists indicate the origins of a number of English words: Welsh—flannel, coracle, cromlech, eisteddfod; Cornish—gull, brill, dolmen; Gaelic and Irish—shamrock, brogue, leprechaun, ogham, Tory, galore, blarney, hooligan, clan, claymore, bog, plaid, slogan, sporran, cairn, whisky, pibroch; Breton—menhir, penguin; Norwegian—ski, ombudsman; Finnish—sauna; Russian—kvass, ruble, tsar, verst, mammoth, ukase, astrakhan, vodka, samovar, tundra (from Lapp), troika, pogrom, duma, soviet, bolshevik, intelligentsia (from Latin through Polish), borscht, balalaika, sputnik, soyuz, salyut, lunokhod; Polish—mazurka; Czech—robot; Hungarian—goulash, paprika; Portuguese—marmalade, flamingo, molasses, veranda, port (wine), dodo; Basque—bizarre; Turkish—janissary, turban, coffee, kiosk, caviar, pasha, odalisque, fez, bosh; Hindi—nabob, guru, sahib, maharajah, mahatma, pundit, punch (drink), juggernaut, cushy, jungle, thug, cheetah, shampoo, chit, dungaree, pucka, gymkhana, mantra, loot, pajamas, dinghy, polo; Persian—paradise, divan, purdah, lilac, bazaar, shah, caravan, chess, salamander, taffeta, shawl, khaki; Tamil—pariah, curry, catamaran, mulligatawny; Chinese—tea (Amoy), sampan; Japanese—shogun, kimono, mikado, tycoon, hara-kiri, gobang, judo, jujitsu, bushido, samurai, banzai, tsunami, satsuma, No (the dance drama), karate, Kabuki; Malay—ketchup, sago, bamboo, junk, amuck, orangutan, compound (fenced area), raffia; Polynesian—taboo, tattoo; Hawaiian—ukulele; African languages—chimpanzee, goober, mumbo jumbo, voodoo; Eskimo—kayak, igloo, anorak, mukluk; Algonkian—totem; Nahuatl—mescal; languages of the Caribbean—hammock, hurricane, tobacco, maize, iguana; Aboriginal Australian—kangaroo, corroboree, wallaby, wombat, boomerang, paramatta, budgerigar.

The alphabet

Orthography. The Latin alphabet originally had 20 letters, the present English alphabet minus *J*, *K*, *V*, *W*, *Y*, and *Z*. The Romans themselves added *K* for use in abbreviations and *Y* and *Z* in words transcribed from Greek. After its adoption by the English, this 23-letter alphabet developed *W* as a ligatured doubling of *U* and later *J* and *V* as consonantal variants of *I* and *U*. The resultant alphabet of 26 letters has both uppercase, or capital, and lowercase, or small, letters.

English spelling is based for the most part on that of the 15th century, but pronunciation has changed considerably since then, especially that of long vowels and diphthongs. The extensive change in the pronunciation of vowels, known as the Great Vowel Shift, affected all of Geoffrey Chaucer's seven long vowels, and for centuries spelling remained untidy. If the meaning of the message was clear, the spelling of individual words seemed unimportant. In the 17th century during the English Civil War, compositors adopted fixed spellings for practical reasons, and in the order-loving 18th century uniformity became more and more fashionable. Since Samuel Johnson's *Dictionary of the English Language* (1755), orthography has remained fairly stable. Numerous tacit changes, such as "music" for "musick" (*c.* 1880) and "fantasy" for "phantasy" (*c.* 1920), have been accepted, but spelling has nevertheless continued to be in part unphonetic. Attempts have been made at reform. Indeed, every century has had its reformers since the 13th, when an Augustinian canon named Orm devised his own method of differentiating short vowels from long by doubling the succeeding consonants or, when this was not feasible, by marking short vowels with a superimposed breve mark (◡). William Caxton, who set up his wooden printing press at Westminster in 1476, was much concerned with spelling problems throughout his working life. Noah Webster produced his *Spelling Book*, in 1783, as a precursor to the first edition (1828) of his *American Dictionary of the English Language*. The 20th century has produced many zealous reformers. Three systems, supplementary to traditional spelling, are actually in use for different purposes: (1) the Initial Teaching (Augmented Roman) Alphabet (ITA) of 44 letters used by educationists in the teaching of children under seven; (2) the Shaw alphabet of 48 letters, designed in implementation of the will of George Bernard Shaw; and (3) the International Phonetic Alphabet (IPA), constructed on the basis of one symbol for one individual sound and used by many trained linguists. Countless other systems have been worked out from time to time, of which R.E. Zachrisson's "Anglic" (1930) and Axel Wijk's *Regularized English* (1959) may be the best.

Obstacles to spelling reform

Meanwhile, the great publishing houses continue unperturbed because drastic reform remains impracticable, undesirable, and unlikely. This is because there is no longer one criterion of correct pronunciation but several standards throughout the world; regional standards are themselves not static, but changing with each new generation; and, if spelling were changed drastically, all the books in English in the world's public and private libraries would become inaccessible to readers without special study.

HISTORICAL BACKGROUND

Among highlights in the history of the English language, the following stand out most clearly: the settlement in Britain of Jutes, Saxons, and Angles in the 5th and 6th centuries; the arrival of St. Augustine in 597 and the subsequent conversion of England to Latin Christianity; the Viking invasions of the 9th century; the Norman Conquest of 1066; the Statute of Pleading in 1362 (this required that court proceedings be conducted in English); the setting up of Caxton's printing press at Westminster in 1476; the full flowering of the Renaissance in the 16th century; the publishing of the King James Bible in 1611; the completion of Johnson's *Dictionary* of 1755; and the expansion to North America and South Africa in the 17th century and to India, Australia, and New Zealand in the 18th.

Old English. The Jutes, Angles, and Saxons lived in Jutland, Schleswig, and Holstein, respectively, before settling in Britain. According to the Venerable Bede, the first historian of the English people, the first Jutes, Hengist and Horsa, landed at Ebbsfleet in the Isle of Thanet in 449; and the Jutes later settled in Kent, southern Hampshire, and the Isle of Wight. The Saxons occupied the rest of England south of the Thames, as well as modern Middlesex and Essex. The Angles eventually took the remainder of England as far north as the Firth of Forth, including the future Edinburgh and the Scottish Lowlands. In both Latin and Common Germanic the Angles' name was Angli, later mutated in Old English to Engle (nominative) and Engla (genitive). "Engla land" designated the home of all three tribes collectively, and both King Alfred (known as Alfred the Great) and Abbot Aelfric, author and grammarian, subsequently referred to their speech as Englisc. Nevertheless, all the evidence indicates that Jutes, Angles, and Saxons retained their distinctive dialects.

Four Old English dialects

The River Humber was an important boundary, and the Anglian-speaking region developed two speech groups: to the north of the river, Northumbrian, and, to the south, Southumbrian, or Mercian. There were thus four dialects: Northumbrian, Mercian, West Saxon, and Kentish (see Figure 1). In the 8th century, Northumbrian led in literature and culture, but that leadership was destroyed by the Viking invaders, who sacked Lindisfarne,

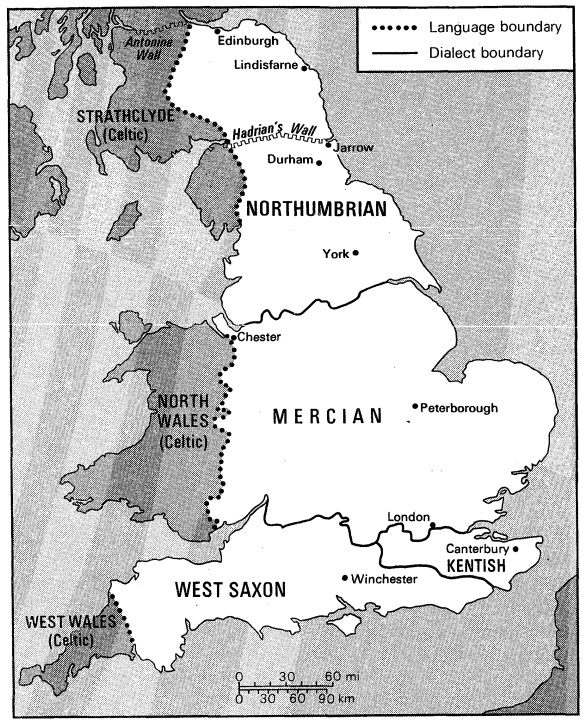


Figure 1: Old English dialects.

an island near the Northumbrian mainland, in 793. They landed in strength in 865. The first raiders were Danes, but they were later joined by Norwegians from Ireland and the Western Isles who settled in modern Cumberland, Westmorland, northwest Yorkshire, Lancashire, north Cheshire, and the Isle of Man. In the 9th century, as a result of the Norwegian invasions, cultural leadership passed from Northumbria to Wessex. During King Alfred's reign, in the last three decades of the 9th century, Winchester became the chief centre of learning. There the Parker Chronicle (a manuscript of the Anglo-Saxon Chronicle) was written; there the Latin works of the priest and historian Paulus Orosius, St. Augustine, St. Gregory, and the Venerable Bede were translated; and there the native poetry of Northumbria and Mercia was transcribed into the West Saxon dialect. This resulted in West Saxon's becoming "standard Old English"; and later, when Aelfric (*c.* 955–*c.* 1010) wrote his lucid and mature prose at Winchester, Cerne Abbas, and Eynsham, the hegemony of Wessex was strengthened.

In standard Old English, adjectives were inflected as well as nouns, pronouns, and verbs. Nouns were inflected for four cases (nominative, genitive, dative, and accusative) in singular and plural. Five nouns of first kinship—*faeder*, *mōdor*, *brōthor*, *sweostor*, and *dohtor* ("father," "mother," "brother," "sister," and "daughter," respectively)—had their own set of inflections. There were 25 nouns such as *mon*, *men* ("man," "men") with mutated, or umlauted, stems. Adjectives had strong and weak declensions, the strong showing a mixture of noun and pronoun endings and the weak following the pattern of weak nouns. Personal, possessive, demonstrative, interrogative, indefinite, and relative pronouns had full inflections. The pronouns of the 1st and 2nd persons still had distinctive dual forms:

iċ	"I"	*wit*	"we two"	*wē*	"we"
thū (*þū*)	"thou"	*ġit*	"you two"	*ġē*	"you"

There were two demonstratives: *sē*, *sēo*, *thaet*, meaning "that," and *thes*, *thēos*, *this*, meaning "this," but no articles, the definite article being expressed by use of the demonstrative for "that" or not expressed at all. Thus, "the good man" was *sē gōda mon* or plain *gōd mon*. The function of the indefinite article was performed by the numeral *ān* "one" in *ān mon* "a man," by the adjective–pronoun *sum* in *sum mon* "a (certain) man," or not expressed, as in *thū eart gōd mon* "you are a good man."

Verbs had two tenses only (present–future and past), three moods (indicative, subjunctive, and imperative), two numbers (singular and plural), and three persons (1st, 2nd, and 3rd). There were two classes of verb stems. (A verb stem is that part of a verb to which inflectional changes—changes indicating tense, mood, number, etc.—are added.) One type of verb stem, called vocalic because an internal vowel shows variations, is exemplified by the verb for "sing": *singan*, *singth*, *sang*, *sungon*, *gesungen*. The word for "deem" is an example of the other, called consonantal: *dēman*, *dēmth*, *dēmde*, *dēmdon*, *gedēmed*. Such verbs are called strong and weak, respectively. Inflection of verbs

All new verbs, whether derived from existing verbs or from nouns, belonged to the consonantal type. Some verbs of great frequency (antecedents of the modern words "be," "shall," "will," "do," "go," "can," "may," and so on) had their own peculiar patterns of inflections.

Grammatical gender persisted throughout the Old English period. Just as Germans now say *der Fuss*, *die Hand*, and *das Auge* (masculine, feminine, and neuter terms for "the foot," "the hand," and "the eye"), so, for these same structures, Aelfric said *sē fōt*, *sēo hond*, and *thaet ēaġe*, also masculine, feminine, and neuter. The three words for "woman," *wīfmon*, *cwene*, and *wīf*, were masculine, feminine, and neuter, respectively. *Hors* "horse," *sċēap* "sheep," and *maeġden* "maiden" were all neuter. *Eorthe* "earth" was feminine, but *lond* "land" was neuter. *Sunne* "sun" was feminine, but *mōna* "moon" was masculine. This simplification of grammatical gender resulted from the fact that the gender of Old English substantives was not always indicated by the ending but rather by the terminations of the adjectives and demonstrative pronouns used with the substantives. When these endings were lost, all outward marks of gender disappeared with them. Thus, the weakening of inflections and loss of gender occurred together. In the North, where inflections weakened earlier, the marks of gender likewise disappeared first. They survived in the South as late as the 14th century.

Because of the greater use of inflections in Old English, word order was freer than today. The sequence of subject, verb, and complement was normal, but when there were outer and inner complements the second was put in the dative case after *tō*: *Sē biscop hālgode Ēadrēd tō cyninge* "The bishop consecrated Edred king." After an introductory adverb or adverbial phrase the verb generally took second place as in modern German: *Nū bydde iċ ān thing* "Now I ask [literally, "ask I"] one thing"; *Thȳ ilcan gēare gesette Aelfrēd cyning Lundenburg* "In that same year Alfred the king occupied London." Impersonal verbs had no subject expressed. Infinitives constructed with auxiliary verbs were placed at the ends of clauses or sentences: *Hīe ne dorston forth bī thǣre ēa siglan* "They dared not sail beyond that river" (*siglan* is the infinitive); *Iċ wolde thās lytlan bōc āwendan* "I wanted to translate this little book" (*āwendan* is the infinitive). The verb usually came last in a dependent clause—*e.g.*, *āwrītan wile* in *gif hwā thās bōc āwrītan wile* (*gerihte hē hīe be thǣre bysene*) "If anyone wants to copy this book (let him correct his copy by the original)." Prepositions (or postpositions) frequently followed their objects. Negation was often repeated for emphasis.

Middle English. One result of the Norman Conquest of 1066 was to place all four Old English dialects more or less on a level. West Saxon lost its supremacy and the centre of culture and learning gradually shifted from Winchester to London. The old Northumbrian dialect became divided into Scottish and Northern, although little is known of either of these divisions before the end of the 13th century (Figure 2). The old Mercian dialect was split into East and West Midland. West Saxon became slightly diminished in area and was more appropriately named the South Western dialect. The Kentish dialect was considerably extended and was called South Eastern accordingly. All five Middle English dialects (Northern, West Midland, East Midland, South Western, and South Eastern) went their own ways and developed their own characteristics. The so-called Katherine Group of writ-

ings (1180–1210), associated with Hereford, a town not far from the Welsh border, adhered most closely to native traditions, and there is something to be said for regarding this West Midland dialect, least disturbed by French and Scandinavian intrusions, as a kind of Standard English in the High Middle Ages.

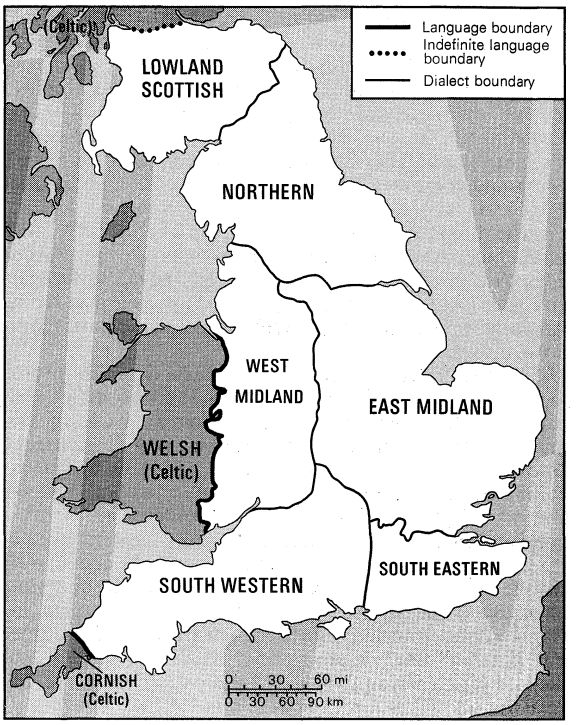


Figure 2: Middle English dialects.

Another outcome of the Norman Conquest was to change the writing of English from the clear and easily readable insular hand of Irish origin to the delicate Carolingian script then in use on the Continent. With the change in appearance came a change in spelling. Norman scribes wrote Old English *y* as *u*, *ȳ* as *ui*, *ū* as *ou* (*ow* when final). Thus, *mycel* ("much") appeared as *muchel*, *fȳr* ("fire") as *fuir*, *hūs* ("house") as *hous*, and *hū* ("how") as *how*. For the sake of clarity (*i.e.*, legibility) *u* was often written *o* before and after *m*, *n*, *u*, *v*, and *w*; and *i* was sometimes written *y* before and after *m* and *n*. So *sunu* ("son") appeared as *sone* and *him* ("him") as *hym*. Old English *cw* was changed to *qu*; *hw* to *wh*, *qu*, or *quh*; *ċ* to *ch* or *tch*; *sċ* to *sh*; *-ċġ-* to *-gg-*; and *-ht* to *ght*. So Old English *cwēn* appeared as *queen*; *hwaet* as *what*, *quat*, or *quhat*; *dīċ* as *ditch*; *sċip* as *ship*; *secge* as *segge*; and *miht* as *might*.

Norman changes in spelling

For the first century after the Conquest, most loanwords came from Normandy and Picardy, but with the extension south to the Pyrenees of the Angevin empire of Henry II (reigned 1154–89), other dialects, especially Central French, or Francien, contributed to the speech of the aristocracy. As a result, Modern English acquired the forms canal, catch, leal, real, reward, wage, warden, and warrant from Norman French side by side with the corresponding forms channel, chase, loyal, royal, regard, gage, guardian, and guarantee, from Francien. King John lost Normandy in 1204. With the increasing power of the Capetian kings of Paris, Francien gradually predominated. Meanwhile, Latin stood intact as the language of learning. For three centuries, therefore, the literature of England was trilingual. *Ancrene Riwle*, for instance, a guide or rule (*riwle*) of rare quality for recluses or anchorites (*ancren*), was disseminated in all three languages.

The sounds of the native speech changed slowly. Even in late Old English short vowels had been lengthened before *ld*, *rd*, *mb*, and *nd*, and long vowels had been shortened before all other consonant groups and before double consonants. In early Middle English short vowels of whatever origin were lengthened in the open stressed syllables of disyllabic words. An open syllable is one ending in a vowel. Both syllables in Old English *nama* "name," *mete* "meat, food," *nosu* "nose," *wicu* "week," and *duru* "door" were short, and the first syllables, being stressed, were lengthened to *nāme*, *mēte*, *nōse*, *wēke*, and *dōre* in the 13th and 14th centuries. A similar change occurred in 4th-century Latin, in 13th-century German, and at different times in other languages. The popular notion has arisen that final mute *-e* in English makes a preceding vowel long; in fact, it is the lengthening of the vowel that has caused *e* to be lost in pronunciation. On the other hand, Old English long vowels were shortened in the first syllables of trisyllabic words, even when those syllables were open; *e.g.*, *hāligdaeg* "holy day," *ārende* "message, errand," *crīstendōm* "Christianity," and *sūtherne* "southern," became *hŏliday* (Northern *hăliday*), *ĕrrende*, *chrĭstendom*, and *sŭtherne*. This principle still operates in current English. Compare, for example, trisyllabic derivatives such as the words chastity, criminal, fabulous, gradual, gravity, linear, national, ominous, sanity, and tabulate with the simple nouns and adjectives chaste, crime, fable, grade, grave, line, nation, omen, sane, and table.

There were significant variations in verb inflections in the Northern, Midland, and Southern dialects as shown in Table 4. The Northern infinitive was already one syllable (*sing* rather than the Old English *singan*), whereas the past participle *-en* inflection of Old English was strictly kept. These apparently contradictory features can be attributed entirely to Scandinavian, in which the final *-n* of the infinitive was lost early in *singa*, and the final *-n* of the past participle was doubled in *sunginn*. The Northern unmutated present participle in *-and* was also of Scandinavian origin. Old English mutated *-ende* (German *-end*) in the present participle had already become *-inde* in late West Saxon (Southern in the accompanying table), and it was this Southern *-inde* that blended with the *-ing* suffix (German *-ung*) of nouns of action that had already become near-gerunds in such compound nouns as *athswering* "oath swearing" and *writingfether* "writing feather, pen." This blending of present participle and gerund was further helped by the fact that Anglo-Norman and French *-ant* was itself a coalescence of Latin present participles in *-antem*, *-entem*, and Latin gerunds in *-andum*, *-endum*. The Northern 2nd person singular *singis* was inherited unchanged from Common Germanic. The final *t* sound in Midland *-est* and Southern *-st* was excrescent, comparable with the final *t* in modern "amidst" and "amongst" from older *amiddes* and *amonges*. The Northern 3rd person singular *singis* had a quite different origin. Like the *singis* of the plural, it resulted almost casually from an inadvertent retraction of the tongue in enunciation from an interdental *-th* sound to postdental *-s*. Today the form "singeth" survives as a poetic archaism. Shakespeare used both *-eth* and *-s* endings ("It [mercy] blesseth him that gives and him that takes," *The Merchant of Venice*). The Midland present plural inflection *-en* was taken from the subjunctive. The past participle prefix *y-* developed from the Old English perfective prefix *ge-*.

Table 4: Variations in Verb Inflections

	Northern	Midland	Southern
Infinitive	sing	singe(n)	singen
Present participle	singand	singende	singinde
Present singular			
1st person	singe	singe	singe
2nd person	singis	singes(t)	singst
3rd person	singis	singeth-es	singeth
Present plural	singis	singen	singeth
Past participle	sungen	(y)sunge(n)	ysunge

Chaucer's language

Chaucer, who was born and died in London, spoke a dialect that was basically East Midland. Compared with his contemporaries, he was remarkably modern in his use of language. He was in his early 20s when the Statute of Pleading (1362) was passed, by the terms of which all

court proceedings were henceforth to be conducted in English, though "enrolled in Latin." Chaucer himself used four languages; he read Latin (Classical and Medieval) and spoke French and Italian on his travels. For his own literary work he deliberately chose English.

Transition from Middle English to Early Modern English. The death of Chaucer at the close of the century (1400) marked the beginning of the period of transition from Middle English to the Early Modern English stage. The Early Modern English period is regarded by many scholars as beginning in about 1500 and terminating with the return of the monarchy (John Dryden's *Astraea Redux*) in 1660. The 15th century witnessed three outstanding developments: the rise of London English, the invention of printing, and the spread of the new learning.

The speech of London

Although the population of London in 1400 was only about 40,000, it was by far the largest city in England. York came second, followed by Bristol, Coventry, Plymouth, and Norwich. The Midlands and East Anglia, the most densely peopled parts of England, supplied London with streams of young immigrants. The speech of the capital was mixed, and it was changing. The seven long vowels of Chaucer's speech had already begun to shift. Incipient diphthongization of high front /i:/ (the *ee* sound in "meet") and high back /u:/ (as in "fool") led to instability in the other five long vowels. (Symbols within slash marks are taken from the International Phonetic Alphabet.) This remarkable event, known as the Great English Vowel Shift, changed the whole vowel system of London English. As /i:/ and /u:/ became diphthongized to /ai/ (as in "bide") and /au/ (as in "house") respectively, so the next highest vowels, /e:/ (this sound can be heard in the first part of the diphthong in "name") and /o:/ (a sound that can be heard in the first part of the diphthong in "home"), moved up to take their places, and so on. The whole process is summarized in Table 5.

Table 5: Vowel Shifts in London English

Chaucer's spelling	Chaucer's pronunciation*	Shakespeare's pronunciation	present pronunciation*	present spelling
lyf	li:f	leif	laif	life
deed	de:d	di:d	di:d	deed
deel	de:l	de:l	di:l	deal
name	na:mə	nɛ:m	neim	name
hoom	hɔ:m	ho:m	houm	home
mone	mo:n	mu:n	mu:n	moon
hous	hu:s	hous	haus	house

*Expressed in the International Phonetic Alphabet.

Every one of the sounds appearing in this table can still be heard somewhere in living English dialects.

When Caxton started printing at Westminster in the late summer of 1476, he was painfully aware of the uncertain state of the English language. In his prologues and epilogues to his translations he made some revealing observations on the problems that he had encountered as translator and editor. At this time, sentence structures were being gradually modified, but many remained untidy. For the first time, nonprofessional scribes, including women, were writing at length.

The revival of classical learning was one aspect of that Renaissance, or spiritual rebirth, that arose in Italy and spread to France and England. It evoked a new interest in Greek on the part of learned men such as William Grocyn and Thomas Linacre, Sir Thomas More and Desiderius Erasmus. John Colet, dean of St. Paul's in the first quarter of the 16th century, startled his congregation by expounding the Pauline Epistles of the New Testament as living letters. The deans who had preceded him had known no Greek, because they had found in Latin all that they required. Only a few medieval churchmen, such as Robert Grosseteste, bishop of Lincoln, and the Franciscan Roger Bacon could read Greek with ease. The names of the seven liberal arts of the medieval curricula (the trivium and the quadrivium), it is true, were all Greek—grammar, logic, and rhetoric; arithmetic, geometry, astronomy, and music—but they had come into English by way of French.

Renaissance scholars adopted a liberal attitude to language. They borrowed Latin words through French, or Latin words direct; Greek words through Latin, or Greek words direct. Latin was no longer limited to Church Latin: it embraced all Classical Latin. For a time the whole Latin lexicon became potentially English. Some words, such as consolation and infidel, could have come from either French or Latin. Others, such as the terms abacus, arbitrator, explicit, finis, gratis, imprimis, item, memento, memorandum, neuter, simile, and videlicet, were taken straight from Latin. Words that had already entered the language through French were now borrowed again, so that doublets arose: benison and benediction; blame and blaspheme; chance and cadence; count and compute; dainty and dignity; frail and fragile; poor and pauper; purvey and provide; ray and radius; sever and separate; strait and strict; sure and secure. The Latin adjectives for "kingly" and "lawful" have even given rise to triplets; in the forms real, royal, and regal and leal, loyal, and legal, they were imported first from Anglo-Norman, then from Old French, and last from Latin direct.

After the dawn of the 16th century, English prose moved swiftly toward modernity. In 1525 Lord Berners completed his translation of Jean Froissart's *Chronicle*, and William Tyndale translated the New Testament. One-third of the King James Bible (1611), it has been computed, is worded exactly as Tyndale left it; and between 1525 and 1611 lay the Tudor Golden Age, with its culmination in Shakespeare. Too many writers, to be sure, used "inkhorn terms," newly-coined, ephemeral words, and too many vacillated between Latin and English. Sir Thomas More actually wrote his *Utopia* in Latin. It was translated into French during his lifetime but not into English until 1551, some years after his death. Francis Bacon published *De dignitate et augmentis scientiarum* (*On the Dignity and Advancement of Learning*, an expansion of his earlier *Advancement of Learning*) in Latin in 1623. William Harvey announced his epoch-making discovery of the circulation of the blood in his Latin *De Motu Cordis et Sanguinis in Animalibus* (1628; *On the Motion of the Heart and Blood in Animals*). John Milton composed polemical treatises in the language of Cicero. As Oliver Cromwell's secretary, he corresponded in Latin with foreign states. His younger contemporary Sir Isaac Newton lived long enough to bridge the gap. He wrote his *Principia* (1687) in Latin but his *Opticks* (1704) in English.

Plans to regulate English

Restoration period. With the restoration of the monarchy in 1660, men again looked to France. John Dryden admired the Académie Française and greatly deplored that the English had "not so much as a tolerable dictionary, or a grammar; so that our language is in a manner barbarous" as compared with elegant French. After the passionate controversies of the Civil War, this was an age of cool scientific nationalism. In 1662 the Royal Society of London for the Promotion of Natural Knowledge received its charter. Its first members, much concerned with language, appointed a committee of 22 "to improve the English tongue particularly for philosophic purposes." It included Dryden, the diarist John Evelyn, Bishop Thomas Sprat, and the poet Edmund Waller. Sprat pleaded for "a close, naked, natural way of speaking; positive expressions; clear senses, a native easiness; bringing all things as near the mathematical plainness" as possible. The committee, however, achieved no tangible result, and failed in its attempt to found an authoritative arbiter over the English tongue. A second attempt was made in 1712, when Jonathan Swift addressed an open letter to Robert Harley, earl of Oxford, then Lord Treasurer, making "A Proposal for Correcting, Improving, and Ascertaining [fixing] the English Tongue." This letter received some popular support, but its aims were frustrated by a turn in political fortunes. Queen Anne died in 1714. The Earl of Oxford and his fellow Tories, including Swift, lost power. No organized attempt to found a language academy on French lines has ever been made since.

With Dryden and Swift the English language reached

its full maturity. Their failure to found an academy was partly counterbalanced by Samuel Johnson in his *Dictionary* (published in 1755) and by Robert Lowth in his *Grammar* (published in 1761).

Johnson's *Dictionary*

Age of Johnson. In the making of his *Dictionary*, Johnson took the best conversation of contemporary London and the normal usage of reputable writers after Sir Philip Sidney (1554–86) as his criteria. He exemplified the meanings of words by illustrative quotations. Johnson admitted that "he had flattered himself for a while" with "the prospect of fixing our language" but that thereby "he had indulged expectation which neither reason nor experience could justify." The two-folio work of 1755 was followed in 1756 by a shortened, one-volume version that was widely used far into the 20th century. Revised and enlarged editions of the unabbreviated version were made by Archdeacon Henry John Todd in 1818 and by Robert Gordon Latham in 1866. See also the articles JOHNSON, SAMUEL and DICTIONARY.

It was unfortunate that Joseph Priestley, Robert Lowth, James Buchanan, and other 18th-century grammarians (Priestley was perhaps better known as a scientist and theologian) took a narrower view than Johnson on linguistic growth and development. They spent too much time condemning such current "improprieties" as "I had rather not," "you better go," "between you and I," "it is me," who is this for?", "between four walls," "a third alternative," "the largest of the two," "more perfect," and "quite unique." Without explanatory comment they banned "you was" outright, although it was in widespread use among educated people (on that ground it was later defended by Noah Webster). "You was" had, in fact, taken the place of both "thou wast" and "thou wert" as a useful singular equivalent of the accepted plural "you were."

As the century wore on, grammarians became more numerous and aggressive. They set themselves up as arbiters of correct usage. They compiled manuals that were not only descriptive (stating what people do say) and prescriptive (stating what they should say) but also proscriptive (stating what they should not say). They regarded Latin as a language superior to English and claimed that Latin embodied universally valid canons of logic. This view was well maintained by Lindley Murray, a native of Pennsylvania who settled in England in the very year (1784) of Johnson's death. Murray's *English Grammar* appeared in 1795, became immensely popular, and went into numerous editions. It was followed by an *English Reader* (1799) and an *English Spelling Book* (1804), long-favourite textbooks in both Old and New England.

19th and 20th centuries. In 1857 Richard Chenevix Trench, dean of St. Paul's, lectured to the Philological Society on the theme, "On some Deficiencies in our English Dictionaries." His proposals for a new dictionary were implemented in 1859, when Samuel Taylor Coleridge's grandnephew, Herbert Coleridge, set to work as first editor. He was succeeded by a lawyer named Frederick James Furnivall, who in 1864 founded the Early English Text Society with a view to making all the earlier literature available to historical lexicographers in competent editions. Furnivall was subsequently succeeded as editor by James A.H. Murray, who published the first fascicle of *A New English Dictionary on Historical Principles* in 1884. Later Murray was joined successively by three editors: Henry Bradley, William Alexander Craigie, and Charles Talbut Onions. Aside from its *Supplements*, the dictionary itself fills 12 volumes, has over 15,000 pages, and contains 414,825 words, illustrated by 1,827,306 citations. It is a dictionary of the British Commonwealth and the United States, a fact symbolized by the presentation of first copies in the spring of 1928 to King George V and Pres. Calvin Coolidge. It exhibits the histories and meanings of all words known to have been in use since 1150. From 1150 to 1500 all five Middle English dialects, as has been seen, were of equal status. They are therefore all included. After 1500, however, dialectal expressions are not admitted, nor are scientific and technical terms not in general use. Otherwise, the written vocabulary is comprehensive. A revised edition of this dictionary, known as *The Oxford English Dictionary*, was published in 1933.

Proposals for a new dictionary

VARIETIES OF ENGLISH

British English. The abbreviation RP (Received Pronunciation) denotes the speech of educated people living in London and the southeast of England and of other people elsewhere who speak in this way. If the qualifier educated be assumed, RP is then a regional (geographical) dialect, as contrasted with London Cockney, which is a class (social) dialect. RP is not intrinsically superior to other varieties of English; it is itself only one particular regional dialect that has, through the accidents of history, achieved more extensive use than others. Although acquiring its unique status without the aid of any established authority, it may have been fostered by the public schools (Winchester, Eton, Harrow, Rugby, and so on) and the ancient universities (Oxford and Cambridge). Other varieties of English are well preserved in spite of the levelling influences of film, television, and radio. In the Northern dialect RP /ɑ:/ (the first vowel sound in "father") is still pronounced /æ/ (a sound like the *a* in "fat") in words such as laugh, fast, and path; this pronunciation has been carried across the Atlantic into American English.

Features of the Northern dialect

In the words run, rung, and tongue, the received-standard pronunciation of the vowel is /ʌ/, like the *u* in "but"; in the Northern dialect it is /u/, like the *oo* in "book." In the words bind, find, and grind, the received standard pronunciation of the vowel sound is /ai/, like that in "bide"; in Northern, it is /i/, like the sound in "feet." The vowel sound in the words go, home, and know in the Northern dialect is /ɔ:/, approximately the sound in "law" in some American English dialects. In parts of Northumberland, RP "it" is still pronounced "hit," as in Old English. In various Northern dialects the definite article "the" is heard as *t*, *th*, or *d*. In those dialects in which it becomes both *t* and *th*, *t* is used before consonants and *th* before vowels. Thus, one hears "t'book" but "th'apple." When, however, the definite article is reduced to *t* and the following word begins with *t* or *d*, as in "t'tail" or "t'dog," it is replaced by a slight pause as in the RP articulation of the first *t* in "hat trick." The RP /tʃ/, the sound of the *ch* in "church," becomes *k*, as in "thack," ("thatch, roof") and "kirk" ("church"). In many Northern dialects strong verbs retain the old past-tense singular forms band, brak, fand, spak for RP forms bound, broke, found, and spoke. Strong verbs also retain the past participle inflection *-en* as in "comen," "shutten," "sitten," and "getten" or "gotten" for RP "come," "shut," "sat," and "got."

In some Midland dialects the diphthongs in "throat" and "stone" have been kept apart, whereas in RP they have fallen together. In Cheshire, Derby, Stafford, and Warwick, RP "singing" is pronounced with a *g* sounded after the velar nasal sound (as in RP "finger"). In Norfolk one hears "skellington" and "solintary" for "skeleton" and "solitary," showing an intrusive *n* just as does "messenger" in RP from French *messager*, "passenger" from French *passager*, and "nightingale" from Old English *nihtegala*. Other East Anglian words show consonantal metathesis (switch position), as in "singify," and substitution of one liquid or nasal for another, as in "chimbly" for "chimney," and "synnable" for "syllable." "Hantle" for "handful" shows syncope (disappearance) of an unstressed vowel, partial assimilation of *d* to *t* before voiceless *f*, and subsequent loss of *f* in a triple consonant group.

In South Western dialects, initial *f* and *s* are often voiced, becoming *v* and *z*. Two words with initial *v* have found their way into RP: "vat" from "fat" and "vixen" from "fixen" (female fox). Another South Western feature is the development of a *d* between *l* or *n* and *r*, as in "parlder" for "parlour" and "carnder" for "corner." The bilabial semivowel *w* has developed before *o* in "wold" for "old," and in "wom" for "home," illustrating a similar development in RP by which Old English *ān* has become "one," and Old English *hāl* has come to be

South Western dialects

spelled "whole," as compared with Northern *hale*. In South Western dialects "yat" comes from the old singular *geat*, whereas RP "gate" comes from the plural *gatu*. Likewise, "clee" comes from the old nominative *clea*, whereas RP "claw" comes from the oblique cases. The verbs keel and kemb have developed regularly from Old English *cēlan* "to make cool" and *kemban* "to use a comb," whereas the corresponding RP verbs cool and comb come from the adjective and the noun, respectively.

In Wales, people often speak a clear and measured form of English with a musical intonation inherited from ancestral Celtic. They tend to aspirate both plosives (stops) and fricative consonants very forcibly; thus, "true" is pronounced with an audible puff of breath after the initial *t*.

Lowland Scottish was once a part of Northern English, but two dialects began to diverge in the 14th century. Today Lowland Scots trill their *r*'s, shorten vowels, and simplify diphthongs. A few Scottish words, such as bairn, brae, canny, dour, and pauky, have made their way into RP. Lowland Scottish is not to be confused with Scottish Gaelic, a Celtic language still spoken by about 50,000 people (all bilingual) in the Highlands and the Western Isles. Thanks to Robert Burns and Sir Walter Scott, many Scottish Gaelic words have been preserved in English literature.

Irish pronunciation

Northern Ireland has dialects related in part to Lowland Scottish and in part to the southern Irish dialect of English. Irish pronunciation is conservative and is clearer and more easily intelligible than many other dialects. Its literature has reached worldwide audiences, whether written by Englishmen born in Ireland, such as Jonathan Swift, Laurence Sterne, Oliver Goldsmith, Sir Richard Steele, Edmund Burke, Oscar Wilde, and George Bernard Shaw, or by authentic Irishmen, such as James Joyce, William Butler Yeats, and John Millington Synge. The influence of Irish Gaelic on the speech of Dublin is most evident in the syntax of drama and in the survival of such picturesque expressions as "We are after finishing," "It's sorry you will be," and "James do be cutting corn every day."

American and Canadian English. The dialect regions of the United States are most clearly marked along the Atlantic littoral, where the earlier settlements were made. Three dialects can be defined: Northern, Midland, and Southern. Each has its subdialects.

The Northern dialect is spoken in New England. Its six chief subdialects comprise northeastern New England (Maine, New Hampshire, and eastern Vermont), southeastern New England (eastern Massachusetts, eastern Connecticut, and Rhode Island), southwestern New England (western Massachusetts and western Connecticut), the inland north (western Vermont and upstate New York), the Hudson Valley, and metropolitan New York (Figure 3).

The Midland dialect is spoken in the coastal region from Point Pleasant, in New Jersey, to Dover, in Delaware. Its seven major subdialects comprise the Delaware Valley, the Susquehanna Valley, the Upper Ohio Valley, northern West Virginia, the Upper Potomac and Shenandoah, southern West Virginia and eastern Kentucky, western Carolina, and eastern Tennessee.

The Southern dialect area covers the coastal region from Delaware to South Carolina. Its five chief subdialects comprise the Delmarva Peninsula, the Virginia Piedmont, northeastern North Carolina (Albemarle Sound and Neuse Valley), Cape Fear and Pee Dee valleys, and the South Carolina Low Country, around Charleston.

These boundaries, based on those of the *Linguistic Atlas of the United States and Canada*, are highly tentative. To some extent these regions preserve the traditional speech of southeastern and southern England, where most of the early colonists were born. The first settlers who came to Virginia (1607) and Massachusetts (1620) soon learned to adapt old words to new uses, but they were content to borrow names from the local Indian languages for unknown trees, such as hickory and persimmon, and for unfamiliar animals, such as raccoons and

Prepared by Mrs. Raven I. McDavid, Jr., for *The Structure of American English*, by W. Nelson Francis; copyright © 1958, The Ronald Press Company, New York. Atlantic Seaboard dialect areas after Hans Kurath, *A Word Geography of the Eastern United States*; copyright © 1949, University of Michigan Press, Ann Arbor

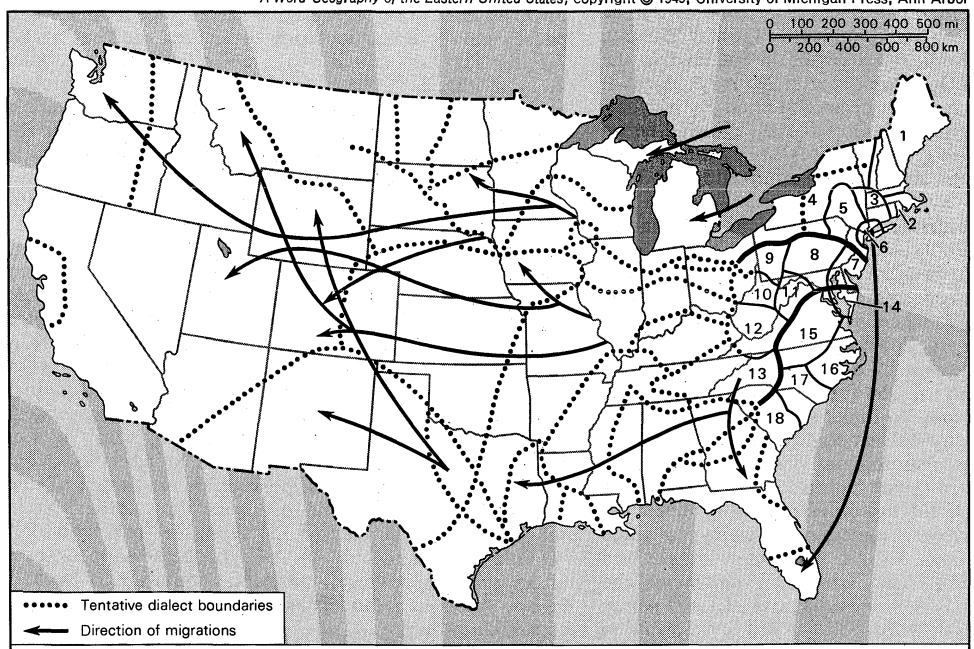


Figure 3: Dialect areas of English in the United States.

woodchucks. Later they took words from foreign settlers: "chowder" and "prairie" from the French, "scow" and "sleigh" from the Dutch. They made new compounds, such as "backwoods" and "bullfrog," and gave new meanings to such words as "lumber" (which in British English denotes disused furniture, or junk) and "corn" (which in British English signifies any grain, especially wheat).

Before the Declaration of Independence (1776), two-thirds of the immigrants had come from England, but after that date they arrived in large numbers from Ireland. The potato famine of 1845 drove 1,500,000 Irishmen to seek homes in the New World, and the European revolutions of 1848 drove as many Germans to settle in Pennsylvania and the Middle West. After the close of the American Civil War, millions of Scandinavians, Slavs, and Italians crossed the ocean and eventually settled mostly in the North Central and Upper Midwest states. In some areas of South Carolina and Georgia the American Negroes who had been imported to work the rice and cotton plantations developed a contact language called Gullah, or Geechee, that made use of many structural and lexical features of their native languages. This remarkable variety of English is comparable to such "contact languages" as Sranan (Taki-Taki) and Melanesian Pidgin. The speech of the Atlantic Seaboard shows far greater differences in pronunciation, grammar, and vocabulary than that of any area in the North Central States, the Upper Midwest, the Rocky Mountains, or the Pacific Coast. Today, urbanization, quick transport, and television have tended to level out some dialectal differences in the United States.

The Gullah dialect

The boundary with Canada nowhere corresponds to any boundary between dialects, and the influence of United States English is strong, being felt least in the Maritime Provinces and Newfoundland. Nevertheless, in spite of the effect of this proximity to the United States, British influences are still potent in some of the larger cities; Scottish influences are well sustained in Ontario. Canada remains bilingual. One-third of its people, living mostly in the province of Quebec, have French as their home language. Those provinces in which French is spoken as a mother tongue by 10 percent or more of the population are described as "federal bilingual districts" in the Official Languages Bill of 1968.

Australian and New Zealand English. Unlike Canada, Australia has no European language other than English within its borders. There are still many Aboriginal languages, though they are spoken by only a few hundred speakers each and their continued existence is threatened. Over 90 percent of the population is British. By the mid-20th century, with rapid decline of its Aboriginal tongues, English was without rivals in Australia.

During colonial times the new settlers had to find names for a fauna and flora (*e.g.*, banksia, iron bark, whee whee) different from anything previously known to them: trees that shed bark instead of leaves and cherries with external stones. The words brush, bush, creek, paddock, and scrub acquired wider senses, whereas the terms brook, dale, field, forest, and meadow were seldom used. A creek leading out of a river and entering it again downstream was called an anastomizing branch (a term from anatomy), or an anabranch, whereas a creek coming to a dead end was called by its native name, a billabong. The giant kingfisher with its raucous bray was long referred to as a laughing jackass, later as a bushman's clock, but now it is a kookaburra. Cattle so intractable that only roping could control them were said to be ropable, a term now used as a synonym for "angry" or "extremely annoyed."

A deadbeat was a penniless "sundowner" at the very end of his tether, and a no-hoper was an incompetent fellow, hopeless and helpless. An offsider (strictly, the offside driver of a bullock team) was any assistant or partner. A rouseabout was first an odd-job man on a sheep station and then any kind of handyman. He was, in fact, the "down-under" counterpart of the wharf labourer, or roustabout, on the Mississippi River. Both words originated in Cornwall, and many other terms, now exclusively Australian, came ultimately from British dialects. "Dinkum," for instance, meaning "true, authentic, genuine," echoed the "fair dinkum," or fair deal, of Lincolnshire dialect. "Fossicking" about for surface gold, and then rummaging about in general, perpetuated the term fossick ("to elicit information, ferret out the facts") from the Cornish dialect of English. To "barrack," or jeer noisily, recalled Irish "barrack" ("to brag, boast"), whereas "skerrick" in the phrase "not a skerrick left" was obviously identical with the "skerrick" meaning "small fragment, particle," still heard in English dialects from Westmorland to Hampshire.

"Rouseabout" and "roustabout"

Some Australian English terms came from Aboriginal speech: the words boomerang, corroboree (warlike dance and then any large and noisy gathering), dingo (reddish-brown half-domesticated dog), galah (cockatoo), gunyah (bush hut), kangaroo, karri (dark-red eucalyptus tree), nonda (rosaceous tree yielding edible fruit), pokutukawa (evergreen bearing brilliant blossom), wallaby (small marsupial), and wallaroo (large rock kangaroo). Australian English has slower rhythms and flatter intonations than RP. Although there is remarkably little regional variation throughout the entire continent, there is significant social variation. The neutral vowel /ə/ (as the *a* in "sofa") is frequently used, as in London Cockney: "arches" and "archers" are both pronounced [ɑ:tʃəz], and the pronunciations of RP "day" and "go" are, respectively, [dəi] and [gəu].

Although New Zealand lies over 1,000 miles away, much of the English spoken there is similar to that of Australia. The blanket term Austral English is sometimes used to cover the language of the whole of Australasia, or Southern Asia, but this term is far from popular with New Zealanders because it makes no reference to New Zealand and gives all the prominence, so they feel, to Australia. Between North and South Islands there are observable differences. For one thing, Maori, which is still a living language (related to Tahitian, Hawaiian, and the other Austronesian [Malayo-Polynesian] languages), has a greater number of speakers and more influence in North Island.

The English of India–Pakistan. In 1950 India became a federal republic within the Commonwealth of Nations, and Hindi was declared the first national language. English, it was stated, would "continue to be used for all official purposes until 1965." In 1967, however, by the terms of the English Language Amendment Bill, English was proclaimed "an alternative official or associate language with Hindi until such time as all non-Hindi states had agreed to its being dropped." English is therefore acknowledged to be indispensable. It is the only practicable means of day-to-day communication between the central government at New Delhi and states with non-Hindi speaking populations, especially with the Deccan, or "South," where millions speak Dravidian (non-Indo-European) languages—Telugu, Tamil, Kannada, and Malayalam. English is widely used in business, and, although its use as a medium in higher education is decreasing, it remains the principal language of scientific research.

English an alternative official language

In 1956 Pakistan became an autonomous republic comprising two states, East and West. Bengali and Urdu were made the national languages of East and West Pakistan, respectively, but English was adopted as a third official language and functioned as the medium of interstate communication. (In 1971 East Pakistan broke away from its western partner and became the independent state of Bangladesh.)

African English. Africa is the most multilingual area in the world, if people are measured against languages. Upon a large number of indigenous languages rests a slowly changing superstructure of world languages (Arabic, English, French, and Portuguese). The problems of language are everywhere linked with political, social, economic, and educational factors.

The Republic of South Africa, the oldest British settlement in the continent, resembles Canada in having two recognized European languages within its borders: English and Afrikaans, or Cape Dutch. Both British and

Dutch traders followed in the wake of 15th-century Portuguese explorers and have lived in widely varying war-and-peace relationships ever since. Although the Union of South Africa, comprising Cape Province, Transvaal, Natal, and Orange Free State, was for over a half century (1910–61) a member of the British Empire and Commonwealth, its four prime ministers (Botha, Smuts, Hertzog, and Malan) were all Dutchmen. In 1970 Afrikaners outnumbered Britishers by 3 to 2. The Afrikaans language began to diverge seriously from European Dutch in the late 18th century and has gradually come to be recognized as a separate language. Although the English spoken in South Africa differs in some respects from standard British English, its speakers do not regard the language as a separate one. They have naturally come to use many Afrikanerisms, such as *kloof*, *kopje*, *krans*, *veld*, and *vlei*, to denote features of the landscape and occasionally employ Bantu names to designate local animals and plants. The words trek and commando, notorious in South African history, have acquired almost worldwide currency.

Afrikanerisms in South African English

Elsewhere in Africa, English helps to answer the needs of wider communication. It functions as the official language of administration in the Bantu states of Botswana, Lesotho, and Swaziland and in Rhodesia, Zambia, Malawi, Burundi, Rwanda, Uganda, and Kenya. The University of East Africa, inaugurated in 1963, has three interstate constituent colleges in which the language of instruction is English, at Kampala (Makerere), Uganda; at Nairobi, Kenya; and at Dar es Salaam, Tanzania.

The West African states of Gambia, Sierra Leone, Ghana, and Nigeria, independent members of the Commonwealth, have English as their official language. They are all multilingual. The official language of Liberia is also English, although its tribal communities constitute four different linguistic groups. Its leading citizens regard themselves as Americo-Liberians, being descendants of those freed Negroes whose first contingents arrived in West Africa in 1822. South of the Sahara Desert, indigenous languages are extending their domains and are competing healthily and vigorously with French and English.

THE FUTURE OF ENGLISH

Geographically, English is the most widespread language on Earth, and it is second only to Mandarin Chinese in the number of people who speak it. The ITU (the International Telecommunication Union) has four official languages: English, French, Spanish, and Russian. The influence of these languages upon one another will inevitably increase.

It is reasonable to ask if changes in English can be predicted. There will doubtless be modifications in pronunciation, especially in that of long vowels and diphthongs. In weakly stressed syllables there is already a discernible tendency, operating effectively through radio and television, to restore the full qualities of vowels in these syllables. This tendency may bring British English more into line with American English and may bring them both a little nearer to Spanish and Italian. Further, it may help to narrow the gap between pronunciation and spelling. Other factors will also contribute toward the narrowing of this gap: advanced technological education, computer programming, machine translation, and expanding mass media. Spelling reformers will arise from time to time to liven up proceedings, but, all in all, traditional orthography may well hold its own against all comers, perhaps with some regularization. Printing houses, wielding concentrated power through their style directives, will surely find it in their best interests to agree on uniformity of spelling. Encyclopaedic dictionaries—computerized, universal, and subject to continuous revision—may not go on indefinitely recording such variant spellings as "connection" and "connexion," "judgment" and "judgement," "labor" and "labour," "medieval" and "mediaeval," "plow" and "plough," "realise" and "realize," "thru" and "through."

Factors leading toward conformity

Since Tudor days, aside from the verb endings *-est* and *-eth*, inflections have remained stable because they represent the essential minimum. The abandonment of the forms thou and thee may encourage the spread of yous and youse in many areas, but it is not necessarily certain that these forms will win general acceptance. The need for a distinctive plural can be supplied in other ways (*e.g.*, the forms "you all, you fellows, you people"). The distinctions between the words "I" and "me," "he" and "him," "she" and "her," "we" and "us," "they" and "them" seem to many authors to be too important to be set aside, in spite of a growing tendency to use objective forms as emphatic subjective pronouns and to say, for instance, "them and us" instead of "they and we" in contrasting social classes. Otherwise, these distinctive forms may remain stable; they are all monosyllabic, they are in daily use, and they can bear the main stress. They will therefore resist levelling processes.

Considerable changes will continue to be made in the forms and functions of auxiliary verbs, catenative (linking) verbs, phrasal verbs, and verb phrases. Indeed, the constituents of verbs and verb groups are being more subtly modified than those of any other word class. By means of auxiliaries and participles, a highly intricate system of aspects, tenses, and modalities is gradually evolving.

In syntax the movement toward a stricter word order seems to many to be certain to continue. The extension of multiple attributives in nominal groups has probably reached its maximum. It cannot extend further without incurring the risk of ambiguity.

In vocabulary there will be further increases, if the present trends continue. Unabbreviated general dictionaries already contain 500,000 entries; even larger dictionaries, with 750,000 entries, will be required soon. Coiners of words will probably not confine themselves to Greek and Latin in creating new terms, but they will exercise their inventive powers in developing an international technical vocabulary that is increasingly shared by Russian, French, and Spanish and that is slowly emerging as the universal scientific language of tomorrow.

BIBLIOGRAPHY

Dictionaries: The Oxford English Dictionary, 13 vol. (1933), is the reissue of *The New English Dictionary on Historical Principles* (1884–1928). Derivative dictionaries include *The Shorter Oxford English Dictionary on Historical Principles*, 2nd ed., 2 vol. (1939); *The Concise Oxford Dictionary of Current English*, 5th ed. (1964); *The Pocket Oxford Dictionary of Current English* (1969); *The Little Oxford Dictionary of Current English*, 4th ed. (1969); *Oxford Illustrated Dictionary* (1962); and *The Advanced Learner's Dictionary of Current English*, 2nd ed. (1963). Other one-volume dictionaries include *Chambers' Twentieth Century* (1972); *The Universal Dictionary of the English Language*, rev. by E.H. PARTRIDGE (1952); *Longmans English Larousse* (1968); and P. HANKS, *Encyclopedic World Dictionary* (1971).

The leading American dictionary is *Webster's Third New International Dictionary of the English Language* (1961), actually 8th in the series since the first appeared in 1828. *Webster's New Collegiate Dictionary* (1973) is an abbreviated version. Other comprehensive dictionaries are *The New Century Dictionary of the English Language*, 2 vol. (1959) and *Funk and Wagnalls New Standard Dictionary of the English Language* (1963). Two comprehensive dictionaries are outstanding: *The Random House Dictionary of the English Language* (1966); and *The American Heritage Dictionary of the English Language* (1969).

Up-to-date and reliable etymological dictionaries include ERNEST WEEKLEY, *An Etymological Dictionary of Modern English*, 2 vol. (1921, reprinted 1967); E.H. PARTRIDGE, *Origins*, 5th rev. ed. (1971); and ERNEST KLEIN, *A Comprehensive Etymological Dictionary of the English Language*, 2 vol. (1966–67). *The Oxford Dictionary of English Etymology* (1966) will long remain the most authenic work in this field.

The two great historical dictionaries of American English are: *A Dictionary of American English on Historical Principles*, ed. by SIR WILLIAM A. CRAIGIE and JAMES R. HULBERT, 4 vol. (1936–44); and *A Dictionary of Americanisms on Historical Principles*, ed. by MITFORD M. MATHEWS, 2 vol. (1951).

Modern usage: H.W. FOWLER'S somewhat eccentric *A Dictionary of Modern English Usage* (1926) was thoroughly updated by SIR ERNEST GOWERS (1965). It has its transatlantic counterpart in the following two works: BERGEN and CORNELIA EVANS, *A Dictionary of Contemporary American Usage*

(1957); and MARGARET NICHOLSON, *A Dictionary of American-English Usage* (1957).

Grammar and structure of English: HAROLD WHITEHALL, *Structural Essentials of English* (1956); A.A. HILL, *Introduction to Linguistic Structures: From Sound to Sentence in English* (1958); J.H. SLEDD, *A Short Introduction to English Grammar* (1959); R.B. LONG, *The Sentence and Its Parts* (1961); PAUL ROBERTS, *English Sentences* (1962); MARTIN JOOS, *The English Verb* (1964); H.A. GLEASON, *Linguistics and English Grammar* (1965); N.C. STAGEBERG, *An Introductory English Grammar* (1965); A.E. DARBYSHIRE, *A Description of English* (1967); R. QUIRK *et al.*, *A Grammar of Contemporary English* (1972); B.M.H. STRANG, *Modern English Structure*, 2nd ed. rev. (1968).

Phonetics of English: Handbooks include HANS KURATH and R.I. MCDAVID, *The Pronunciation of English in the Atlantic States* (1961); and A.C. GIMSON, *An Introduction to the Pronunciation of English* (1963).

Histories: The fullest account of the "external history" of the language is given by A.C. BAUGH in *A History of the English Language*, 2nd ed. (1963). FERDINAND MOSSE, *Esquisse d'une histoire de la langue anglaise* (1947), is a masterpiece—brief, lucid, and profound. KARL BRUNNER, *Die englische Sprache: Ihre geschichtliche Entwicklung*, 2nd ed., 2 vol. (1960–62), is indispensable to advanced students.

Two brief surveys written early in the century are recognized classics and remain stimulating: HENRY BRADLEY, *The Making of English* (1904, rev. by SIMEON POTTER, 1968); and J.O.H. JESPERSON, *Growth and Structure of the English Language* (1905, reprinted 1971). Other fairly substantial histories include STUART ROBERTSON and F.G. CASSIDY, *The Development of Modern English*, 2nd ed. (1955); G.L. BROOK, *A History of the English Language* (1958); M.M. BRYANT, *Modern English and Its Heritage*, 2nd ed. (1962); M.W. BLOOMFIELD and L.D. NEWMARK, *A Linguistic Introduction to the History of English* (1963); W.N. FRANCIS, *The English Language, an Introduction* (1965); THOMAS PYLES, *The Origins and Development of the English Language* (1964); SIMEON POTTER, *Our Language*, rev. ed. (1968); J.W. CLARK, *Early English: A Study of Old and Middle English* (1967); A.C. PARTRIDGE, *Tudor to Augustan English* (1969); J.A. SHEARD, *The Words We Use*, rev. ed. (1970); and B.M.H. STRANG, *A History of English* (1971). F.T. VISSER, *An Historical Syntax of the English Language*, 3 vol. (1963–73), provides copious illustrations and bibliographies.

Special studies: GEORGE W. TURNER, *The English Language in Australia and New Zealand* (1966); SIMEON POTTER, *Changing English* (1969); JOHN W. SPENCER (ed.), *The English Language in West Africa* (1971); MITFORD M. MATHEWS (ed.), *The Beginnings of American English* (1931); THOMAS PYLES, *Words and Ways of American English* (1952); ALBERT H. MARCKWARDT, *American English* (1958).

Bibliographies: ARTHUR G. KENNEDY, *A Bibliography of Writings on the English Language from the Beginning of Printing to the End of 1922* (1927); HAROLD B. ALLEN, *Linguistics and English Linguistics* (1966). New books are recorded in the *Annual Bibliography of English Language and Literature* (annual), edited for the Modern Humanities Research Association, and in *The Year's Work in English Studies* (annual), edited for the English Association. Books, theses, and current studies are announced in *Research in Progress in the Modern Languages and Literatures* (7/year), a supplement to Publications of the Modern Language Association of America (PMLA).

(S.P.)

Enlightenment

The Enlightenment was a movement of thought and belief concerned with the interrelated concepts of God, reason, nature, and man that claimed wide assent among European intellectuals in the 17th and 18th centuries. Although diverse in emphases and interests, the Enlightenment attacked the established ways of European life and, in its conviction that right reason could discover useful knowledge, aspired to the conquest of man's happiness through freedom.

ANTECEDENTS OF THE ENLIGHTENMENT

Ancestral roots. The roots of the Enlightenment reach back to the days when Greek philosophers discovered a regularity in nature and concluded that its governing principle was reasoning mind and when, under the promptings of Socrates, they turned to consider man and ascribed a high value to his intellectual powers. Individual philosophers and their schools raised Greece to a brilliant eminence, but its glory faded and Rome emerged as the heir. The Romans saved much, adapted much, and made contributions of their own; but the Greco-Roman culture finally lost its vitality. Unheralded, there came in a tide of Eastern religions preaching doctrines of personal salvation. Christianity won the competition for converts, and before the end of the 4th century it became the exclusive religion of the empire. Christianity sprang from an environment totally alien to that of the Greeks and Romans, and its purposes were so contrary to theirs that many fervent Christians wanted to repudiate pagan writings altogether. But the overwhelming majority of the converts were Greeks and Romans, for whom a radical break with their past was unthinkable; pagan literature was the only available source for philosophical justifications and for educational materials. And so Christianity had to make a cultural accommodation, but it tried to suppress the most immoral aspects of paganism and to bend the rest to the special purposes of the faith. The greatest achievement was that in the 13th century when Thomas Aquinas took up newly found manuscripts of Aristotle and made them into the very fundament of Christian philosophy.

Again, intellectual history moved from triumphs to reverses. More scholastic disputation badly injured Thomas' dominion; the papacy was moved to Avignon in 1309 (the discreditable "Babylonian Captivity"), and, worse still, expanding secular concerns in the 14th century shook the old order, nowhere more disruptingly than in Italy. Symptomatic of this change of atmosphere was Petrarch's newborn zest for classical letters and the response to it on the part of countless enthusiasts who roamed far and wide in search of forgotten manuscripts. Several revivals of classical letters during the Middle Ages had rescued many works from destruction, but the old religious purposes were not that of Petrarch and his friends, who wanted the ancient authors stripped of the distortions wrought upon them through the centuries of Christian editing and interpretation. These new men, the Humanists, sought to recapture the spirit of Greece and Rome, which had been expressed with such wisdom and beauty; and they produced a magnificent body of criticism and letters, historiography and science that made a glorious epoch in European history.

Petrarch's role in the revival of classical learning

But after about a century, intellectual life in Italy suffered ruinously from political turmoil, and Humanism went over the Alps to find congenial homes in Germany and elsewhere in western Europe. There the central purpose of Humanism was to serve religion, and this was especially true for Erasmus, who wanted to perfect a philosophy of Christ. Erasmus saw many things to reform, but when, in 1517, Martin Luther raised the standard of revolt, Erasmus drew back in horror.

The turbulent Luther unleashed more furies than he could have anticipated. A century later the old ecumenical church of the West was broken. The Protestants were appealing to the authority of the Scriptures, and the Catholics were invoking the authority of the organic church; and the Council of Trent (1545–63) closed the door on useful dialogue between them. Meanwhile, the course of the Reformation had thrown up problems of its own, and debates over would-be solutions had scattered cross-purposes and confusion. By the 17th century Europe was without a community of thought. The Renaissance had spread the diverse stream of Greek philosophizing, and Platonism, Stoicism, and Epicureanism competed for a hearing. Catholics and Protestants discharged their guns across an unbridgeable chasm; newer political questions raised up a veritable Pandora's box of conflicting opinions.

The decision of the Renaissance had been to keep the synthesis between classical letters and the Christian faith, but more recent developments had injured that hope. Inescapably, the need for an intelligible and accepted medium of communication sent men back to classical points of view, points of view not necessarily in conflict with Christian judgment but at the same time independent of it. One notable appeal to the minds of all Europeans was couched in terms of revived Stoicism. It was made by

Grotius' appeal to natural law

Hugo Grotius, a noble and scholarly Dutch jurist. In his *De Jure Belli ac Pacis* (1625) he invoked natural law as the rule governing international relations. For 2,000 years no one had denied the existence of natural law, but there was a fresh significance in the emphasis Grotius placed on its primacy and on its independence of Christian theology. This law, said the devoutly Christian Grotius, was so immutable that God himself could not change it.

The scientific revolution. Another development was working out a new frame of reference—the beginning of modern science, a discipline that soon lost patience with religious quibbling and, no less, with the absurdities of Greek and Roman science.

Bacon and Descartes. The great prophet of this newest learning, the man who stirred the imagination of the next generations by painting the enrichment of human life through new discoveries, was Francis Bacon. Suspicious of the mathematics that had been making great technical strides, he proposed an all-comprehending method of induction. In contrast, Descartes drove from his mind everything except simple ideas, whose truth was undeniable, and he proceeded by rigorous deduction to work out his whole philosophic system. Despite the gulf between these two methods, Bacon and Descartes were allied in basic purpose and in influence. Both were at war with what they considered a harmful past; both dreamed of improvement in man's lot as the duty of science; both separated science from theology. Both, too, saw the need for practical experiment, and neither had a useful notion of how to go about it.

Galileo. The honour of fundamental pioneering belongs to Galileo Galilei, a man of genius, piety, and truculence. Taking up the old figure of speech that God had given man two books, Holy Writ and nature, he insisted emphatically that the language of the book of nature was mathematics. For his purposes, Galileo limited his world of nature to properties capable of mathematical analysis, such as extension and motion, and therefore capable of being described in terms of precise natural laws. His method of research he elegantly demonstrated in working out the law of falling bodies by mathematical analysis of hypotheses and checking by experiment.

Such achievements should have been renown enough for any man, but Galileo had yet another triumph. As a starting point for it he had the astronomical work of Copernicus, Tycho Brahe, and Kepler. Their studies had done mortal damage to the Ptolemaic notion of a universe rotating around the Earth, but these great explorers had left behind them not only a storm of emotional resistance but also a number of crucial unanswered questions. Report had come to Galileo of a new invention called a telescope, and he made one for himself. In a few nights of peering into the heavens he saw a system of structure and motion that proved the Copernican heliocentric explanation to all but the most hidebound defenders of the old order. With the discredit of the Ptolemaic astronomy, the whole of Aristotelian natural philosophy became untenable, and Galileo in 1632 recklessly set down his views in *Dialogo sopra i due massimi sistemi del mondo, tolemaico e copernicano.* He dedicated his book to the Pope, but that act of courtesy availed him little. He was called before the Inquisition to hear a verdict already prepared; but the broader significance of the trial was that Galileo's new scientific way of thinking had gone so far from inherited modes that no reasoned interchange was possible, and the church based its condemnation on fear.

Newton. Another difficult problem remained: the need to construct a general system of mechanics that could account for the motions of the stars in terms of observed behaviour of matter on Earth. A number of scientists worked at it, but it fell to Isaac Newton to win the honour of success. In 1687 the Royal Society, under the authorization of its secretary Samuel Pepys, published Newton's *Philosophiae Naturalis Principia Mathematica.* In this work, with prodigious mathematical reasoning, Newton presented the law of universal gravitation.

Newton modestly confessed that he did not know what gravitation was, and when his calculations indicated some irregularities that seemed to require the direct intervention of God to set them right, he was honest enough to say so. The German scientist and philosopher Gottfried Wilhelm Leibniz, who had quarrelled with him over priority in the invention of the calculus, attacked Newton harshly for picturing God as a clumsy watchmaker, but few had patience with such small matters. Newton had explained the universe; he had vindicated the rationality of nature; and he was, said the zealous Voltaire, the greatest man who had ever lived. Behind this enthusiasm for Newton and for natural philosophy there lay a serious question. The mounting tide of discoveries in physics pressed toward the autonomy and the moral neutrality of science, but the spirit of the time could not allow this kind of isolation for so wonderful an instrument of knowledge. The by-product of science had been the destruction of wide domains of medieval error; perhaps its methods, its experimental reasoning, could be used for all of man's problems. Indeed, Newton himself had suggested the possibility.

The effect of Newton's discovery of the law of gravitation

THE AGE OF THE ENLIGHTENMENT

Enlightened religion. The Christian faith showed a powerful tenacity when confronted by this avalanche of new facts and new conclusions, but it could not fail to be acutely sensitive to the changing intellectual fashions. By the early 18th century, radical ideas hitherto submerged came into the open and precipitated a debate fraught with great consequences. Hitherto, the rationalists had had no base for comparison except their knowledge of the ancient world. Latterly, they had become acquainted (in considerable measure through Jesuit missionaries) with pious Egyptians and Siamese and with Chinese, who, they said, had no real religion at all but were still men of exemplary virtue. More recently still, another stock character came upon the European stage: the noble savage, truly a child of innocence who worshipped his god in simple faith and obeyed the precepts of Mother Nature. And, finally, this growing breed of scientists was steadily grinding out disturbing implications. Good Christians though most of them were, they showed that Christianity's views of nature could not stand up under the tests of experimental reason.

The rise of Deism. Such considerations led dissatisfied men to assume that, beneath the world's religious diversities, there was a common body of beliefs, a religion planted by nature in all men everywhere. In 1624, while Grotius was preparing his case for natural law in international relations, Lord Herbert of Cherbury published his *De Veritate*, a definition of natural religion. He said that certain beliefs are so manifest that all men of reason accept them: the existence of one God who dispenses rewards and punishments, and the obligation resting on man to worship him in repentant piety and virtue. These articles sufficed for a religious life in the present world and for salvation in the hereafter. Deism, as this doctrine came to be called, did not always radiate Herbert's lofty Stoic tone, but his statement was so representative that it persisted until Thomas Paine, in his *Age of Reason* (1794–96), wrote the swan song of the movement.

The Deists' views

The Deists maintained the old synthesis between God, reason, and nature, but, reasoning from a mathematically ordered nature to its architect, they allowed the architect only the qualities manifest in his handiwork. Grand metaphysical systems fell under the ban as mere extravagant nonsense; philosophizing, said the dominant voice, should follow the method of the natural sciences. Such convictions led inescapably to relentless warfare on those beliefs that were unique to Christianity. The Holy Writ of the Christians suffered denunciation as groundless historical confusion; but the Old Testament came in for special attention—it was scornfully dismissed as a repository of the crimes and obscenities in which God connived with the so-called chosen people. With equal vigour the Deists ridiculed biblical prophecy out of court; the notion that God would make revelations in such a book, or in later manifestations, was utterly offensive to the rational mind. Miracles, therefore, were unthinkable. It would do

violence to reason to suppose that God would, or could, set aside the laws that gave expression to his own eternal being.

Deism was never an organized movement. Its first home was among English intellectuals. Around 1700 a series of books appeared that hit ruthlessly at what they held to be the misconceptions and offenses of orthodox Christianity and argued for a depersonalized God, whom anyone could discern for himself. Some of the authors offended good taste and the law by their blasphemies; some wrote with calculated prudence. But during a half century the central core of their argument remained essentially unchanged: Christianity was an indefensible burden on the back of natural religion. But then English Deism went quietly into decline, in part because public disputation was reduced to tedious repetition and in part because the Anglican Church was too preoccupied with political advantages and social dominance to wish to disturb its fat slumber by theological contentiousness. England in 1688–89, with its Glorious Revolution, had settled its wracking political issue, and it was time to relax. The clergy preached cooly on the prudential morality of the Christian life, while Joseph Addison taught his readers the affable urbanity appropriate to a civilized society in an age of reason. Deism moved its headquarters to the Continent.

France had its own native roots of Deism, which grew vigorously on the disillusionments and discontents of the reign of Louis XIV. Voltaire was already committed to an anti-Christian Rationalism before he sought a haven in England in 1726. On his return after two years of exile he wrote a sparkling report, *Lettres philosophiques*, which reaffirmed his commitment to the Deist view of religion. Through his remaining years, his faith remained inflexible. Pressed by more radical opinions, he held steadfastly, in the end somewhat frenziedly, to the God whose evidence he saw in nature and whose existence he deemed necessary to preserve the social order. His one change of emphasis was in the growing vehemence of his anti-clericalism. In the 1760s he completed his *Dictionnaire philosophique*, a witty but earnest compendium of malice toward Christianity, and at the same time, poltroon though he frequently was, he entered the public lists to bring justice to several persons he considered to be victims of ecclesiastical murder. By then he had adopted in rage the motto *Écrasez l'infâme!:* "Every sensible man, every honourable man, must hold the Christain sect in horror."

No one was more temperamentally removed from Voltaire than Jean-Jacques Rousseau, but when he included a profession of Deist faith in his novel *Émile*, Voltaire sent him an enthusiastic compliment. Rousseau's profession was a manifesto for reason as the only reliable test for religious truth and equally an attack on all doctrines and clerical institutions that stood between autonomous man and God. Rousseau characteristically made a place for heightened emotion in such passages as the invitation to go to the sublime book of nature and learn there to worship the Creator; but the overtone of anti-intellectualism did not obscure the rational doctrine of Deism.

The spread of Deism

Throughout France opposition to the church spread, drawing strength from the debilities of the *ancien régime*, and the Deist faith went on to the culturally less well prepared soil of Germany and to the British colonies in America. The movement kept up its momentum far into the second half of the 18th century, but it then paid the price of its lack of organization and the weakness of its emotional impact on uncritical minds. Without a sacred text, without a compelling leader, Deism was always a matter of individual judgment and thereby vulnerable to centrifugal forces.

The Christian response

Naturally, when established societies felt the shock of Deist subversion, they resorted to the time-honoured methods of defense. First came Christianity's dogmatic assertion of truth from pulpit and press (*e.g.*, in 1770 French publishers brought out 90 books in support of the Catholic faith). Behind such gentle measures, however, there were, in considerable diversity, grimmer ways of suppressing dissent. Grimmest was the vengeance of the Inquisition, but hardly less so was the possibility of civil judgments, including the death penalty. And then came the *Index Librorum Prohibitorum* (*Index of Forbidden Books*), ecclesiastical and state censorship and licenses to print, and book burnings. Much skilled and harsh determination went into devising means to stifle corrupting thought. Many individuals suffered many punishments, but heresy was a hydra-headed monster, thriving on persecution and on the stupidities of its enemies.

In Vienna in 1772 the political authorities suddenly realized that they had to put their own *Catalogue of Forbidden Books* in that self-same *Catalogue*, because the impure of heart were finding it a useful guide to spirited reading. Administrative inefficiency and, notably in France, competing authorities helped to dent the cutting edge of tyranny, but the greatest aid for those dealing in illicit materials was their own ingenuity in adapting the arts of the smuggler and the forger into their affairs. Books sold under the table, books without authors, books with false places of publications—all were commodities that did well. In these deceptions Voltaire was by no means alone, but he was a superior master. His brilliant style was more clearly his own than was an acknowledged signature, but in edition after edition he denied writing the *Dictionnaire philosophique*, and no policeman challenged his audacity.

Such cat-and-mouse antics were not the whole story of this confrontation. The real trouble was more subtle. Christianity was rooted in both reason and revelation, and, according to the Fathers and the doctors of the church, these sources were not in conflict; revelation simply had the higher truths. Now came Deism, professing to have the full credentials of reason and demanding that the Christians show how they could, in reason, defend revelation. Again the first battleground was England. Increasingly, supporters of the Christian faith found themselves driven to admit the validity of natural religion and then to search out reasonable grounds for adding revelation. John Tillotson, future archbishop of Canterbury, spoke for many when he stated that Christianity was needed to complete the findings of natural religion: its service was to test the reasoning process and to give a supplementary motive for a moral life. Tillotson conceded much, but his contemporary and fellow Christian John Locke went even further when he accepted reason as the judge of things above reason. Bishop Butler's *Analogy of Religion Natural and Revealed* (1736) offered a fresh line of Christian defense. It became a famous book, but, despite the kindly Bishop's Christian dedication, it did not demolish many Deists. In the pinch, he did not venture into dogmatic assertions based on the usual Christian sources; he took his stand on probability, and he reminded his readers that there were things in natural religion and in nature just as difficult to believe as was orthodox Christian doctrine: after all, Newton could not explain gravitation. Comparable gestures toward a friendly view of the rational standard appeared in Holland, in Germany with a group of prominent Lutheran pastors, and in France, where the Jesuits had a high skill in reasoned debate until their expulsion in 1762. The consequence was that, on the Continent as in England, the theological content of practical, daily religion declined significantly. The popular preachers discoursed in muted tones of reasonableness on the beauties of Christian virtue.

While Deists and Christians were exchanging blows, they tended to agree that God in his benevolence had designed the Earth for man's happiness; "the best of all possible worlds," Leibniz said. Another said that the sun rises and shines for men, and still others pointed to the evidence of the Creator's forethought—the density of water to facilitate navigation; the shape of the melon to make it easy to cut. Even Voltaire, having no conception of change in the universe, surveyed the mountains and saw in them the wisdom and the benevolence of God.

Atheism. Both Deists and Christians heard with anxiety that their happy assumptions were threatened by the upsurge of atheism. Materialism was an ancient philoso-

The rise of Materialism

phy. Epicurus had propounded it, and his Roman disciple Lucretius had put it into his poem *De rerum natura*, which, still later, had excited the Renaissance enthusiasts for classical letters. Throughout the whole of its history, Epicureanism had been a scandal for the mass of literate people, but a notable moment came in 1563, when a French Humanist published a new edition of Lucretius and urged his readers to forget the absurd ideas and concentrate on the poetry. The plea helped a bit, but a far more significant work of salvage was done in the 17th century by Pierre Gassendi, a Catholic priest of blameless repute. Boldly, Gassendi said that Epicurus had not been an advocate of vice but had cultivated the virtues of honesty and inner peace. His Materialism, Gassendi went on, was erroneous; but he shrewdly suggested that it could be made palatable by postulating Divine Providence behind Epicurus' atoms and voids. Descartes would not admit a void in his universe, but he made a contribution, however unwillingly, by his dualism of spirit and matter, which made all animals, including physical man, into purely mechanical structures. With the growth of science, the 18th century took kindly to Gassendi's revival of atomistic philosophy and to the Cartesian mechanistic view of nature. In England David Hartley rewrote Lockean psychology and explained mental life in terms of physical vibrations building complex patterns by the equally mechanical processes of association. In France another physician, La Mettrie, brazenly spelled out the title of his book, *L'Homme-machine* (1747), and stated that man, like everything else, was reducible to matter in motion. "We are no more committing a crime," he wrote, "when we obey our primitive instincts than the Nile is committing a crime with its floods, or the sea with its ravages." Voltaire would not entertain what was to him an offensive notion, but the Encyclopaedist Denis Diderot was tempted and succumbed. The most powerful atheist influence was that of Baron d'Holbach, a wealthy German who had settled in Paris to wage war against God. He commissioned the publication of a whole library of books in addition to those he wrote himself; his only criterion was that a book should serve the cause. In his eyes, religion was the main source of degradation. Rumours circulated about his dinner parties, and ordinary folk felt uneasy when they passed his house, but d'Holbach went further in his anger against Christianity than many men were willing to go.

Skepticism. One further, and more powerful, threat to both Deism and Christianity was skepticism, also a philosophic doctrine that had ancient roots. Because of the lasting popularity of Montaigne's essays of the 16th century, one must see in the meditations of this gentle devotee of ancient letters a fresh source for modern doubt. As he mused on human folly and fanaticism, he repeatedly threw in his personal conclusion, "What do I know?" It was a diffident question, but over several generations it prodded many minds.

The influence of Pierre Bayle

A century later, subdued reflection gave way to a noisy uproar. The cause was Pierre Bayle, whose course from Protestantism to Catholicism and back to Protestantism sent him into a life of exile and of hostility to all religions. His instrument of vengeance was his *Dictionnaire historique et critique*, the first edition of which appeared in 1697. In form it was a collection of biographical sketches with comments, but the individuals chosen and the amount of space given to each reflected Bayle's deadly purpose—to demolish the vices of religion. With learning and wit and with a titillating spice of salaciousness, Bayle drove his sword at every exposed surface of the erstwhile eternal verities. Perhaps nowhere did he hit harder than at the Old Testament when he asked how it was that, when Holy Writ chronicles the morally offensive record of King David, it goes on to call him a man after God's heart.

Bayle's *Dictionnaire* had an extraordinary success. It was soon expanded in new editions, translated into English and German, and boiled down into abridgments and digests. The rollcall of those who read it was a veritable *Who's Who* of the 18th century. Frederick II of Prussia commissioned two abridgments; Benjamin Franklin and Thomas Jefferson recommended it to their friends; Voltaire drew inspiration from it for his own *Dictionnaire philosophique;* Diderot wrote a lyrical article on Bayle for the *Encyclopédie* and was enraged when the rascally publisher suppressed the main substance. No one, however, was more persuaded than the Scottish former Presbyterian David Hume.

At 28 Hume published the first two volumes of *A Treatise of Human Nature*. The subtitle showed the cast of a mind that never changed: *an Attempt to introduce the experimental Method of Reasoning into moral subjects*. From this first, and unsuccessful, book to his posthumous *Dialogues concerning Natural Religion* (1779), Hume, with amiable reasonableness, dissected all certainty from human affairs. Down went the concept of cause and effect, an idea seemingly essential to all rational thinking; down went the celebrated powers of human reason. But the realm of belief suffered most notoriously in this raging storm of doubt. The miracles of so-called revealed religion he subjected to the test of experience, and they dissolved beyond rescue except for the most willful believers. Even natural religion could not stand up to the test of fact; there was no universal consensus supporting it. According to Hume, primitive men, in their fear, invented gods, many gods, to gain protection, and it was only a long time later that monotheism emerged from ignoble calculations. The great trouble, said Hume, was that men assumed that the Supreme Being gave his qualities to them as an endowment of the creation, whereas in reality men ascribed to God their own values. Here, then, was a repugnant philosophy; nothing was left but mere uncertainty and probability at best. The only saving grace, Hume puckishly admitted, was that one could not be a skeptic all the time.

Christian defenses. In the eyes of orthodox believers in Christ, the new trends were all manifestations of a perverse spirit that incited defiance as well as consternation. From Christians little moved to dispute came an assertion that their religion did not need to submit to these debasing tests of rationality. The mystics and others of unassailable conviction went their own ways, indifferent to foolish clamours. Still, it was not always easy. Blaise Pascal, mathematical genius and convert to the rigorous faith of the Jansenists of France, had felt agonizingly the tension between the old ways and the new. He had been convinced that human reason could not explain the ultimate philosophic questions, and he sought for moral certainty in Christian revelation. The universe was vast and men were insignificant, but they had souls and they were dependent on God. In his fear of the vastness of space, Pascal committed himself to a supreme act of faith: "The heart has its reasons which reason does not know." As a student of the law of probability derived from the gambling table, Pascal took refuge in his wager on the existence of God: "if I lost I would have lost little; if I had won I would have gained eternal life."

Chief of those who a century after Pascal would not accept the threatened demolition of the Christian faith was John Wesley, an Oxford student of the classics, early given to piety and a second-time convert through the Moravian Brethren (a German pietist sect). It was no novelty when Wesley said that reason could not produce the love of God, but it was strangely exhilarating when he told his spiritually hungry followers, "A true and living faith in Christ is inseparable from a sense of pardon from all past and freedom from all present sins." For 50 years Wesley preached his simple message of the love of God manifested in Christ, and the humble folk found consolation as Wesley travelled up and down England preaching an average of 1,000 sermons a year.

And so before the end of the 18th century something less than success had crowned the efforts of those earnest men who used emancipated reason and the methods of science to explain God and the ancient mysteries of the universe. The old faith had been buffeted, but Wesley was only the first announcement of a religious trend toward the unleashed emotional flights of the Romantics.

The investigation of man. Meanwhile, the attempt to examine the nature of man in the light of new thought

raised difficult problems. The old and continuing Christian view had a harsh consistency: man carried in him the stamp of Adam's sin until the hour of redemption through Christ's atonement arrived. The Renaissance concept of the dignity of man was an implied criticism of the old belief. But now, in the heady days of rationalistic deductions, it seemed axiomatic that God had implanted his reason and his goodness in man. In this view there was no place for original sin; were it otherwise, the creation would have been monstrously unjust, and that was unthinkable. But the difficulty about this happy analysis was to find men who were convincing demonstrations of the thesis.

The notion of the goodness of man

Locke's epistemology. Because a priori reasoning about man was not very profitable, it was in order to utilize Newton's new scientific method to study human nature. The pioneer was John Locke, and his vehicle was *An Essay Concerning Human Understanding* (1690). Locke's purpose was to determine the scope and limitation of human knowledge through an investigation of the workings of the mind. In the end he set a low limit to knowledge, but what made him the great teacher of the 18th century was his description of the mental process: all knowledge arises from experience, from sensations sorted and combined by reflection. There are no innate ideas; every child is born with a mind like a blank sheet of paper on which the impingement of the external world inscribes the data of life's accumulation of learning. Bishop Berkeley attacked Locke for assuming a world of substance as the source of sensations when that world was beyond the reach of knowledge; but aside from Berkeley's criticism, the psychology of the next century was essentially a gloss on Locke.

The Lockean investigation had demolished not only innate ideas but also innate depravity. Environment made the man, and the most obvious corollary was that the way to improve man was to improve the environment. It was a lesson that gave new dimension to human possibilities and therefore a strong impetus to the zeal for humanitarian reform.

Ethical questions. Locke was primarily concerned with epistemology and gave little attention to the nonmental aspects of man. The language of the day spoke of the passions—the emotions—and their role became increasingly important as the study of human behaviour continued. The rationalist deduction held to man's inherent goodness by virtue of his reason: to know what was right was to do what was right. This statement hardly fitted the observable facts, and a qualification was added: when reason was not obscured, it would tell man what was right and he would act accordingly. But, again, this revision was short of full accuracy: it was established fact that a man who clearly knows the truth may not serve it. It was therefore in order to investigate more carefully the inner sources of moral action. Even exalted rationalists had to admit the existence of ethical promptings in the nonintellectual side of human nature (*i.e.*, in the passions), and two generations and more sought a precise explanation. Early in the 18th century, the 3rd Earl of Shaftesbury found it in an instinctive love for the good because of its inherent worth, its beauty, and its reflection of nature. Bishop Butler pointed to the conscience, the voice of God speaking to the human soul. Even David Hume turned from his skepticism long enough to ascribe to man an innate sentiment for humanity, and his friend Adam Smith, the calculating economist, gained his first celebrity in 1759 from a book that asserted that sympathy was a guiding emotion in human conduct. Across the channel, Voltaire, that alleged prince of cynics, found in man a natural benevolence never seen in the beasts, and the radical Morelly wrote in his *Code de la nature* (1775), "In the natural order the idea of active or passive benevolence precedes every other idea, even that of Divinity."

Running through these comforting reassurances was perforce a question of the relation between this ethical faculty, rooted in the emotional impulses, and reason. Shaftesbury was persuaded that rationality and moral emotion could never be at odds because both were the gifts of nature. Butler contended that conscience had its own power of reasoning. In the second half of the century the common voice began to falter. Adam Smith could not say that sympathy and reason would invariably counsel the same moral decision, and Hume, returned to his most destructive vein, wrote in his *Treatise*, "We speak not strictly and philosophically when we talk of the combat of passion and of reason. Reason is, and ought only to be, the slave of the passions, and can never pretend to any other office than to serve and obey them." Many of Rousseau's passages implied, however unphilosophically from Hume's point of view, an open conflict between reason and the passions. Under the inspiration of Rousseau's idealized man of virtue, novelists were soon playing variations on the theme that goodness was merely a matter of opening the heart's emotional spigots.

Conflict between reason and the passions

Meanwhile, the discussion of ethics had addressed itself to an even more important theme. Christianity had always made its basic appeal to man's desire for the happiness of heaven and his fear of the torments of hell. It was easy to generalize this human predilection and say simply that the dynamic force in life was the desire to secure pleasure and to avoid pain. Christian John Locke agreed: "Pleasure and pain and that which causes them, good and evil, are the hinges on which our passions turn." With a boldness that threatened to become unseemly, numerous writers, notably the exuberant Diderot, sang praises of the passions and in so doing obliquely discounted reason.

It was necessary to think of the ethical implications. Few were ready to make the pursuit of pleasure a moral code in itself. The extravagant determinists could argue that man's decisions and acts were simply the consequences of matter in motion and that there could thus be no question of morality in human life; but most reflections on ethics, despite Hume, assigned to reason some kind of brake on personal misbehaviour. Others contended that the experience of living with people taught men the need they had of each other and the ensuing necessity for reciprocal obligations. In addition, as a justification to some and a reassurance to others, there circulated an assumption of a mysterious harmony between self-interest and the common good. Montesquieu wrote, "Everyone pursues the common good under the impression that he is following his own private advantage." The impudent Bernard de Mandeville expressed the same conviction in his *Fable of the Bees* and summarized the whole matter in his pithy aphorism: "Private vices, Publick benefits." Alexander Pope was more eloquent: "Thus God and Nature link'd the gen'ral frame/ And bade Self-love and Social be the same."

The brave Newtonians had confidently set out to create a science of morals, but they had only created confusion. Another question arose: if there was this anxiety about the capacity of the individual to find within himself a reassuring incentive to ethical conduct, was it possible to find a source external to the individual? For believing Christians the answer was simple: God through the Holy Scriptures and his providential governance of the world has given his commands, and it was man's duty to obey them. It was a far more difficult task for the would-be scientific thinkers, who had broken relations with the Christian God. Thomas Hobbes in the mid-17th century had left a suggestion. In a bit of hypothetical history he told of people who in the state of nature were, each and every one, under the sovereignty of pleasure and pain. The consequence was war of all against all, and life was "solitary, poor, nasty, brutish, and short." To escape from this deadly plight, these people set up a state and consigned to it an all-comprehensive power to rule. The compact created obligation. Hobbes's successors faced the same problem. To restrain moral anarchy it was imperative to design a code based on the welfare of the community. The expression "the greatest happiness for the greatest number" had first appeared in the 1730s, and Claude-Adrien Helvétius made an ominous suggestion. In his book *De l'esprit* (1758), burned by the parlement of Paris and anathematized in Rome, Helvétius explained how to take this self-centred creature man and make out

of him a social being: "It is solely through good laws that one can form virtuous men. Thus the whole art of the legislator consists of forcing men, by the sentiment of self-love, to be always just to one another." Out of such materials Jeremy Bentham drew his blueprint for a laissez-faire democracy.

ASPECTS OF ENLIGHTENMENT THOUGHT

Humanitarianism. It was perhaps less important that these 18th-century thinkers could not fashion an agreed science of morals than that so many people were deeply concerned. The Deist search for natural religion had led to wide-ranging investigations of peoples all over the Earth, with investigations centred not on manifest differences in colour and customs but on what people had in common. And they found much. "It is universally acknowledged," Hume wrote in 1748, "that there is a great uniformity among the acts of men, in all nations and ages, and that human nature remains still the same in its principles and operations. History informs us of nothing new or strange in this particular. Its chief use is only to discover the constant and universal principles of human nature." The sum totality of all men made up humanity, and humanity, by some subtlety, inspired loyalty and compassion. Under this influence, vaunted reason was turning itself unobtrusively into reasonableness. When the concept of natural law was applied to social problems it underwent a significant modification. It no longer stated what was, but what ought to be; and since in so many respects the human condition fell far short of the ought, the Enlightenment became critical, reformist, and eventually revolutionary.

Universal principles of human nature

Locke's demonstration that environment made men provided the dynamic inspiration for a vast attack on prevailing abuses. In the presence of disinterested compassion, those laboured analyses of man chained forever to his pleasures and pains seemed sterile and irrelevant. The long, ceaseless campaigns waged by the intellectuals for freedom of religion and speech might appear to a cynic to be breast-beating exercises to secure their own privileges and rights, but certainly the cynic's yardstick could not measure a countless list of self-jeopardizing acts by men working to increase individual freedom for the simple reason that individual freedom was good. Out of that same simple reasoning came an attack on the inhumanity of Europe's brutal administration of brutal law.

One blind spot, as later generations would call it, was the Enlightenment's little thought for the misfortunes of the lesser rungs of the European social ladder. The poor Europe had had with it always. Sympathy leaped more readily over the horizon to embrace people of dark hue whose status touched the European conscience. Atheist and Christian equally agreed that slavery was morally wrong, and they made common cause against it.

The religious radicals and their Christian enemies also shared an insistent sense of all men linked together in universal brotherhood. Standing in the way of the ideal were many grim realities, but the most baffling was the sovereign state, which demanded the obedience and, far too often, the blood of its people. Despite early warnings of a ferment of nationalism, men of many divergent persuasions toward other issues sounded their defiant calls to world citizenship. Invoking the spirit of Grotius, the Swiss lawyer Vattel pleaded for a society of states living together in peace under the morally binding prescriptions of natural law. Other men who loved peace recalled a proposal attributed to Henry IV of France to establish a formal international confederation that would make war in Europe impossible; the Abbé de Saint-Pierre revived this project in 1713 and worked at his cause for 30 years without gaining many converts. Rousseau published a rather lukewarm restatement, and there the peace campaign rested until 1795, when Immanuel Kant spoke the final word for the Enlightenment in "Zum ewigen Frieden." But Kant's eloquence was to no purpose, for at the time Europe was already two years into a war that would not end until 1815.

Social thinking. The application of the Enlightenment's rationalistic and scientific tools to social fields led to no settled agreement, either as to method or to conclusions.

Political thought. The Reformation had thrown up doctrines ranging from primitive communism to the divine right of kings. Bishop Bossuet restated with the best French eloquence of the 17th century the special relation between God and Louis XIV, but the ancient principle of the natural law went on talking about the state of nature, the social contract, and natural rights. The weight of these theories were, by their very nature, on the side of radicalism. Yet a curious thing happened—the first notable statement of political philosophy in the new age of mathematical rationalism was a secularized theory of absolutism. The author was Thomas Hobbes, who became a rigorous reasoner when, in middle age, he chanced upon Euclid's *Elements*. Hobbes started with natural law but, as indicated above, he admitted no "ought to be." The law imposed a sleepless search for pleasure and the passionate desire to escape pain. The people caught in the desperate consequences of this rule of life made a social compact whereby all resigned their individual powers into the hands of a sovereign outside the compact, whose authority for all practical purposes was unlimited. Hobbes had many acute insights into human affairs, but Charles II, whom he was trying to support in the distress of his exile in France, was so scandalized that Hobbes had to flee the royal presence and take refuge with Oliver Cromwell.

The social compact

Half a century later John Locke accepted Hobbes's strong sense of individualism and his conception of the state as a mere mechanical convenience rather than a divine institution, but in all other respects his arguments went back to pre-Hobbesian terms. The state of nature, said Locke in the second of his *Two Treatises of Government*, was a social regime of men of good will; in it, natural law had endowed men with rights, principal among which were life, liberty, and property. Certain confusions arose when every man defined for himself and tried to protect his own rights, and it was minor difficulties and not desperation that inspired the formation of a political community specifically and exclusively for the protection of individual rights. Government was derived from the consent of the people, and it functioned by majority rule. Locke's purpose had been to compose a challenge to Charles II's brother, who was heir apparent; but the *Treatises* were not published until 1690, and by that time the Glorious Revolution had closed a troubled period and Locke's revolutionary tract was appropriated as a defense of the Revolution. Thereafter, England settled down for a long time to enterprises requiring no new political philosophizing.

While England was relaxing into complacency, most of the Continent still remained too bound to a lethargic inheritance to think of formal political theory. A few Germans, a few men of the Low Countries, an occasional Italian rose sufficiently above their bleak routines to more systematic thought, but it is best, for summary purposes, to concentrate on the affairs of France. Long before the death of Louis XIV, times were out of joint and remained so, with few respites, to the end of the monarchy. There were paralyzing difficulties, however, in finding the causes of the disturbed temper of the 18th century and proposing suitable remedies. The French monarchy had done too well in that it had stifled all prospects of generally popular institutions and had made public discussion a hazardous occupation. The monarchy had done badly in that it had not rooted out the mentality and the vested interests and important agencies of feudalism. The monarchy had also done badly in that its close ties with the Catholic Church made every grievance against the church either obliquely or overtly a grievance against the royal structure of power. The consequences were disastrous for judicious thinking about politics.

All his life Voltaire unremittingly poured his energy into supporting the powers of the crown against the parlements, the medieval law courts that had become citadels of aristocratic resistance. Montesquieu wanted a greater role in government for the nobility of which he was a member; for that reason he held the British House of

Lords in high esteem. Diderot, not by instinct a political animal, never understood the contest between king and nobles and always blew hot or cold as immediate circumstances prompted him. And yet for him, as for Voltaire, realism bade him acknowledge that all kings of France, even Louis XIV, had faults; but he never revealed doubts about the monarchy. He had courted Louis XV for preferments and had got them, and he took pride in the fact that Frederick II of Prussia had courted him. A new trend stretching across Europe from Spain to Russia gave plausibility to the thought that man's best hope for reforms lay in the vigorous sovereign. In the most unexpected capitals, rulers were reorganizing their governments, removing tax and other burdens on trade, and restraining the brutalities of the law. No enlightened Philosophe in France could complain until he realized that these so-called enlightened despots had another, and a repugnant, face turned toward war and conquest.

What was left for France to talk about, and to write about on condition of prudence, were the abuses of political power along with those of ecclesiastical power. Despite this dejecting atmosphere, in 1762 a new kind of political book appeared in Paris. It was called *Du contrat social* and its author was the difficult, emotionally unstable Jean-Jacques Rousseau, who had first won fame by a perfervid attack on the evil consequences of civilization. Rousseau proposed a society able to cultivate the individual's moral stature without injury to his freedom. It was a fresh effort at that knottiest problem of political philosophy. For a solution he returned to the old, and now somewhat derided, idea of a social contract. According to Rousseau, each individual ceded his rights and powers to a general will of which he was part—that is, an equal participant in the political life. As he stated the terms, they logically ruled out a distinction between rights and powers reserved to the individual and the authority and functions assigned to the community, that issue which was of paramount importance to Locke. By Rousseau's reckoning, the individual was as free after the contract as he had been in the state of nature; he had merely substituted for natural freedom a civic freedom by which he was elevated into a moral being. Some later readers have found in *Du contrat social* dangerous undertones of totalitarianism because of the equivocal place of the individual in relation to the general will, because of the axiom that the general will was right, and because of a prescription of a civic religion with penalties for defiance. Other latter-day students have rejected the imputation. It may be permissible to conclude that Rousseau, in writing ostensibly a book on political theory, did not really have his mind on politics. In one respect, nonetheless, he gained in clarity over John Locke. The *Two Treatises of Government* had referred to the powers of an undefined majority; *Du contrat social* wound up in an emotionally suffused claim for political democracy.

Rousseau's *Du contrat social*

Both men had their triumphs. In America Thomas Jefferson incorporated strict Lockean doctrine in the Declaration of Independence. When shortly thereafter the French revolutionaries wrote their own declaration, they made a place for Rousseau's general will and sovereignty of the nation. On both sides of the Atlantic, however, there was fumbling over the fundamental question of popular government: who was to have the right to vote.

The mainstream of Enlightened thought had no quarrel with Locke's assertion that property was a natural right. Yet Rousseau, in agreement with Hume, was uneasy. Although he did not write down specific remedies, he was persuaded that political democracy could not work without a relative economic equality. In the seething atmosphere of the second half of the 18th century, a little band of dissenters took their criticisms far beyond Rousseau's doubts. In 1755 the obscure Morelly, in *Le Code de la nature*, attacked property as the mother of all crimes tormenting the world, and he made a proposal that was to have a long history: let every man contribute to the commonweal according to his abilities and receive according to his needs. Two decades later the Abbé de Mably, starting with equality as the law of nature, contended along with Morelly—and, for that matter, with the early Rousseau—that the introduction of property had destroyed the golden age of man. Soon, François-Noël Babeuf, civil servant turned radical journalist, tried to organize a revolution within a revolution to secure the true equality ordained by nature. In England William Godwin, taking his stand on the immutable laws of the universe, rigorously pushed reason into condemning not only property but even the state and marriage. Such scandalous utterances, however, were only tiny specks in the sky. As the old century was dying, property seemed to be resting on a firm ground of support.

The law. For the Enlightenment, the law under which men actually lived required urgent consideration. There could hardly have been a greater contrast between the law of the courts and that which reason allegedly perceived in nature. Once Grotius had given his definition, the law of all lands was summoned to the supreme tribunal of reason. Man-made law, said the indictment, was valid only insofar as it conformed with nature's law. Even the famous jurist William Blackstone, steeped in the English common law, accepted the argument. On the Continent, where there was little ground for contentment, the spirit of legal thought was uncompromising; the law, said the zealots, was not made by edicts and legislative enactments but was discovered by right reason. The instrument of reform was a rigorous code, a summation of legal wisdom that could be administered with mechanical precision. The outcome was partial victory in the Prussian and Austrian codes and in the Code Napoléon.

Economic theory. Another aspect of human affairs that the Enlightenment sought to judge was the economic. The practical background was a long history of extensive state intervention in economic pursuits—mercantilism—and a recent history of middle class dissatisfactions. In France the wars of Louis XIV had wrought great misery, and even before the King's death, critics had dared to propose reforms. Among the first to apply rationalist criteria was the Sieur de Boisguillebert, in whose judgment state interference did violence to the law of nature, which prescribed a harmony of interests; right reasoning under the stimulus of free competition was the proper guide to economic behaviour. A half century later Louis XV's physician, François Quesnay, founded the physiocratic school on essentially the same assumptions. The physiocrats' economic man was motivated by self-interest but lived under nature's dictate of harmony; it followed, therefore, that unfettered private activity was the certain means to the commonweal.

The question of state interference

Adam Smith's *Inquiry into the nature and causes of the Wealth of Nations* (1776), a declaration of economic independence, criticized the physiocrats' emphasis on the unique importance of land but otherwise revealed Smith's kinship. Smith began with the butcher and the baker intent on their own good and went on to demonstrate the merit of "the system of natural liberty." Sharing the British dislike of extremes, he was not dogmatic; but behind his array of observations and utilitarian arguments there lurked the old rational conviction about "the great Director of Nature" leading men "by an invisible hand" to serve a wider cause than their immediate own. Smith had no illusions about the way men conducted their business affairs, and he wanted the state to protect the weak against the strong. Nonetheless, his bias was for the free individual determining his own destiny. Smith spoke at the right moment and became a revered teacher of the following generations.

The meaning of history. The Enlightenment's thought about history began with a dilemma. The basic rationalist values were timeless, and the practical purpose of examining the world's past experience was to strip away the accidents of time and place and lay bare Hume's "constant and universal principles of human nature." Another service from studying history was required: when real men were so far removed from the man of reason and nature, it was the task of the historians to explain the causes of the gap. To their labours most of the students of the past brought a point of view that suggested the answers in advance. As a result, numerous historical

works glossed the theme of the evil consequences of religious and political benightedness. The histories of such as Voltaire, Guillaume Raynal, Hume, and Edward Gibbon, who recorded the triumph of barbarism and religion, were heavy guns in the arsenal of the Enlightenment. They were not, however, simply camouflaged propaganda. The Reformation had inspired solid historical scholarship, even if its purpose had been partisan, and the grave scholars of the 17th century had built on it to amass documentary materials and to broaden the skills of criticism. The luminous names of the 18th century had a surer grasp, and their masterpieces—for such many were—made Europe conscious of the pleasure and the importance of reading history.

This quietly intruding sense of historical change prompted reflection, not only on the relation of the Enlightenment to the past but also on what mankind could expect in the future. Earlier speculations were not useful guides for the latter. The ancient world had generally thought in terms of cycles in which there was a deterioration of ages from gold to iron. The Christians accepted from their predecessors the idea of cosmic cataclysm, but they would admit just one, which would come at God's own appointed time; in the meanwhile, the only significance of human history was the drama of salvation for individual souls. The Renaissance tended to return to the idea of the cycle, but it was that of youth, maturity, and decay of states after the manner of Aristotle's history of the Greek polis rather than the total upheaval of the cosmos assumed by the ancient world. The happy rebirth in 15th-century Italy was, then, merely the beginning of a new cycle. Still in the 18th century, when Montesquieu reflected on the causes of the greatness and decadence of the Romans, his interpretation continued the Renaissance approach to history.

The harbinger of a new way of thinking was Francis Bacon, the aforementioned prophet of science. The Greeks and Romans, he contended, belonged to the youth of the world, and his own generation was the heir of all accumulated knowledge. It was a mistake to be chained to the hopelessness of the notion of the cycle; by the proper method of inquiry, man could move on to great benefits through the conquest of nature. It was significant that whereas Thomas More had constructed his *Utopia* in a remote setting, Bacon put his *New Atlantis* in the future.

The idea of progress

Bacon had been thinking in terms of useful knowledge, and he found the Greeks and Romans the children and not the adults of time. Such talk was offensive to the literate community that had been raised on classical literature. For such people a comparison between the two ages had to be a comparison of their literatures, and for those spiritually suckled on Homer and Virgil there was no comparison. By the late 17th century, brazen writers began to defend the literature of their own time, and for a whole century cultural Europe resounded with the cut and thrust of the warfare between the "ancients" and the "moderns." The battles raged on the subjective field of taste, and a clear-cut victory was manifestly out of the question. Long before the end of this contest Bernard Fontenelle, permanent secretary of the Académie des Sciences, had shifted the ground from letters to scientific achievement, and few could deny him his claim for progress up to his day. But more significantly, Fontenelle was primarily concerned with the future, and for its course he detected a law of progress that would guide man's destiny to still better things.

As social dissatisfaction, and hopes, mounted late in the 18th century, the idea of scientific and intellectual progress quietly slipped into a belief in the general progress of mankind, moral no less than material. Hard-headed realists like Voltaire were too oppressed with the continuing perversities of the world to fall into an easy optimism; knowledge could be used for bad purposes as well as for good, and future improvements were dependent on the role of sound reason. Such a qualification, however, had little restraining value for those whose vision of the future had become assured. Young Anne-Robert-Jacques Turgot, the future minister of Louis XVI, surveyed all human history and found it a record of the progress of the race toward perfection; even in ages of apparent barbarism, men and society had moved, however slowly, in the right direction and in the future would continue to do so under a law of acceleration. Turgot's friend the Marquis de Condorcet converted these calm anticipations into fervent prophecy. Hiding in 1793 from the revolution that he had joyfully served, Condorcet hastily composed his *Esquisse d'un tableau historique des progrès de l'esprit humain*. He saw mankind facing its tenth and culminating epoch—a new day in which the steady march toward perfectibility could not be stopped. The time would come, said this man under sentence of death, when the sun would shine only on free men who knew no master but their reason.

The end of the Enlightenment. A chronology of the disintegration of the Enlightenment is impossible. For all of its belief in eternals, it was from the beginning a flow of ideas in which emphases shifted and trains of thought broke under the impact of new evidence and new insights. Newton's physics and Locke's epistemology were absorbed into the uneasy synthesis, but eventually they bore a heavy responsibility for the decay. Natural law as perceived by right reason was, for many of the latter-day lovers of science, sinking into a bloodless cliché, while the pleasure–pain doctrine built up its empire of utilitarianism. At the high noon of religious rationalism, John Wesley cried defiance with his emotional Methodism. Nature, once considered a synonym of reason and visible proof of the existence of God and his benevolence, broke up into something to be studied with scientific objectivity and something to be enjoyed in romantic indulgence. Not by any means the least blow to the Enlightenment's timeless verities was the revolutionary growth of historical mindedness, of the turn toward rethinking all mortal affairs as well as scientific questions in terms of ceaseless change. In the dying days of the Enlightenment the Western mind was setting out on the road that led to Edmund Burke and Hegel and to Darwin and Marx.

Later times have ridiculed a naïve epoch, but the Enlightenment's moral and intellectual significance is high because it tried to rethink in new idioms the oldest values and beliefs in the history of Western civilization.

BIBLIOGRAPHY. NORMAN HAMPSON, *The Enlightenment*, vol. 4 of the *Pelican History of European Thought* (1968), an introduction to the 18th century—full of knowledge and spirit; PRESERVED SMITH, *A History of Modern Culture*, 2 vol. (1930–34), wisdom expressed in good style is not soon out of date; PETER GAY, *The Enlightenment*, vol. 1, *The Rise of Modern Paganism* (1966), vol. 2, *The Science of Freedom* (1969), a work of distinguished scholarly labour mindful of the complexities in the century's struggle to repudiate Christianity and to rise above classical thought to modernity—the critical bibliographical essays are indispensable for any serious student; ERNST CASSIRER, *Die Philosophie der Aufklärung* (1932; Eng. trans., *The Philosophy of the Enlightenment*, 1951), a classic work by a profound German scholar; PAUL HAZARD, *La Crise de la conscience européenne* (1935; Eng. trans., *The European Mind, 1680–1715*, 1953) and *La Pensée européenne au XVIIIème siècle, de Montesquieu à Lessing* (1946; Eng. trans., *European Thought in the Eighteenth Century, from Montesquieu to Lessing*, 1954), by a contrasting French scholar who puts a light touch on learning; CARL LOTUS BECKER, *The Heavenly City of the Eighteenth Century Philosophers* (1932; paperback reprint, 1959), a much appreciated set of lectures in which a gentle cynic finds it amusing that the philosophers were doing essentially what their forebears had been doing (PETER GAY has reproved him in *The Party of Humanity: Essays in the French Enlightenment*, 1964); GEORGE R. HAVENS, *The Age of Ideas* (1955), scholarly but perhaps over-enthusiastic biographies of the principal French *philosophes;* GEORGE H. SABINE, *A History of Political Theory*, 3rd ed. (1961), the best survey of this vast field by a philosopher of great and dispassionate learning; A.R. HALL, *The Scientific Revolution, 1500–1800*, 2nd ed. (1962), a standard work on an essential aspect of the Enlightenment; J.B. BURY, *The Idea of Progress: An Inquiry into Its Origin and Growth* (1920), the famous classic, which might profitably be read with CHARLES FRANKEL, *The Faith of Reason: The Idea of Progress in the French Enlightenment* (1948); and HENRY VYVERBERG, *Historical Pessimism in the French Enlightenment* (1958).

(Da.H.)

Entoprocta

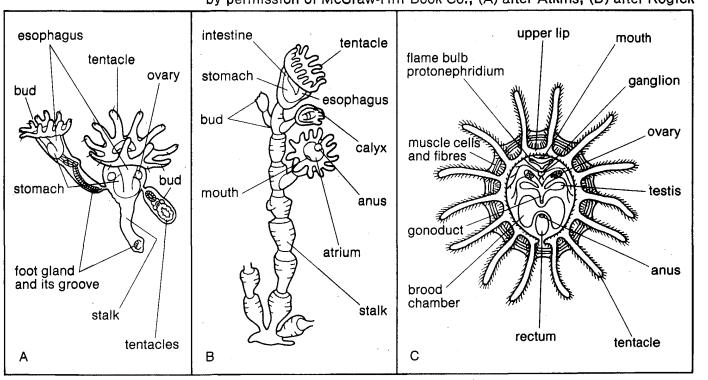
The Entoprocta, a small phylum of aquatic invertebrates, are found attached to some object (*i.e.*, are sedentary). Some entoprocts live as individuals; others form colonies. Most are marine: *Pedicellina*, for example, forms clumps on seaweeds or bryozoans (small marine invertebrates) that are found near low tidemark.

General features

Entoproct individuals are called zooids; a *Pedicellina* colony, for example, consists of a series of distinct zooids, which originate from a creeping, branching stolon (or stemlike attachment structure) and attain a height of about two millimetres, occasionally four or five millimetres. Although they are small, zooids may be so numerous that the substance supporting them has a fuzzy appearance. The body of a zooid consists of a stalk bearing a tentacle-crowned head, called the calyx. The calyx is soft; the stalk may be soft or hard. Each *Pedicellina* zooid, for example, consists of a globular calyx that is linked to the stolon by a flexible supporting stalk, which can move from its base in a jerky, sweeping manner. The outer surface of the calyx bears a circle of tentacles; most of the interior of the calyx contains the alimentary canal, which is U shaped and has both mouth and anus encircled by tentacles. The anus is raised on a papilla, leaving a space (or atrium) between the rectum and the mouth. Related to *Pedicellina* is *Barentsia*, which has a jointed stalk of rigid segments, and the freshwater *Urnatella*, which has a nodular stalk.

Less familiar genera, although they are represented by many more species, are *Loxosoma* and *Loxosomella;* individuals are not joined by stolons but often occur in dense aggregations known as loxosome populations. The individuals are very small (often only 0.5 millimetre high) and are of biological interest because they live in intimate associations with other animals (*e.g.*, polychaetes, sipunculids, sponges, and bryozoans). Loxosomes apparently are energy commensals; they obtain food from the respiratory or feeding currents of the host animal.

From *McGraw-Hill Encyclopaedia of Science and Technology* (1960); reproduced by permission of McGraw-Hill Book Co., (A) after Atkins, (B) after Rogick



Structure of entoprocts.
(A) *Loxosoma singulare* zooid with two buds. (B) *Urnatella gracilis* colony consisting of a flexible stalk with urnlike segments. (C) *Pedicellina cernua* calyx (top view).

Budding and reproduction

Natural History. Entoprocts reproduce asexually and sexually. Asexual reproduction, called budding or blastogenesis, gives rise to both the pedicellinid colony and the loxosome population. In pedicellinids the stolons are blastogenic; the colony grows by the linear extension and the branching of stolons, which produce at their tips a succession of buds that differentiate into zooids.

In the loxosomes, which lack stolons, the calyx is blastogenic and commonly produces two series of buds. The daughter zooids break free and may either swim away, stalk foremost (resembling miniature torpedoes), or use a "foot" on the end of the stalk to move over the host's surface. In *Loxosomella* the foot contains a cement gland. In *Loxosoma*, however, the stalk has a sucking disc at the end, and a zooid of this genus retains the ability to move throughout its life, using foot and tentacles alternately as points of attachment to a surface.

Entoprocts can disperse themselves over large regions by sexually produced larvae. Zooids may be of one sex or hermaphroditic, having both male and female sexual organs in one animal, in which case the testes mature first. In both types, ducts from the paired gonads unite and open into the atrium (*i.e.*, the space between rectum and mouth), and fertilization is internal. Several young embryos attach themselves to a part of the atrial wall in front of the rectum that supplies nutriment for the embryos, which develop and eventually are released as larvae. In *Urnatella* one larva develops inside an ovary.

The entoproct larva resembles an annelid larva; in the entoproct larva, however, the girdle of locomotory cilia (*i.e.*, hairlike processes) is not located in an equatorial position but is found near the basal pole. The *Pedicellina* larva, which has a structural resemblance to an inverted and nontentaculate adult calyx, attaches base downwards (with mouth and anus toward the surface). In the metamorphosis (change) of a larva into a zooid, the gut is reoriented so that mouth and anus are in the adult positions. Although loxosome larvae may metamorphose, daughter zooids usually are budded off (*i.e.*, set free) directly from the larvae. Daughter zooids also may form internally, the larva rupturing to release the zooid.

Entoprocts with stolons can regenerate (*i.e.*, replace) lost parts of their bodies. Lost calyces, for example, are replaced quickly. *Barentsia* and *Urnatella* colonies survive the winter as dormant buds or stolon sections which germinate into new colonies in the spring.

Feeding and digestion

Form and function. All entoprocts feed on organic particles that are filtered from the surrounding water. Ciliated tentacles on the outer margin of the calyx spread widely. Long lateral cilia drive water inward between the tentacles. The inner surface of each tentacle bears a tract of short frontal cilia. The tract descends into a ciliated groove, which passes the tentacle bases and extends to the mouth. Food particles, thrown by the long lateral cilia onto the short frontal cilia are trapped in mucus, carried down the tentacles, around the ciliated groove, and into the mouth. Unsuitable particles are removed by the water flow. If disturbed, the tentacles can be folded inward so that they lie within a muscular membrane that unites their bases.

The body wall consists of an outer, cuticle-covered epidermis (or skin), below which (particularly in the stalk) is a layer of muscle fibres. The primitive body cavity is filled with connective tissue and fluid. In colonial entoprocts, muscle cells at the junction of the calyx and the stalk contract rhythmically every few seconds and may serve as pumps to assist the movement of nutrients from the calyces toward the growing tips of the stolons.

In marine entoprocts two flame-cell nephridia (*i.e.*, excretory tubules) share an opening into the atrium. The freshwater genus, *Urnatella*, has many nephridia, which have an osmoregulatory (*i.e.*, water-removing) function.

From a double ganglion above the stomach, nerves radiate to the tentacles and to some parts of the viscera.

Distinguishing characteristics

Evolution and classification. Entoprocts are characterized by stalked, pseudocoelomate zooids that contain a U-shaped gut (*i.e.*, both mouth and anus open within a circle of 6 to 36 ciliated tentacles). The larva and the pattern of its early development indicate that the entoprocts may be linked remotely with the Annelida and their related phyla. There is no evidence of a close link with any phylum.

First described in 1774, entoprocts were classified with the Bryozoa, in which filter-feeding, tentaculated zooids also are present. The similarities between the two groups, however, are superficial. Methods of food capture, for example, differ fundamentally; in addition, bryozoan zooids are coelomate, and the anus opens outside the tentacles. The name Entoprocta was introduced in 1869, and the group was first recognized as a phylum in 1888. Alternative names (Calyssozoa and Kamptozoa) were proposed in 1921 and 1929.

The phylum is small and has no known fossil members. The subdivision into three families does not accord fully with the considerable differences between colonial and

noncolonial forms. It might be more accurate if two orders were recognized; the first order would include the families Pedicellinidae (about 30 species classified into 6 genera according to stalk structure) and Urnatellidae (2 species), and the second order would contain only the Loxosomatidae (about 100 species).

BIBLIOGRAPHY. Detailed, though no longer fully up-to-date, accounts are: L.H. HYMAN, *The Invertebrates*, vol 3, *Acanthocephala, Aschelminthes and Entoprocta*, pp. 521–554 (1951); P. BRIEN, "Classe des endoproctes ou kamptozoaires," in P.P. GRASSE (ed.), *Traité de zoologie*, vol. 5, pp. 927–1007, (1959), in French.

(J.S.R.)

Enzyme

Enzymes comprise an important group of proteins that make possible the chemical reactions that constitute metabolism. They are efficient and specific—that is, they accelerate the rate of one kind of chemical reaction of one type of compound, and they do so far more efficiently than man-made catalysts. They act as biological catalysts, or accelerators, of the thousands of successive complex biochemical reactions that take place in all living things. Enzymes are controlled by activators and inhibitors that initiate or block reactions. All cells contain enzymes, which usually vary in number and composition, depending on the cell type; an average mammalian cell, for example, is about one one-billionth (10^{-9}) the size of a drop of water and contains about 3,000 enzymes.

GENERAL FEATURES

Historical background. Although little was known about the chemical nature of enzymes until the 20th century, enzymatic reactions have been used by man since prehistoric times. Among man's earliest discoveries was that the juice of grapes ferments and turns into wine, a process attributed by the Greeks to the god Bacchus. Not until the 19th century was it understood that the conversion of grape sugar to alcohol is a process catalyzed by enzymes found in yeast. Another enzymatic action was discovered by chance when primitive herdsmen filled canteens consisting of the stomachs of slaughtered animals with milk instead of water and found that the milk soon formed clumps of cheese. The reason for the transformation, the action on milk of an enzyme (rennin) released from the lining of animal stomachs, was not known until centuries later. Rennin from calves' stomachs is still widely used in the production of cheese. Throughout history, accidental discoveries of enzymatic reactions have enabled man to carry out other processes; *e.g.*, the leavening of bread, the brewing of beer, the formation of vinegar, and the tenderizing of meat.

The existence of enzymes was established in the middle of the 19th century by scientists studying the process of fermentation. Their role as catalysts of all living things followed rapidly. Developments before 1850 included (in 1833) the separation from malt of the enzyme amylase, which converts starch into sugar, and (in 1836) the isolation from the stomach wall of animals of a component of gastric juice that could partially digest food in a test tube, the enzyme pepsin.

Enzymes were known for many years as ferments, from the Latin word for yeast. In 1878 the name enzyme, from the Greek words meaning "in yeast," was introduced and has since been universally used.

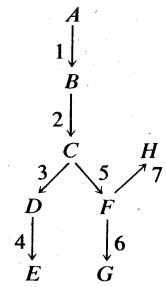
Importance. *In biology.* Almost all of the numerous and complex biochemical reactions that take place in animals, plants, and micro-organisms are regulated by enzymes.

The role of enzymes in metabolism

Some enzymes help to break down large nutrient molecules, such as proteins, fats, and carbohydrates, into smaller molecules. This process occurs during the digestion of foodstuffs in the stomach and intestines of animals. Other enzymes guide the small molecules through the intestinal wall into the bloodstream. Still other enzymes promote the formation of large, complex molecules from the small, simple ones to produce cellular constituents. Enzymes are also responsible for other functions; *e.g.*, the storage and release of energy, the course of reproduction, the processes of respiration, vision, etc. They are indispensable to life.

Each enzyme is able to promote only one type of chemical reaction. The compounds on which it acts are called substrates. Enzymes operate in tightly organized metabolic systems called pathways. A seemingly simple biological phenomenon—the contraction of a muscle, for example, or the transmission of a nerve impulse—actually involves a number of chemical steps in which one or more chemical compounds (substrates) are converted to substances called products; the product of one step in a metabolic pathway serves as the substrate for the succeeding step.

The role of enzymes in metabolic pathways can be illustrated diagrammatically. The chemical compound represented by *A* (see diagram) is converted to product *E* in a series of enzyme-catalyzed steps, in which intermediate compounds represented by *B*, *C*, and *D* are formed in succession. They act as substrates for enzymes represented by 2, 3, and 4. Compound *A* may also be converted by another series of steps, some of which are the same as those in the pathway for the formation of *E*, to products represented by *G* and *H*.



The letters represent chemical compounds; numbers represent enzymes that catalyze individual reactions. The relative heights represent the thermodynamic energy of the compounds; *e.g.*, compound *A* is more energy-rich than *B*, *B* more energy-rich than *C*. Compounds *A*, *B*, etc., change very slowly in the absence of a catalyst but do so rapidly in the presence of catalysts 1, 2, 3, etc.

The regulatory role of enzymes in metabolic pathways can be clarified by using a simple analogy: that between the compounds, represented by letters in the diagram, and a series of connected water reservoirs on a slope. Similarly, the enzymes represented by the numbers are analogous to the valves of the reservoir system. The valves control the flow of water in the reservoir; that is, if only valves 1, 2, 3, and 4 are open, the water in *A* flows only to *E*, but, if valves 1, 2, 5, and 6 are open, the water in *A* flows to *G*. In a similar manner, if enzymes 1, 2, 3, and 4 in the metabolic pathway are active, product *E* is formed, and, if enzymes 1, 2, 5, and 6 are active, product *G* is formed. The activity or lack of activity of the enzymes in the pathway therefore determines the fate of compound *A; i.e.*, it either remains unchanged or is converted to one or more products. In addition, if products are formed, the activity of enzymes 3 and 4 relative to that of enzymes 5 and 6 determines the quantity of product *E* formed compared with product *G*.

Both the flow of water and the activity of enzymes obey the laws of thermodynamics; hence, water in reservoir *F* cannot flow freely to *H* by opening valve 7, because water cannot flow uphill. If, however, valves 1, 2, 5, and 7 are open, water flows from *F* to *H*, because the energy conserved during the downhill flow of water through valves 1, 2, and 5 is sufficient to allow it to force the water up through valve 7. In a similar way, enzymes in the metabolic pathway cannot convert compound *F* directly to *H* unless energy is available; enzymes are able to utilize energy from energy-conserving reactions in order to catalyze reactions that require energy. During the enzyme-catalyzed oxidation of carbohydrates to carbon dioxide and water, energy is conserved in the form of an energy-rich compound, adenosine triphosphate (ATP). The energy in ATP is utilized during an energy-consuming process such as the enzyme-catalyzed contraction of muscle.

Because the needs of cells and organisms vary, not only the activity but also the synthesis of enzymes must be regulated; *e.g.*, the enzymes responsible for muscular activity in a leg muscle must be activated and inhibited at appropriate times. Some cells do not need certain enzymes; a liver cell, for example, does not need a muscle enzyme. A bacterium does not need enzymes to metabolize substances that are not present in its growth medium. Some enzymes, therefore, are not formed in certain cells, others are synthesized only when required, and still others are found in all cells (see also GENE). The formation and activity of enzymes are regulated not only by genetic mechanisms but also by organic secretions (hormones) from endocrine glands and by nerve impulses. Small molecules also play an important role (see below *Enzyme flexibility and allosteric control*).

The significance of enzymes in disease

If an enzyme is defective in some respect, disease may occur. The enzymes represented by the numbers 1 to 4 in the diagram must function during the conversion of the starting substance *A* to the product *E*. If one step is blocked because an enzyme is unable to function, product *E* may not be formed; if *E* is necessary for some vital function, disease results. Many inherited diseases of man result from a deficiency of one enzyme. Some of these are listed in Table 1. The disease called albinism, for example, results from an inherited lack of ability to synthesize the enzyme tyrosinase, which catalyzes one step in the pathway by which the pigment for hair and eye colour is formed (see also METABOLISM, DISEASES OF; BIRTH DEFECTS AND CONGENITAL DISORDERS).

Table 1: Enzymes Identified with Hereditary Diseases

disease name	defective enzyme
Albinism	tyrosinase
Phenylketonuria	phenylalanine hydroxylase
Fructosuria	fructokinase
Methemoglobinemia	methemoglobin reductase
Galactosemia	galactose-1-phosphate uridyl transferase

In medicine. Enzymes play an increasingly important role in medicine. The enzyme thrombin is used to promote the healing of wounds. Other enzymes are used to diagnose certain kinds of disease, to cause the remission of some forms of leukemia—a disease of the blood-forming organs—and to counteract unfavourable reactions in people who are allergic to penicillin. The enzyme lysozyme, which destroys cell walls, is used to kill bacteria. Research concerning medical applications of enzymes may lead to their use as preventives of tooth decay and as anticoagulants in the treatment of thrombosis, a disease characterized by the formation of a clot, or plug, in a blood vessel. Enzymes may eventually be used to control enzyme deficiencies and abnormalities resulting from diseases.

In industry. The most efficient catalysts known to man, enzymes are used in industrial processes involving the preparation of certain chemical compounds and the tanning of leather; they are valuable in analytical procedures involving the detection of very small quantities of specific substances. Enzymes are necessary in such food-related industries as cheese making, the brewing of beer, the aging of wine, and the baking of bread. Enzymes also may be used to clean clothes. For industrial use of enzymes see BAKING AND BAKERY PRODUCTS; BREWING; and WINE MAKING.

GENERAL PROPERTIES

Classification and nomenclature. The first enzyme name, proposed in 1833, was diastase. Sixty-five years later, it was suggested that all enzymes be named by adding "-ase" to a root indicative of the nature of the substrate of the enzyme. Although enzymes are no longer named in such a simple manner, with the exception of a few—*e.g.*, pepsin, trypsin, chymotrypsin, papain—most enzyme names do end in "-ase."

Any systematic classification of enzymes should be based on a common property or quality that varies sufficiently to be useful as a distinguishing feature. In this regard, three properties of enzymes could serve as a basis for enzyme classification—the exact chemical nature of the enzyme, the chemical nature of the substrate, and the nature of the reaction catalyzed. Adequate information about the detailed chemical structures of more than a few enzymes does not yet exist. In addition, although, as indicated above, early attempts at enzyme classification were based on the nature of broad groups of substrates (*e.g.*, enzymes called carbohydrases act on carbohydrates), close functional similarities among enzymes in different groups were often obscured. By general agreement, then, enzymes now are classified according to their substrates and the nature of the reaction they catalyze.

Systematic and trivial names

In an attempt to devise a rational system of enzyme nomenclature, two names are given to an enzyme. One, known as the systematic name, is based on logical principles but is often long and awkward; the other, "trivial" name is short and generally used but not usually exact or systematic (see Table 2). In the scheme of systematic nomenclature, six main groups of enzymatic reactions are recognized; each catalyzes one reaction type and is subdivided on the basis of detailed definitions of the reaction catalyzed and of the substrate involved in the reaction. Enzymes that catalyze reactions in which hydrogen is transferred belong to the group known as oxidoreductases; those that catalyze the introduction of the elements of water at a specific site in a molecule are called hydrolases. The other four groups of reactions are the transferases—which catalyze reactions in which substances other than hydrogen are transferred—the lyases, the isomerases, and the ligases. Oxidoreductases and transferases account for about 50 percent of the approximately 1,000 enzymes recognized thus far. Table 2 lists a few enzymes, their trivial names, their systematic names, and their biological roles.

Chemical nature. Little was known about the chemical nature of enzymes until the beginning of the 20th century, although scientists were almost convinced that they were proteins. In 1926, the enzyme urease was the first to be crystallized and clearly identified as a protein. Within the next few years, the digestive enzymes pepsin, trypsin, and chymotrypsin were shown to be proteins. Since that time, hundreds of enzymes, all of them proteins, have been prepared and characterized by chemical methods. Much of the knowledge of protein chemistry has, in fact, resulted from studies involving enzymes and from attempts to understand their nature and mode of action.

Although some enzymes consist of a single chain of the amino acids (*i.e.*, simple organic molecules containing nitrogen), most enzymes are composed of more than one chain. Each chain is called a subunit. Many enzymes have two, four, or six subunits, and some consist of as many as 12 to 60 subunits. In many cases, the subunits have identical structures; in others, however, several different types of subunit chains are involved. (See PROTEIN for a detailed discussion of the chemical nature of proteins.)

Much of the dry weight of plants and animals is attributable to proteins; with the exception of proteins that act as structural elements, most of the proteins in physiologically active tissues such as kidney and liver are enzymes. Regardless of the exact amount of enzymatic protein in an organism, it is clear that hundreds of different enzymes must be present in each tissue to account for the myriad reactions comprising metabolism.

Cofactors. Although some enzymes consist only of protein, many are complex proteins; *i.e.*, they have a protein component and a so-called cofactor. A complete enzyme is called a holoenzyme; if the cofactor is removed, the protein, no longer enzymatically active, is called the apoenzyme. A cofactor may be a metal—such as iron, copper, or magnesium—a moderately sized organic molecule called a prosthetic group, or a special type of substrate molecule known as a coenzyme. The cofactor may aid in the catalytic function of an enzyme, as do metals and prosthetic groups, or take part in the enzymatic reaction, as do coenzymes.

Coenzymes

A coenzyme serves as a type of substrate in certain enzymatic reactions and thus reacts in the exact propor-

Table 2: Classification of Some Enzymes

systematic name*		trivial name	reaction catalyzed	biological role
code number†	name‡			
1.1.1.1	alcohol: NAD oxidoreductase	alcohol dehydrogenase	alcohol + NAD → acetaldehyde NADH	alcoholic fermentation
1.1.1.27	L-lactate: NAD oxidoreductase	lactic dehydrogenase	lactate + NAD → pyruvate + NADH	carbohydrate metabolism
2.7.1.40	ATP: pyruvate phospho transferase	pyruvate kinase	pyruvic acid + ATP → phosphoenol-pyruvic acid + ADP	carbohydrate metabolism
3.1.1.7	acetylcholine: acetyl hydrolase	acetylcholinesterase	acetylcholine + H_2O → acetate + choline	nerve-impulse conduction
3.4.4.13	peptide peptido hydrolase	thrombin	peptide bond + $H_2O \rightarrow -\overset{O}{\overset{\|}{C}}-OH + H_3{}^+N-$	blood-clotting mechanism

*Based on recommendations (1964) of the International Union of Biochemistry. †The numbering system is as follows: the first number places the enzyme in one of six general groups—1, oxidoreductases; 2, transferases; 3, hydrolases; 4, lyases; 5, isomerases and 6, ligases. The second number places the enzyme in a subclass based on substrate type or reaction type; *e.g.*, the enzyme may act on molecules with —CHOH groups. The third number places the enzyme in a subsubclass, which specifies the reaction type more fully; *e.g.*, NAD coenzyme required. The fourth number is the serial number of the enzyme in its subsubclass. ‡NAD and NADH represent the oxidized and reduced forms of nicotinamide adenine dinucleotide (NAD), respectively; ATP and ADP represent adenosine triphosphate and adenosine diphosphate, respectively.

tions (*i.e.*, stoichiometrically) required for reaction, rather than in catalytic quantities. A coenzyme may, for example, assume the role of a hydrogen acceptor, as does nicotinamide adenine dinucleotide (NAD), which accepts hydrogen from the substrate, or a chemical-group donor, as does adenosine triphosphate (ATP), which donates phosphoric acid to the substrate. After ATP has donated a phosphoric acid molecule to the substrate, the phosphoric acid can be reacquired in a second stoichiometric reaction catalyzed by a second enzyme. The catalytic nature of a coenzyme is apparent only when it couples the activities of two enzymes in this way. Coenzymes thus are the links, or shuttles, in metabolic pathways that enable substances—*e.g.*, hydrogen, phosphoric acid—to be exchanged.

An interesting aspect of coenzymes is that many of them are closely related to vitamins, which are necessary for the proper functioning of life processes. Many vitamins are essential for the synthesis of coenzymes; since coenzymes play a vital role in metabolism, it is not difficult to understand why certain vitamins are essential. Table 3 lists some coenzymes, the type of enzymes with which they react, and the vitamins required for their synthesis in mammals (see also VITAMIN; NUCLEOTIDES).

Table 3: The Relationship Between Coenzymes and Vitamins

name of coenzyme	some enzymes with which it reacts	vitamin needed for synthesis
Adenosine triphosphate	kinases, muscle enzymes	none
Nicotinamide adenine dinucleotide	dehydrogenases	nicotinic acid
Pyridoxal phosphate	transaminases	pyridoxine
Biotin	carboxylases	biotin
Coenzyme A	acyl transferases	pantothenic acid

THE NATURE OF ENZYME-CATALYZED REACTIONS

The nature of catalysis. In a chemical reaction—for example, one in which substance *A* is converted into product *B*—a point of equilibrium eventually is reached at which no further chemical change occurs; *i.e.*, the rate of conversion of *A* to *B* equals the rate of conversion of *B* to *A*. The so-called thermodynamic-equilibrium constant expresses this chemical equilibrium. A catalyst may be defined as a substance that accelerates a chemical reaction but is not consumed in the process. The amount of catalyst has no relationship to the quantity of substance altered; very small amounts of enzymes are very efficient catalysts. Because the presence of an enzyme accelerates the rate of conversion of a compound to a product, it accelerates the approach to equilibrium; it does not, however, influence the equilibrium point attained.

The molecules in the watery medium of the cell are in constant thermal motion but, because they are more or less stable compounds, they would react only occasionally to form products in the absence of enzymes. There exists an energy barrier to the reaction of a molecule. The energy required to overcome the barrier to reaction is called the energy of activation. A reaction proceeds to equilibrium only if the molecules have sufficient energy of activation to form an activated complex, from which products can be derived. Enzymes greatly increase the chances for reactions by their ability to make large numbers of specific molecules more reactive (*i.e.*, unstable) by forming intermediate compounds with them. The unstable intermediates quickly break down to form stable products, and the enzymes, unchanged by the reaction, are able to catalyze the formation of additional products.

The role of the active site. That the compound on which an enzyme acts (substrate) must combine in some way with it before catalysis can proceed is an old idea, now supported by much experimental evidence. The combination of substrate molecules with enzymes involves collisions between the two. Enzymes are large molecules, the molecular weights of which (based on the weight of a hydrogen atom as 1) range from several thousand to several million. The substrates on which enzymes act usually have molecular weights of several hundred. Because of the difference in size between the two, only a fraction of the enzyme is in contact with the substrate; the region of contact is called the active site. Usually, each subunit of an enzyme has one active site capable of binding substrate.

The characteristics of an enzyme derive from the sequence of amino acids, which determine the shape of the enzyme (*i.e.*, the structure of the active site) and hence the specificity of the enzyme.

Figure 1A shows the way in which the substrate fits into the active site of an enzyme. A small portion of the structure of the enzyme is represented by the various dark lines and symbols. The triangles labeled *A* and *B* represent the amino acids at the active site that actually catalyze the reaction in which the bond (indicated by the indentation) in the substrate molecule is broken as a product forms. The circles in the enzyme structure represent amino acids that function in attracting certain regions of the substrate, so that it is held in the proper position to allow the reaction at the active site to proceed.

Forces attracting enzyme and substrate

The forces that attract the substrate to the surface of an enzyme may be of a physical or a chemical nature. Electrostatic bonds may occur between oppositely charged groups—the circles containing plus and minus signs on the enzyme are attracted to their opposites in the substrate molecule. Such electrostatic bonds can occur with groups that are completely positively or negatively charged (*i.e.*, ionic groups) or with groups that are partially charged (*i.e.*, dipoles). The attractive forces between substrate and enzyme may also involve so-called hydrophobic bonds, in which the oily, or hydrocarbon, portions of the enzyme (represented by H-labeled circles) and the

substrate are forced together in the same way as oil droplets tend to coalesce in water.

Modifications in the structure of the amino acids at or near the active site usually affect the enzyme's activity, because these amino acids are intimately involved in the fit and attraction of the substrate to the enzyme surface. The characteristics of the amino acids near the active site determine whether or not a substrate molecule will fit into the site. A molecule that is too bulky in the wrong places, as indicated by the arrow in Figure 1B, cannot fit into the active site and thus cannot react with the enzyme. In a similar manner, a molecule lacking essential attractive forces (indicated by the arrow in 1C) or the appropriately charged regions (indicated by the arrow in 1D) might not be bound to the enzyme. On the other hand, a molecule with a bulky group (1E) at a position such that it does not interfere with the binding of the molecule to the enzyme or with the function of the active site is able to serve as a substrate for the enzyme. The idea of a fit between substrate and enzyme, called the "key–lock" hypothesis, was proposed by a German chemist, Emil Fischer, in 1899 and explains one of the most important features of enzymes, their specificity. In most of the enzymes studied thus far, a cleft, or indentation, into which the substrate fits (see 1A), is found at the active site.

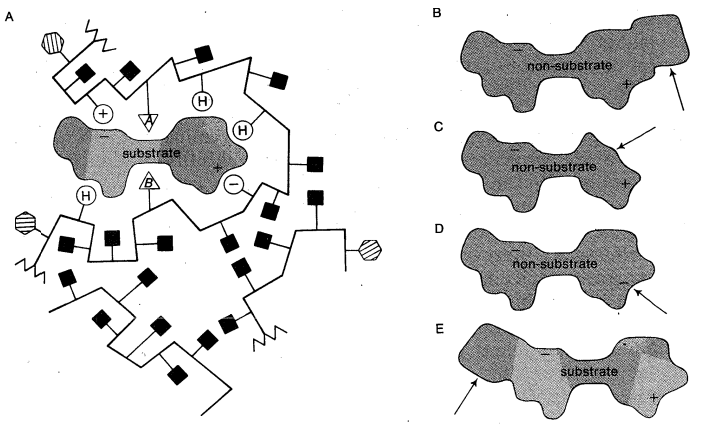


Figure 1: The role of active site in lock and key fit of a substrate—the key—to an enzyme—the lock (see text).

The specificity of enzymes. *Substrate specificity.* Since the substrate must fit into the active site of the enzyme before catalysis can occur, only properly designed molecules can serve as substrates for a specific enzyme; in many cases, an enzyme will react with only one naturally occurring molecule. Two oxidoreductase enzymes from Table 2 will serve to illustrate the principle of enzyme specificity. One (alcohol dehydrogenase) acts on alcohol, the other (lactic dehydrogenase) on lactic acid; the activities of the two, even though both are oxidoreductase enzymes, are not interchangeable—*i.e.*, alcohol dehydrogenase will not catalyze a reaction involving lactic acid or vice versa, because the structure of each substrate differs sufficiently to prevent its fitting into the active site of the alternative enzyme. Enzyme specificity is essential because it keeps separate the many pathways, involving hundreds of enzymes, that function during metabolism.

Figure 1 illustrates enzyme specificity. As mentioned above, the molecules in B, C, and D cannot serve as substrates for the enzyme because they are either too large (B), too small (C), or too repulsive because of charge (D) to bind to the enzyme's active site. Molecules the structural differences of which do not affect the active site (see arrow, Figure 1E), are able to react with the enzyme and are substrates.

Not all enzymes are as highly specific as the example in Figure 1; digestive enzymes such as pepsin and chymotrypsin, for example, are able to act on almost any protein, as they must if they are to act upon the varied types of proteins consumed as food. On the other hand, thrombin, which reacts only with the protein fibrinogen, is part of a very delicate blood-clotting mechanism and thus must act only on one compound in order to maintain the proper functioning of the system.

When enzymes were first studied, it was thought that most of them were "absolutely specific"—*i.e.*, that they would react with only one compound. In most cases, however, a molecule other than the natural substrate can be synthesized in the laboratory; it is enough like the natural substrate to react with the enzyme. Use of these synthetic substrates has been valuable in understanding enzymatic action. It must be remembered, however, that, in the living cell, many enzymes are absolutely specific for the compounds found there.

Reaction specificity. All enzymes isolated thus far are specific for the type of chemical reaction they catalyze—*i.e.*, oxidoreductases do not catalyze hydrolase reactions, and hydrolases do not catalyze reactions involving oxidation and reduction. An enzyme therefore catalyzes a specific chemical reaction but may be able to do so on several similar compounds.

The mechanism of enzymatic action. An enzyme attracts substrates to its active site, catalyzes the chemical reaction by which products are formed, and then allows the products to dissociate—*i.e.*, separate from the enzyme surface. This sequence is shown in Figure 2. The combination formed by an enzyme and its substrates is called the enzyme–substrate complex. When two substrates and one enzyme are involved, the complex is called a ternary complex; one substrate and one enzyme are called a binary complex. The substrates are attracted to the active site by electrostatic and hydrophobic forces, which are called noncovalent bonds because they are physical attractions and not chemical bonds (see above *The role of the active site*).

The enzyme–substrate complex

In Figure 2, two substrates (S_1 and S_2) bind to the active site of the enzyme during step (1) and react to form products (P_1 and P_2) during step (2). The products dissociate from the enzyme surface in step (3), releasing the enzyme. The enzyme, unchanged by the reaction, is able to react with additional substrate molecules in this manner many times per second to form products. The step in which the actual chemical transformation occurs—step (2) of Figure 2—is of great interest, and, although much is known about it, it is not yet fully understood. In general there are two types of enzymatic mechanisms, one in which a so-called covalent intermediate forms (see below) and one in which none forms (Figure 2B). Figure 2A and B show in detail the catalytic events that occur during an enzyme-catalyzed reaction.

Figure 2A shows the mechanism by which a covalent intermediate—*i.e.*, an intermediate with a chemical bond between substrate and enzyme—forms. One substrate, $B - X$ (in 2A), reacts with the group N on the enzyme surface to form an enzyme-B intermediate compound. The intermediate compound then reacts with the second substrate, Y, to form the products $B - Y$ and X.

Many enzymes catalyze reactions by this type of mechanism. Acetylcholinesterase (Table 2 and Figure 2A) is used as a specific example in the sequence described below. The two substrates (S_1 and S_2) for acetylcholin-

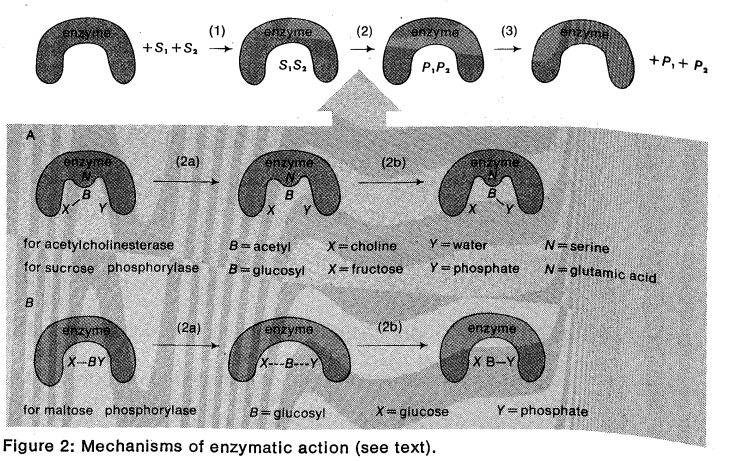


Figure 2: Mechanisms of enzymatic action (see text).

esterase are acetylcholine (the $B - X$ of Figure 2A) and water (the Y of Figure 2A). After acetylcholine ($B - X$) binds to the enzyme surface, a chemical bond forms between the acetyl moiety (B) of acetylcholine and the group N (part of the amino acid serine) on the enzyme surface. The result of the formation of this bond, called an acyl–serine bond, is one product, choline (X), and the enzyme-B intermediate compound (an acetyl–enzyme complex). The water molecule (Y) then reacts with the acyl–serine bond to form the second product, acetic acid ($B - Y$), which dissociates from the enzyme. Acetylcholinesterase is regenerated and is again able to react with another molecule of acetylcholine. This kind of reaction, involving the formation of an intermediate compound on the enzyme surface, is generally called a double displacement reaction.

Sucrose phosphorylase (see Figure 2A) acts in a similar way. The substrate for sucrose phosphorylase is sucrose, or glucosyl-fructose ($B - X$), and the group N on the enzyme surface is a chemical group called a carboxyl group (COOH). The enzyme-B intermediate, a glucosyl–carboxyl compound, reacts with phosphate (Y) to form glucosyl-phosphate ($B - Y$). The other product (X) is fructose.

In double displacement reactions, the covalent intermediate between enzyme and substrate apparently influences the reaction to proceed more rapidly. Because the enzyme is unaltered at the end of the reaction, it functions as a true catalyst, even though it is temporarily altered during the enzymatic process.

Although many enzymes form a covalent intermediate, the mechanism is not essential for catalysis; some reactions proceed as illustrated in Figure 2B. One substrate (Y) reacts directly with the second substrate ($X - B$), in a so-called single displacement reaction. The B moiety, which is transformed in the chemical reaction, is involved in only one reaction and does not form a bond with a group on the enzyme surface. In Figure 2B, the enzyme maltose phosphorylase directly affects the bonds of the substrates ($B - X$ and Y), which, in this case, are maltose (glucosylglucose) and phosphate, to form the products, glucose (X) and glucosylphosphate ($B - Y$).

Covalent intermediates between part of a substrate and an enzyme occur in many enzymatic reactions, and various amino acids—serine, cysteine, lysine, and glutamic acid—are involved.

The rate of enzymatic reactions. *The Michaelis–Menten hypothesis.* If the velocity of an enzymatic reaction is represented graphically as a function of the substrate concentration (S), the curve obtained in most cases is a hyperbola (see Figure 3). The mathematical expression of this curve, shown in the equation below, was developed in 1913 by two German biochemists, L. Michaelis and M.L. Menten. In the equation, V_M is the maximal velocity of the reaction, and K_M is called the Michaelis constant,

The Michaelis constant, K_M

$$\text{velocity} = \frac{V_M(S)}{K_M + (S)}.$$

The shape of the curve is a logical consequence of the active-site concept; *i.e.*, the curve flattens at the maximum velocity (V_M), which occurs when all the active sites of the enzyme are filled with substrate. The fact that the velocity approaches a maximum at high substrate concentrations provides support for the assumption that an intermediate enzyme–substrate complex forms. At the point of half the maximum velocity ($\frac{V_M}{2}$ in Figure 3), the substrate concentration in moles per litre (M) is equal to the Michaelis constant, which is a rough measure of the affinity of the substrate molecule for the surface of the enzyme. K_M values usually vary from about 10^{-8} to 10^{-2} M, and V_M usually varies from 10^5 to 10^9 molecules of product formed per molecule of enzyme per second. The value for V_M is referred to as the turnover number when expressed as moles of product formed per mole of enzyme per minute. The binding of molecules that inhibit or activate the protein surface usually results in similar types of curves.

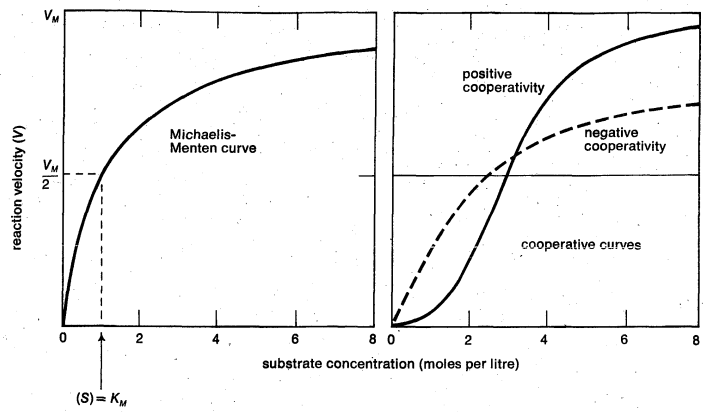
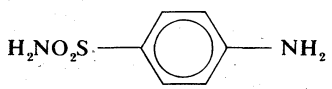


Figure 3: Curves representing enzyme action (see text).

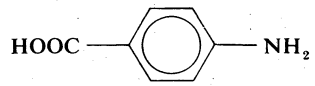
The efficiency of enzymes. Enzymes are more efficient than man-made catalysts operating under the same conditions. Because many enzymes with different specificities occur in a cell, adequate space exists only for a few enzyme molecules catalyzing one specific reaction. Each enzyme, therefore, must be very efficient. One molecule of the enzyme catalase, for example, can produce 10^{12} molecules of oxygen per second. The catalytic groups at the active site of an enzyme act 10^6 to 10^9 times more effectively than do analogous groups in a nonenzymatic reaction.

The reason for the great efficiency of enzymes is not completely understood. It results in part from the precise positioning of the substrates and the catalytic groups at the active site, which serves to increase the probability of collision between the reacting atoms. In addition, the environment at the active site may be favourable for reaction—that is, acidic and basic groups may act together more effectively there, or some strain may be induced in the substrate molecules so that their bonds are broken more easily, or the orientation of the reacting substrates may be optimal at the enzyme surface. The theories that have been formulated to account for the high catalytic efficiency of enzymes, although reasonable, still remain to be proved.

Inhibition of enzymes. Some molecules very similar to the substrate for an enzyme may be bound to the active site but be unable to react. Such molecules cover the active site and thus prevent the binding of the actual substrate to the site. This inhibition of enzyme action is of a competitive nature, because the inhibitor molecule actually competes with the substrate for the active site. The inhibitor sulfanilamide (see below), for example, is similar enough to a substrate (*p*-aminobenzoic acid) of an enzyme involved in the metabolism of folic acid that it binds to the enzyme but cannot react. It covers the active site and prevents the binding of *p*-aminobenzoic acid. This enzyme is essential in certain disease-causing bacteria but is not essential to man; large amounts of sulfanilamide therefore kill the micro-organism but do not harm man. Inhibitors such as sulfanilamide are called anti-metabolites. Sulfanilamide and similar compounds that kill a pathogen without harming its host are now widely used in chemotherapy.



sulfanilamide
(inhibitor)



p-aminobenzoic acid
(substrate)

Some inhibitors prevent, or block, enzymatic action by reacting with groups at the active site. The nerve gas di-

isopropyl fluorophosphate, for example, reacts with the serine at the active site of acetylcholinesterase to form a covalent bond. The nerve-gas molecule involved in bond formation prevents the active site from binding the substrate, acetylcholine, thereby blocking catalysis and nerve action. Iodoacetic acid similarly blocks a key enzyme in muscle action by forming a bulky group on the amino acid cysteine, which is found at the enzyme's active site. This process is called irreversible inhibition.

Some inhibitors modify amino acids other than those at the active site, resulting in loss of enzymatic activity. The inhibitor causes changes in the shape of the active site. Some amino acids other than those at the active site, however, can be modified without affecting the structure of the active site; in these cases, enzymatic action is not affected.

Such chemical changes parallel natural mutations. Inherited diseases frequently result from a change in an amino acid at the active site of an enzyme, thus making the enzyme defective. In some cases, an amino acid change alters the shape of the active site to the extent that it can no longer react; such diseases are usually fatal. In others, however, a partially defective enzyme is formed, and an individual may be very sick but able to live.

Effects of temperature and pH. Enzymes function most efficiently within a physiological temperature range. Since enzymes are protein molecules, they can be destroyed by high temperatures. An example of such destruction, called protein denaturation, is the curdling of milk when it is boiled. Increasing temperature has two effects on an enzyme. First, the velocity of the reaction increases somewhat, because the rate of chemical reactions tends to increase with temperature; second, the enzyme is increasingly denatured. Increasing temperature thus increases the metabolic rate only within a limited range. If the temperature becomes too high, enzyme denaturation destroys life. Low temperatures also change the shapes of enzymes. With enzymes that are cold-sensitive, the change causes loss of activity. Both excessive cold and heat are therefore damaging to enzymes.

Denaturation of enzymes

The degree of acidity or basicity of a solution, which is expressed as pH, also affects enzymes. As the acidity of a solution changes—*i.e.*, the pH is altered—a point of optimum acidity occurs, at which the enzyme acts most efficiently. Although this pH optimum varies with temperature and is influenced by other constituents of the solution containing the enzyme, it is a characteristic property of enzymes. Because enzymes are sensitive to changes in acidity, most living systems are highly buffered; *i.e.*, they have mechanisms that enable them to maintain a constant acidity. This acidity level, or pH, is about 7 in most organisms. Some bacteria function under moderately acidic or basic conditions; and the digestive enzyme pepsin acts in the acid milieu of the stomach. There is no known organism that can survive in either a very acidic or a very basic environment (see also HOMEOSTASIS).

ENZYME FLEXIBILITY AND ALLOSTERIC CONTROL

The induced-fit theory. The key–lock hypothesis (see above *The nature of enzyme-catalyzed reactions*) does not fully account for enzymatic action; *i.e.*, certain properties of enzymes cannot be accounted for by the simple relationship between enzyme and substrate proposed by the key–lock hypothesis. A theory called the induced-fit theory retains the key–lock idea of a fit of the substrate at the active site but postulates in addition that the substrate must do more than simply fit into the already preformed shape of an active site. Rather, the theory states, the binding of the substrate to the enzyme must cause a change in the shape of the enzyme that results in the proper alignment of the catalytic groups on its surface. This concept has been likened to the fit of a hand in a glove, the hand (substrate) inducing a change in the shape of the glove (enzyme). Although some enzymes appear to function according to the older key–lock hypothesis, most apparently function according to the induced-fit theory.

During step 1 in Figure 4, which illustrates the induced-fit theory, the substrate approaches the enzyme surface and induces a change in its shape that results in the correct alignment of the catalytic groups (indicated by triangles *A* and *B*). In the case of the digestive enzyme carboxypeptidase, the binding of the substrate causes a tyrosine molecule at the active site to move by as much as 15 angstroms (one angstrom equals 10^{-8} centimetres). Circles *C* and *D* in the figure represent substrate-binding groups on the enzyme that are essential for catalytic activity. During step 2 the catalytic groups at the active site react with the substrate to form products. The products separate from the enzyme surface during step 3, and the enzyme is able to repeat the sequence.

Nonsubstrate molecules that are too bulky (Figure 4D) or too small (Figure 4E) alter the shape of the enzyme so that a misalignment of catalytic groups *A* and *B* occurs; such molecules are not able to react even if they are attracted to the active site.

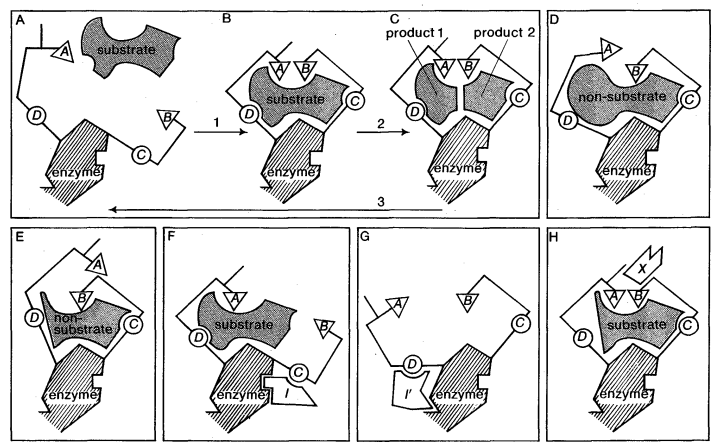


Figure 4: Induced-fit binding of a substrate to an enzyme surface and allosteric effects (see text).

The induced-fit theory explains a number of anomalous properties of enzymes; for example, "noncompetitive inhibition" (Figure 4F), in which a compound inhibits the reaction of an enzyme but does not prevent the binding of the substrate. In this case, the inhibitor compound *I* attracts the binding group *C* so that the catalytic group *B* is too far away from the substrate to react. The site at which the inhibitor binds to the enzyme is not the active site and is called an allosteric site. The inhibitor changes the shape of the active site to prevent catalysis without preventing binding of the substrate.

Figure 4G shows the effect of an inhibitor (*I*), which distorts the active site by affecting the essential binding group *D;* as a result, the enzyme can no longer attract the substrate. In Figure 4H, a so-called activator molecule, *X*, affects the active site so that a nonsubstrate molecule is properly aligned and hence can react with the enzyme; *X* is called an allosteric activator of the reaction. Such activators can affect both binding and catalytic groups at the active site.

Enzyme flexibility is extremely important because it provides a mechanism for regulating enzymatic activity. As shown in Figure 4F and G, the orientation at the active site can be disrupted by the binding of an inhibitor at a site other than the active site. Moreover, the enzyme can be activated by molecules that induce a proper alignment of the active site for a substrate that alone cannot induce this alignment (Figure 4H).

The importance of enzyme flexibility

As mentioned above, the sites that bind inhibitors and activators are called allosteric sites to distinguish them from active sites. Allosteric sites are in fact regulatory sites able to activate or inhibit enzymatic activity by influencing the shape of the enzyme. When the activator or inhibitor dissociates from the enzyme, it returns to its normal shape. Thus, the flexibility of the protein structure allows the operation of a simple, reversible control system similar to a thermostat.

Types of allosteric control. *Feedback inhibition and allosteric activation.* Allosteric control can operate in

many ways; two examples serve to illustrate some general effects. A pathway consisting of ten enzymes is involved in the synthesis of the amino acid histidine. When a cell contains enough histidine, synthesis stops—an appropriate economy move by the cell. Synthesis is stopped by the inhibition of the first enzyme in the pathway by the product, histidine; the mechanism is similar to that of Figure 4G. The inhibition of an enzyme by a product is called feedback inhibition; *i.e.*, a product many steps removed from an initial enzyme blocks its action. Feedback inhibition occurs in many pathways in all living things.

Allosteric control can also be achieved by activators. The hormone adrenaline (epinephrine) acts in this way. When energy is needed, adrenaline is released and activates, by allosteric activation, the enzyme adenyl cyclase. This enzyme catalyzes a reaction in which the compound cyclic adenosine monophosphate (cyclic AMP) is formed from ATP. Cyclic AMP in turn acts as an allosteric activator of enzymes that speed the metabolism of carbohydrate to produce energy. This type of allosteric regulation also is widespread in biological systems.

Thus, a combination of allosteric activation and inhibition allows the production of energy or materials when they are needed and shuts off production when the supply is adequate.

Induction and repression. Allosteric control is a rapid method of regulating products continuously needed by living things. Yet some cells have no need for certain enzymes, and it would be wasteful for the cell to synthesize them. In this case, certain molecules, called repressors, prevent the synthesis of unneeded enzymes. The repressors are proteins that bind to deoxyribonucleic acid (DNA) and prevent the first step in the process resulting in protein synthesis. If certain metabolites are added to cells that need an enzyme, enzyme synthesis occurs—*i.e.*, it is induced. Addition of galactose to a growth medium containing *Escherichia coli* bacteria, for example, induces the synthesis of the enzyme beta galactosidase. The bacteria thus can synthesize this galactose-metabolizing enzyme when it is needed and prevent its synthesis when it is not. The way in which the synthesis of enzymes is induced or repressed in mammalian systems is less well understood but is believed to be similar.

Different types of cells in complex organisms have different enzymes, even though they have the same DNA content. The enzymes actually synthesized are the ones needed in a specific cell and vary not only for different types of cells—*e.g.*, nerve, muscle, eye, and skin cells—but also for different species.

Cooperativity in allosteric enzymes. In an enzyme consisting of several subunits, or chains, alteration in the shape of one chain as a result of the influence either of a substrate molecule or of allosteric inhibitors or activators may change the shape of a neighbouring chain. As a result, the binding of a second molecule of substrate occurs in a different way from the binding of the first, and the third is different from the second. This phenomenon, called cooperativity, is characteristic of allosteric enzymes.

Cooperativity (see Figure 3) is reflected by a sigmoid curve, as compared to the hyperbolic curve of Michaelis–Menten. An enzyme of several subunits that exhibits cooperativity is far more sensitive to control mechanisms than is an enzyme of one subunit and hence one active site.

The first example of cooperativity was observed in hemoglobin, which is not an enzyme but behaves like one in many ways. The absorption of oxygen in the lungs and its deposition in the tissues is far more efficient because the subunits of hemoglobin show positive cooperativity, so-called because the first molecule of substrate makes it easier for the next to bind.

Negative cooperativity (also illustrated in Figure 3), in which the binding of one molecule makes it less easy for the next to bind, also occurs in living things. Negative cooperativity makes an enzyme less sensitive to fluctuations in concentrations of metabolites and may be important for enzymes that must be present in the cell at relatively constant levels of activity.

Some enzymes are closely associated aggregates of several enzyme units; the pyruvate dehydrogenase system, for example, contains five different enzymes, has a total molecular weight of 4,000,000, and consists of four different types of chains. Apparently, the enzymes in cells may be organized by forming complex units, by being absorbed on a cell wall, or by being isolated by membranes in special compartments. Since a pathway involves the stepwise modification of chemical compounds, aggregations of the enzymes in a given pathway facilitate their function in a manner similar to an industrial assembly line (see also METABOLISM).

BIBLIOGRAPHY. ALBERT L. LEHNINGER, *Biochemistry* (1970), a general biochemical text with emphasis on the biological aspects of enzymatic action; HENRY R. MAHLER and EUGENE H. CORDES, *Basic Biological Chemistry* (1968), a general biochemical text with detailed consideration of enzymatic pathways and with emphasis on the chemistry of enzymatic stages; S.G. WALEY, *Mechanisms of Organic and Enzymic Reactions* (1962), a short text emphasizing enzymatic mechanisms; ABRAHAM WHITE, PHILIP HANDLER, and EMIL L. SMITH, *Principles of Biochemistry*, 4th ed. (1968), a general biochemical text with emphasis on medical aspects of enzymatic function.

(D.E.K.)

Epaminondas

A Greek statesman and general who was largely responsible for breaking the military dominance of the Spartans, Epaminondas, by his innovations in the conventional scheme of Greek tactics, opened the way for Alexander the Great's military victories and, by his creation of new political centres in the Peloponnese, permanently altered the balance of power in Greece. His remarkable personality was well on the way to making his native city, Thebes, the leading city of 4th-century Hellas, when his death in battle cut short his progress.

Epaminondas was born in Thebes of an aristocratic family in the last years of the 5th century BC (about 410 BC). His father, though poor, provided his son with a good education. Particularly attracted to philosophy, the boy became a devoted pupil of Lysis of Tarentum, a Pythagorean, who had settled in Thebes. Epaminondas did not at first take any part in political life but served on military expeditions. There is a legend that he saved the life of his colleague Pelopidas in battle in 385.

In 382 the Spartans took advantage of an expedition to northern Greece to conspire with a few Thebans and seize power by a sudden coup. For three and a half years the government was in the hands of this small dictatorship, backed by a Spartan garrison in the Cadmeia (the citadel of Thebes). Many of the previous leaders, including Pelopidas, were driven into exile. Epaminondas remained in private life, but when Pelopidas, returning secretly from Athens, successfully overthrew the dictatorship in 379 and frightened the Spartan garrison into surrender, Epaminondas is said to have been one of those who led the popular uprising in Thebes. No individual part is attributed to him for the next eight years, during which Thebes, in alliance with Athens, successfully fought off Sparta and re-established its traditional leadership in a federation of the cities of Boeotia. In 371 the general war was ended at a peace conference, but Sparta and Athens combined to refuse recognition to the Theban federation by insisting that each city of Boeotia should be a separate party to the treaty, while Thebes claimed that its federation should be treated as a single unit. Epaminondas, who was boeotarch (one of the five magistrates of the federation), maintained this position, even when it led to the exclusion of Thebes from the peace treaty. The Spartans had an army stationed on Thebes's western frontier, waiting to follow up their diplomatic success by a crushing military attack. But in the Battle of Leuctra (371) Epaminondas was ready with a tactical innovation. Instead of the usual advances of heavily armed infantry drawn up in an equal number of ranks over the whole front, he massed his troops on the left wing to the unprecedented depth of 50 ranks against

Epaminondas' first victory

an overall Spartan depth of 12. The Spartans, who according to Greek convention had their best troops on the right wing, were overwhelmed by the force of the Theban advance. The novelty consisted in striking the enemy first at their strongest, instead of their weakest, point, with such crushing force that the attack was irresistible. The defeat of the Spartans inflicted such heavy losses on the very limited numbers of the Spartan soldiers that it seriously threatened the possibility of raising another Spartan army. The Boeotian federation had been saved, and after more than a year the Theban army, once more led by Epaminondas, proceeded to press home its victory. In the winter (a most unusual season for Greek warfare) of 370–369 they invaded the Peloponnese and penetrated the valley of the Eurotas (modern Evrótas). For the first time for at least two centuries an enemy army was in sight of Sparta. The subject population of Helots revolted, and Epaminondas re-created the state of Messenia, which had been enslaved by the Spartans for 300 years. He also encouraged the Arcadians, who had broken from Sparta's league, to found Megalopolis (Big City) as a federal capital. These new political creations served to keep Sparta in check so that it was never again a serious military power outside the Peloponnese. Epaminondas' brilliant success was met with jealousy and political opposition at home. He had stayed abroad over his year of office and was impeached on his return but acquitted. In 369–368 he led a second successful invasion of the Peloponnese, gaining further allies for Boeotia. In 367 he also served as a common soldier in an army sent to rescue his friend Pelopidas, who was a prisoner of Alexander, tyrant of Pherae (Thessaly). The expedition got into difficulties from which it was only rescued when Epaminondas was appointed general. This resulted in his re-election as boeotarch. He then returned to Thessaly and secured the release of Pelopidas. In 366 he invaded the Peloponnese for a third time, with a view to strengthening the Theban position there. He obtained assurances of fidelity from several states and, perhaps because of these assurances, decided not to overthrow the oligarchical governments that had been established by the Spartans. This was not accepted by the Theban government, which was in favour of overthrowing the oligarchs and establishing new democracies.

Third Peloponnesian expedition

Athens had supported Sparta and was at war with Thebes. In 364–363 Epaminondas made a bold attempt to challenge Athens' naval empire. With a new Boeotian fleet, he sailed to Byzantium, with the result that a number of cities in the Athenian Empire rebelled against their now-threatened masters. But the next year the outbreak of civil war in the Arcadian league brought Epaminondas once more to the head of a large allied army in the Peloponnese. He was met by Sparta, Athens, and their allies in the Battle of Mantineia (362). Epaminondas repeated on a large scale the tactics of Leuctra and was once more victorious but died of a wound on the field of battle. With his death all constructive initiative appeared to vanish from Theban policy.

BIBLIOGRAPHY. Of ancient sources, NEPOS, *Epaminondas* is vague and moralizing. PLUTARCH'S *Life of Epaminondas* is lost but can be indirectly approached through his *Life of Pelopidas* and PAUSANIAS. There is no important single work on the subject by a modern scholar, so it is best studied in the standard general histories. See especially M. CARY in the *Cambridge Ancient History*, vol. 6, ch. 4 (1927); and N.G.L. HAMMOND, *A History of Greece to 322 BC*, 2nd ed., ch. 2–3 and appendix 7 (1967), in which special attention is given to the chronological problems.

(H.W.P.)

Ephemeroptera

The order Ephemeroptera comprises the group of insects known as mayflies. Other common names for the winged stages are shadfly, sandfly, dayfly, fishfly, and drake. The aquatic immature stage, called a nymph or naiad, is widely distributed in freshwater; a few species can tolerate the brackish water of marine estuaries.

The winged stages attract attention through mass emergences when they may make roads slippery, clog gutters, and taint the air with an odour of decay. Mayfly nymphs are important in the energy transfer cycle that occurs in freshwaters (winged stages cannot feed). Some species are carnivorous, but the majority of nymphs feed on diatoms, algae, higher plants, and organic detritus; nymphs are devoured in turn by many carnivorous animals, especially fishes.

General features. *Appearance.* Winged mayflies have large compound eyes; short, bristlelike antennae; and functionless mouthparts and digestive tracts. Their membranous wings (usually with many veins) include a front pair and a hind pair, which is much smaller than the front pair or absent; in repose, the wings are held together upright over the body like those of a butterfly. The adult mayfly has two or three threadlike tails, usually as long as, or longer than, the body.

Nymphal characters

Nymphal characters include a single claw terminating each of the six legs. The surface of the thoracic region of the body is strongly rounded outward and bears the developing wings in external pads on the upper surface; the abdominal region usually is long and slender. Gills are attached to the outer edge of the upper surface of some of the ten segments into which the body is divided. The body of the nymph terminates in three, less often two, slender tails. Adult mayflies of North American species range in body length, exclusive of tails, from 2.5 millimetres, (0.1 inch) for *Caenis* to 32 millimetres (more than an inch) for *Hexagenia*.

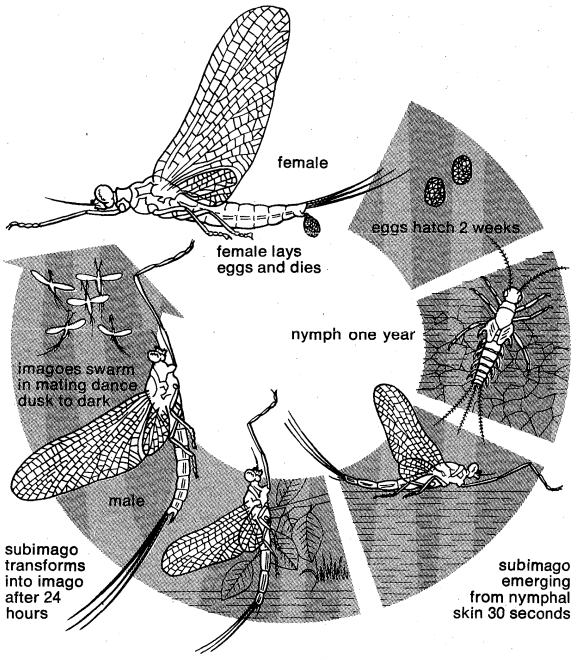
Distribution and abundance. Worldwide, about 2,000 species of mayflies have been described, nearly 600 of them from North America north of Mexico. The order, represented in all continents except Antarctica, occurs sparingly in many islands far distant from major landmasses; *e.g.*, only two species are known from Fiji. In areas of high biological productivity (*e.g.*, gravel-bottomed, hard-water, temperate-zone streams), as many as 1,400 nymphs have been found in one square foot of surface, and one gravel riffle (a shallow area extending across a stream bed) has yielded 33 species.

Natural history. *Life cycle.* The life cycle of mayflies consists of four stages: egg, nymph, subimago, and imago. Eggs, which vary widely in size and surface detail, may be oblong, oval, or rounded. Depending on the species, a female may produce fewer than 50 or more than 10,000 eggs. Eggs are laid in water and either settle to the bottom or adhere to some submerged object; they often hatch in about two weeks but may undergo a period of varying duration in which no growth occurs. This cessation of growth, known as diapause, is an adaptation that enables the insects to avoid conditions hostile to developing nymphs or to emerging winged stages.

Nymphal life may be as short as two weeks or as long as two years; an annual cycle is most common. As many as 50 molts (periodic shedding of skin) may occur, depending on the species and the environment. When growth is complete, the nymphal skin splits down the back and a winged form, called the subimago, or dun, emerges. The subimago flies from the surface of the water to some sheltered resting place nearby; after an interval lasting a few minutes to several days, but usually overnight, the skin is shed for the last time, and the imago, spinner, or adult stage emerges. Mayflies are the only insects that molt after developing functional wings. The subimago resembles the imago in overall appearance, although it is softer and duller in colour than the adult. The wings of the subimago, generally rather opaque, are tinted with gray, blue, yellow, or olive; heavy pigmentation along the veins may give the wings of the subimago a mottled appearance that rarely persists in the imago. Legs and tails of the subimago are shorter than are those of the imago. It is often incorrectly assumed that the two stages are different species.

Subimago and imago stages

Mating and egg deposition. Mating takes place soon after the final molt. In most species death ensues shortly after mating and oviposition (egg deposition). Winged existence may last only a few hours; *Hexagenia* males, however, may live long enough to engage in mating flights on two successive days, and female imagos that retain their eggs may live long enough to mate on either



Life cycle of the mayfly.

of two successive days. Groups of male imagoes perform a mating flight, or dance, over water as dusk approaches, flying into any breeze or air current. Individuals may fly up and forward, then float downward and repeat the performance. Females soon join the swarm, rising and falling as the dance continues. The male approaches the female from below and behind and grasps her thorax with his elongated front legs; mating is completed on the wing. After her release by the male, the female deposits her eggs and dies. A few species are ovoviviparous—*i.e.*, eggs hatch within the body of the female.

Methods of oviposition vary. Some species drop the rounded egg mass from a height of several feet in a manoeuvre suggestive of dive-bombing; in others, the female flies low over the water's surface, striking it at intervals with the tip of her abdomen and washing off a few eggs each time she strikes the water. Still other females extrude the eggs from two oviducts as two long packets, which usually adhere to each other; they may be dropped from a foot or more above the water, but more often, the female falls to the surface with wings extended and squeezes out the eggs as she dies. In a fourth type of oviposition, the female alights on some object protruding from the water and crawls under the surface, depositing the eggs while submerged. Females, unless they drop the eggs from a height of several feet, are vulnerable to feeding fishes. Mayflies sometimes mistake blacktopped roads for streams, forming swarms over them, and drop eggs on road surfaces.

Ecology. Mayfly nymphs are preyed upon by carnivorous invertebrates and fishes. Winged stages are devoured in flight by birds, bats, and predatory insects, including dragonflies, robber flies, and hornets. When at rest, mayflies may be preyed upon by spiders, beetles, birds, and certain mammals, especially flying squirrels in North America. During their transformation to the adult stage and especially during oviposition by females, mayflies are vulnerable to predation by fishes; artificial lures used by fishermen are patterned after them.

Form and function. Adaptations of form and function presumably determine distribution. The legs and jaws of some nymphs are modified for burrowing in silt or sand; other species are flattened, so that they can enter narrow crevices or cling to bottom materials in swift currents; long, slender legs and body adapt others for clambering on submerged vegetation. Strong swimmers are long and slender and occupy a variety of habitats. Gills may be platelike, feathery, or filamentous; often they are modified for specialized functions.

Paleontology and classification. Recognizable mayflies occur in the fossil record of the Upper Carboniferous Period (about 325,000,000 to 280,000,000 years ago), and they appear to have been abundant during the Permian. Represented largely by wing impressions, the fossil record is so incomplete that most systems of classification and interpretations of relationships are based on characteristics of recent forms, chiefly their morphology.

Distinguishing taxonomic features. Characteristics of the male genitalia are the most reliable means for identification of adult species. Many other features, including patterns of veins in the wings, affect generic and other higher categories of classification.

Annotated classification. The classification below is modified from that of George Edmunds, Jr. (1962).

ORDER EPHEMEROPTERA (mayflies)

Soft-bodied insects; life cycle consisting of 4 stages—egg, nymph, subimago, imago; wings membranous, at rest held vertically upward; hind wings reduced; mouthparts and digestive system of adults nonfunctional; only insect to molt after developing functional wings; antennae bristlelike; 5 superfamilies—Heptagenioidea (families Siphlonuridae, Baetidae, Metretopodidae, Oligoneuriidae, Heptageniidae, and Ametropodidae); Leptophlebioidea (families Leptophlebiidae, Ephemerellidae, Tricorythidae); Ephemeroidea (families Behningiidae, Potamanthidae, Euthyplociidae, Polymitarcidae, Ephemeridae, Palingeniidae); Caenoidea (families Neoephemeridae, Caenidae); Prosopistomatoidea (families Baetiscidae, Prosopistomatidae); about 2,000 species on all continents except Antarctica.

Critical appraisal. In addition to the scheme above, another classification increases the number of families to 25. Before the taxonomy is established with certainty, more attention must be given to nymphal characters, not only their morphology but also their physiology and behaviour. Many species remain undescribed.

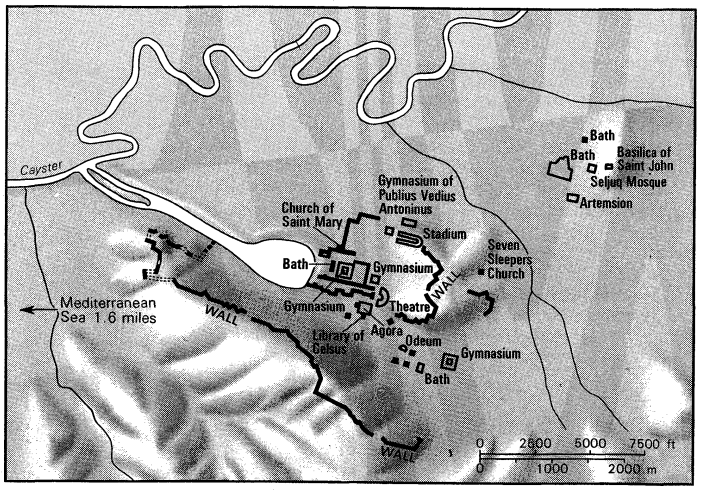
BIBLIOGRAPHY. L. BERNER, *A Tabular Summary of the Biology of North American Mayfly Nymphs (Ephemeroptera)* (1959); B.D. BURKS, *The Mayflies, or Ephemeroptera, of Illinois* (1953), chiefly taxonomic, with good illustrations; A.E. EATON, *A Revisional Monograph of the Ephemeridae, or Mayflies*, 6 pt. (1883–88), taxonomy of world animal life with historical value; G.F. EDMUNDS, JR., "The Principles Applied in Determining the Hierarchic Level of the Higher Categories of Ephemeroptera," *Syst. Zool.*, 11:22–31 (1962), and "Biogeography and Evolution of Ephemeroptera," *A. Rev. Ent.* (1972); J.W. and F.A. LEONARD, *Mayflies of Michigan Trout Streams* (1962), biology and ecology with colour plates; J.G. NEEDHAM, J.R. TRAVER, and Y.C. HSU, *The Biology of Mayflies* (1935, reprinted 1969), comprehensive handbook.

(J.W.L.)

Ephesus

Ephesus was an ancient Ionian Greek city, the ruins of which lie near the modern village of Seljuq in the Izmir *il* (province) of Turkey. In Roman times it was situated on the northern slopes of the hills Coressus and Pion and south of the Cayster (Küçükmenderes) River, the silt from which has formed a fertile plain but has caused the coastline to move ever farther west. The Temple of Artemis, or Diana, to which Ephesus owed much of its fame and which seems to mark the site of the classical Greek city, was probably on the seaboard when it was founded (about 600 BC), one mile east by northeast of Pion (modern Panayir Dağ). In Roman times a sea channel was maintained with difficulty to a harbour well west of Pion. In late Byzantine times this channel in turn became useless, and the coast by the mid-20th century was three miles farther west. Ephesus commanded the west end of one great trade route into Asia, that along the Cayster valley, and had easy access to the other two, along the Hermus (Gediz) and the Maeander (Büyükmenderes) Rivers.

History. Ephesus enters history in the mid-7th century BC, when it was attacked by the Cimmerians. Unlike its neighbour, Magnesia, it survived the attacks. For part of the early 6th century the city was under tyrants. Though allied by marriage to the kings of Lydia, its people could not hold back the Lydian Croesus, who asserted a general suzerainty over the city. He did, however, present many



Principal sites of Hellenistic and Byzantine Ephesus.
Adapted from J. Keil, *Ephesos* (1964); Austrian Archeological Institute, Vienna

columns and some golden cows for a new and splendid rebuilding of the Artemiseum (Temple of Artemis). At this time, according to Strabo, the Ephesians began to live in the plain; and to this period, too, should be allotted the redrafting of the laws, said to have been the work of an Athenian, Aristarchus. Ephesus soon submitted to Cyrus of Persia. Early in the Ionian revolt (499–493 BC) against the Persians, Ephesus served as a base for an Ionian attack on Sardis; but it is not mentioned again until 494, when the Ephesians massacred the Chiot survivors of the battle of Lade. The massacre may have occurred because Ephesus was a commercial rival of the chief rebels, Chios and Miletus. Ephesus maintained friendly relations with Persia for about 50 years: in 478 Xerxes, returning from his failure in Greece, honoured Artemis of Ephesus, although he sacked other Ionian shrines, and left his children for safety in Ephesus; and Themistocles landed there in the 460s on his flight to Persia. But after 454 Ephesus appears as a regular tributary of Athens. Great Ephesians up to this time had been Callinus, the earliest Greek elegist (mid-7th century BC), the satirist Hipponax, and the famous philosopher Heracleitus, one of the Basilids.

The Greco-Persian wars

Ephesus shared in a general revolt of 412 BC against Athens, siding with Sparta in the Second Peloponnesian war, and remained an effective ally of Sparta down to the end of the war. Threatened by Persia after 403, Ephesus served in 396 as the headquarters of King Agesilaus of Sparta. In 394 the Ephesians deserted to Conon's anti-Spartan maritime league, but by 387 the city was again in Spartan hands and was handed by Antalcidas to Persia. There followed the pro-Persian tyranny of Syrphax and his family, who were stoned to death in 333 on Alexander the Great's taking the city. After 50 years of fluctuating fortune, Ephesus was conquered by the Macedonian general Lysimachus, and resettled around Coressus and Pion (286–281 BC). Lysimachus introduced colonists from Lebedus and Colophon and renamed the city after his wife, Arsinoo—a name soon dropped. This was the beginning of Ephesus' Hellenistic prosperity. It became conspicuous for the abundance of its coinage.

After the defeat of Antiochus the Great, king of Syria, by the Romans in 189 BC, Ephesus was handed over by the conquerors to the king of Pergamum. Attalus III of Pergamum bequeathed Ephesus with the rest of his possessions to the Roman people (133 BC). Thenceforth, Ephesus remained subject to Rome, except for a brief time beginning in 88 BC, when, at the instigation of Mithradates the Great of Pontus, the cities of Asia Minor revolted and killed their Roman residents. The Ephesians even killed those Romans who had fled for refuge to the Artemiseum; notwithstanding which they returned in 86 BC to their former masters. Their claim, preserved on an extant inscription, that in admitting Mithradates they had merely yielded to superior force was rudely brushed aside by Sulla, who inflicted a very heavy fine. Although it twice chose the losing side in the Roman civil wars and although it was stoutly opposed by Pergamum and Smyrna, Ephesus became under Augustus the first city of the Roman province of Asia. The geographer Strabo wrote of its importance as a commercial centre in the 1st century BC. The triumphal arch of 3 BC and the aqueduct of AD 4–14 initiated that long series of public buildings, ornamental and useful, that make Ephesus the most impressive example in Greek lands of a city of imperial times.

Meanwhile the Christian Church began to win converts. A famous protest in the theatre against the teachings of St. Paul, described in Acts 19, is dated about AD 57. According to local belief Ephesus was the last home of the Virgin, who was lodged near the city by St. John and died there. The tradition that St. Luke also died there seems to be less strongly supported. Ephesus was one of the seven churches of Asia to which the Revelation to John or Revelation was addressed.

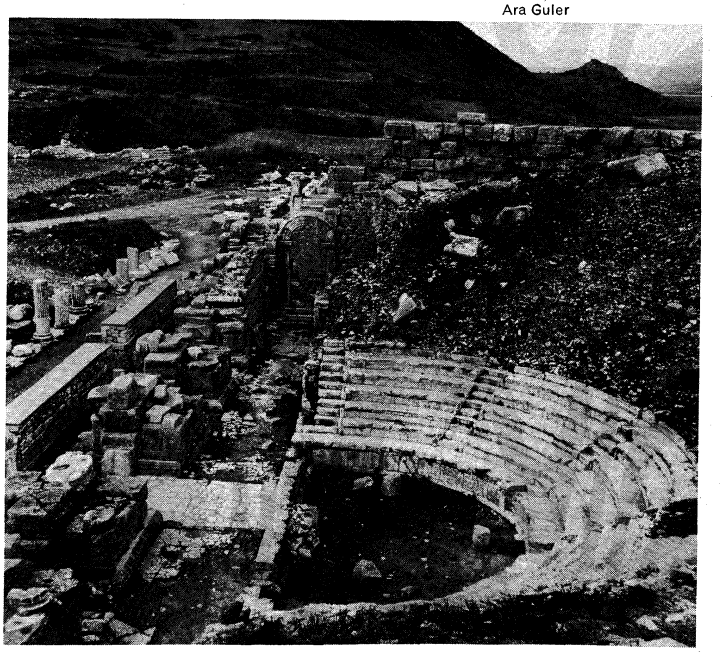
The Goths destroyed both city and temple in AD 262, and neither ever recovered its former splendour. The emperor Constantine, however, erected a new public bath, and Arcadius rebuilt at a higher level the street from the theatre to the harbour, named after him, the Arkadiane. A general council of the church, held at Ephesus in AD 431 in the great double church of St. Mary, condemned Nestorius and justified the cult of the Virgin as Theotokos, "mother of God." A few years later according to legend, the Seven Sleepers of Ephesus (a group of 3rd-century Christian martyrs) were miraculously raised from the dead. They too became the object of a famous cult. The emperor Justinian built the magnificent basilica of St. John in the 6th century. By the early Middle Ages, the city was no longer useful as a port, and fell into decline; late Byzantine Ephesus, conquered by the Seljuqs in 1090, was merely a small town. After brief splendour in the 14th century, even this was deserted, and the true site of the Artemiseum remained unsuspected until 1869.

The Council of Ephesus

Excavations and extant remains. J.T. Wood, working at Ephesus for the British Museum between 1863 and 1874, excavated the odeum and theatre. In May 1869 he struck a corner of the Artemiseum. His excavation exposed to view not only the scanty remains of the latest edifice (built after 350 BC) but the platform below it of an earlier temple of identical size and plan subsequently found to be that of the 6th century BC, to which Croesus contributed. The sculptured fragments of both temples

The Artemiseum

Ara Güler



Ruins of the odeum, Ephesus, built in the 2nd century AD.

were sent to the British Museum. In 1904 D.G. Hogarth, heading another mission from the museum, examined the earlier platform and found beneath its centre the remains of three yet older structures. In its earliest known phase the temple was apparently a small platform of green schist, containing a sealed deposit of primitive coins and other objects. These date from *c.* 600 BC.

It is impossible to assign the various architects named by ancient authors to the respective phases of the temple. At best, Chersiphron and Metagenes can be tentatively assigned to the Temple of Croesus, Chirocrates or Dinocrates to that of the 4th century. There had perhaps been some repairs toward 400 BC, associated with the architects Paeonius and Demetrius and with the prize-winning dedicatory hymn of the famous musician Timotheus. The Artemiseum passed rapidly through three phases before *c.* 550 BC. The temple of Croesus (the fourth phase) was remarkable for its great size (it was more than 300 feet long and 150 feet wide), for the carved figures round the lower drums of its columns (*columnae caelatae*), and for the smaller but elaborate figured friezes along its roof gutter (*sima*). Croesus' temple seems to have been burned down in 356 BC. The new temple built shortly afterward copied the old in its *columnae caelatae*, one of which was by Scopas; but the new *sima*, instead of small, crowded figures, had a more conventional, if vigorous, *rinceau* ornament. The cella contained, among other great works, the Amazons of Polyclitus, Phidias, and Cresilas.

Other ruins

Lysimachean Ephesus has been continuously excavated since 1894 by the Austrian Archaeological institute, but so solid and extensive is the Roman town that by the early 1960s the Austrians had rarely penetrated to Hellenistic levels.

On the hill of Ayasoluk (Hagios Theologos) is Justinian's church of St. John the Theologian, built around a shrine variously associated in the early Middle Ages with the death or bodily assumption of St. John. The church, uncovered since 1922, is a noble structure but badly restored. On the hill there is also a beautiful Seljuq mosque dedicated in 1375.

The public buildings of the city are arranged in a rectangular street pattern going back to Hellenistic days. They include the theatre, capable of seating nearly 25,000 spectators and completed in its present form under Trajan; the agora (market place), surrounded by stoas (sheltered promenades) dating from the time of Severus; the library of Celsus, also Trajanic and well known because of its façade; and an immense array of baths and gymnasiums.

All these buildings are to the west of Pion. On its north side is the stadium, and north of this the gymnasium of Publius Vedius Antoninus, relatively small but very complete and with a notable chapel for the cult of Antoninus Pius. South of Pion were the odeum—another gift of Vedius—a roofed semicircular theatre to hold 1,400 persons; also a series of fountains and aqueducts, notably the aqueduct of Gaius Sextilius Pollio, which crossed the valley from Coressus. The unmortared city wall along the crest of Coressus appears to be that of Lysimachus.

Of the early Byzantine city, besides the stretch of curtain wall on Panajir Dağ, there remain the ruined church of the Seven Sleepers to its east and the long double basilica of the Virgin, the scene of the council, to its west. This basilica was rebuilt several times; it was largely around this building, between the great gymnasiums and the stadium of the classical city, that the early Byzantine Ephesians gathered.

BIBLIOGRAPHY. G.E. BEAN, *Aegean Turkey: An Archaeological Guide* (1966), contains useful material on topography and history. KARL GSCHWIND, *Der ephesische Johannes und die Artemis Ephesia* (1965); and FRANZ MILTNER, *Ephesos, Stadt der Artemis und des Johannes* (1958), treat both St. John's city and Artemis. See also B.F. DEUTSCH, *Our Lady of Ephesus* (1965); and JOSEPH KEIL, *Ephesos: Ein Führer durch die Ruinenstätte und ihre Geschichte*, 5th ed. rev. (1964), a guide to the old city. Artemis is treated more specifically by B.L. TRELL in *The Temple of Artemis at Ephesus* (1945); and by HELLA KRAUSE-ZIMMER in *Artemis Ephesia* (1964), supplemented by HERMAN THIERSCH, *Artemis Ephesia: Eine archäologische Untersuchung* (1935); EMANUEL LOEWY, *Zur Chronologie der frühgriechischen Kunst: Die Artemistempel von Ephesos* (1932); W.R. LETHABY, "The Earlier Temple of Artemis at Ephesus," *Journal of Hellenic Studies*, vol. 37 (1917); and JAMES FERGUSSON, *The Temple of Diana at Ephesus* (1883). Other *Journal of Hellenic Studies* articles include PAUL JACOBSTHAL, "The Date of the Ephesian Foundation—Deposit," and E.S.G. ROBINSON, "The Coins from the Ephesian Artemision Reconsidered" (both vol. 71, 1951). Older material includes T.A. JOHNSTON, *Ephesus and Its Great Council* (1930), on religion; RUDOLPH HEBERDEY, *Das Theater in Ephesos* (1912), dealing with the theatre; GEORGE PHAEDROS, *Guide Book of the Ruins of Ephesus* (1910); and J.T. WOOD, *Modern Discoveries on the Site of Ancient Ephesus* (1890).

Epic

An ambiguous term, "epic" is used most often to designate a long narrative poem recounting heroic deeds, though it has also been loosely used to describe novels, such as Tolstoy's *War and Peace*, and motion pictures, such as Eisenstein's *Ivan the Terrible*. In literary usage, the term encompasses both oral and written compositions. The prime examples of the oral epic are Homer's *Iliad* and *Odyssey*. Outstanding examples of the written epic include Virgil's *Aeneid* and Lucan's *Pharsalia* in Latin; *Chanson de Roland* in medieval French; Ariosto's *Orlando Furioso* and Tasso's *Gerusalemme liberata* in Italian; *Poema* (or *Cantar*) *de mio Cid* in Spanish; and Milton's *Paradise Lost* and Spenser's *Faerie Queene* in English. There are also seriocomic epics, such as the *Morgante* of a 15th-century Italian poet, Pulci, and the pseudo-Homeric *Battle of the Frogs and Mice*. Another distinct group is made up of the so-called beast epics—narrative poems written in Latin in the Middle Ages and dealing with the struggle between a cunning fox and a cruel and stupid wolf. Underlying all of the written forms is some trace of an oral character, partly because of the monumental persuasiveness of Homer's example but more largely because the epic was, in fact, born of an oral tradition. It is on the oral tradition of the epic form that this article will focus. The article SAGA deals with a related form. Discussions of the role of the epic in the development of the various literatures of the world may be found in such articles as LITERATURE, WESTERN; and LITERATURE, EAST ASIAN; and in articles on the arts of various peoples, such as AFRICAN PEOPLES, ARTS OF; and SOUTH ASIAN PEOPLES, ARTS OF.

GENERAL CHARACTERISTICS OF THE EPIC

An epic may deal with such various subjects as myths, heroic legends, histories, edifying religious tales, animal stories, or philosophical or moral theories. Epic poetry has been used by peoples all over the world and in different ages to transmit their traditions from one generation to another, without the aid of writing. These traditions frequently consist of legendary narratives about the glorious deeds of their national heroes. Thus scholars have often identified "epic" with a certain kind of heroic oral poetry, which comes into existence in so-called heroic ages. Such ages have been experienced by many nations, usually at a stage of development in which they have had to struggle for a national identity. This effort, combined with such other conditions as an adequate material culture and a sufficiently productive economy, tend to produce a society dominated by a powerful and warlike nobility, constantly occupied with martial activities, whose individual members seek, above all, everlasting fame for themselves and for their lineages.

The epic as a genre of heroic ages

Uses of the epic. The main function of poetry in heroic-age society, appears to be to stir the spirit of the warriors to heroic actions by praising their exploits and those of their illustrious ancestors, by assuring a long and glorious recollection of their fame, and by supplying them with models of ideal heroic behaviour. One of the favorite pastimes of the nobility in heroic ages in different times and places has been to gather in banquet halls to hear heroic songs, in praise of famous deeds, sung by professional singers as well as by the warriors

themselves. Heroic songs also were often sung before a battle, and such recitations had tremendous effect on the morale of the combatants. Among the Negro tribe of the Fulbe in The Sudan, for instance, whose epic poetry has been recorded, a nobleman customarily set out in quest of adventures accompanied by a singer (*mabo*), who also served as his shield bearer. The singer was thus the witness of the heroic deeds of his lord, which he celebrated in an epic poem called *baudi*.

The aristocratic warriors of the heroic ages were, thus, members of an illustrious family, a link in a long chain of glorious heroes. And the chain could snap if the warrior failed to preserve the honour of the family, whereas, by earning fame through his own heroism, he could give it new lustre. Epic traditions were to a large extent the traditions of the aristocratic families: the Old French word *geste*, used for a form of epic that flourished in the Middle Ages, means not only a story of famous deeds, but also a genealogy.

The passing of a heroic age does not necessarily mean the end of its heroic oral poetry. An oral-epic tradition usually continues for as long as the nation remains largely illiterate. Usually it is after the heroic age has passed that the narratives about its legendary heroes are fully elaborated. Even when the nobility that originally created the heroic epic perishes or loses interest, the old songs can persist as entertainments among the people. Court singers, then, are replaced by popular singers, who recite at public gatherings. This popular tradition, however, must be distinguished from a tradition that still forms an integral part of the culture of a nobility. For when a heroic epic loses its contact with the banquet halls of the princes and noblemen, it cannot preserve for long its power of renewal. Soon it enters what has been called the reproductive stage in the life cycle of an oral tradition, in which the bards become noncreative reproducers of songs learned from older singers. Popular oral singers, like the *guslari* of Yugoslavia, no doubt vary their songs to a certain extent each time they recite them, but they do so mainly by transposing language and minor episodes from one acquired song to another. Such variations must not be confounded with the real enrichment of the tradition by succeeding generations of genuine oral poets of the creative stage. The spread of literacy, which has a disastrous effect on the oral singer, brings about a quick corruption of the tradition. At this degenerate stage, the oral epic soon dies out if it is not written down or recorded.

The popular tradition of the epic

The ancient Greek epic exemplifies the cycle of an oral tradition. Originating in the late Mycenaean period, the Greek epic outlasted the downfall of the typically heroic age culture (*c.* 1100 BC), and maintained itself through the "Dark Age" to reach a climax in the Homeric poems by the close of the Geometric Period (900–750 BC). After Homer, the activity of the *aoidoi*, who sang their own epic songs at the courts of the nobility, slowly declined. During the first half of the 7th century, the *aoidoi* produced such new poems as those of Hesiod and some of the earlier poems of what was to become known as the Epic Cycle. Between 625 and 575 BC, the *aoidoi* gave way to oral reciters of a new type called rhapsodes or "stitchers of songs," who declaimed for large audiences the already famous works of Homer, while holding in their hand, a staff (*rhabdos*), which they used to give emphasis to their words. It seems probable that these rhapsodes, who played a crucial role in the transmission of the Homeric epic, were using some sort of written aids to memory before Homeric recitations were adopted in 6th-century Athens as part of the Panathenaic festivals held each year in honour of the goddess Athena.

Verbal formulas. To compose and to memorize long narrative poems like the *Iliad* and the *Odyssey*, oral poets used a highly elaborate technical language with a large store of traditional verbal formulas, which could describe recurring ideas and situations in ways that suited the requirements of metre. So long as an oral epic tradition remains in its creative period, its language will be continually refined by each generation of poets in opposite directions, called scope and economy. Scope is the addition of new phrases to express a larger number of recurrent concepts in varying metrical values fitting the possible positions in a verse. Economy is the elimination of redundancies that arise as gifted poets invent new set phrases that duplicate, both in a general sense and in metrical value, the formulas already existing in the traditional stock.

Nowhere has this refinement proceeded any nearer to perfection than in the language of the Homeric epic. As has been shown by statistical analysis, it exhibits a remarkable efficiency, both in the rareness of unnecessarily duplicative variants and in the coverage of each common concept by the metrical alternatives useful in the composition of the six-foot metric line the Greeks used for epic poetry.

Thus, for example, if the idea of a ship has to be expressed at the end of a line of verse, the ship may be described as "well-trimmed" (*nēos eisēs*), "curved" (*nēos amphielissa*), or "dark-prowed" (*nēos kyanoprōiros*), depending entirely on the number of feet that remain to be filled by the phrase in the hexameter; if the phrase has to cover the two final feet of the verse and the words have to be put in the dative case, the formula "of a well-trimmed ship" will be replaced by "to a black ship" (*nēi melainē*). The sole occurrence of "Zeus who gathers lightning" (*steropēgereta Zeus*), which is an exact metrical equivalent of the more common "Zeus who delights in thunder" (*Zeus terpikeraunos*), constitutes one of the very few actual duplications of such formulas found in Homer.

Finally, some of the typical scenes in the heroic life, such as the preparation of a meal or sacrifice, or the launching or beaching of a ship, contain set descriptions comprising several lines that are used by rote each time the events are narrated.

This highly formalized language was elaborated by generations of oral poets to minimize the conscious effort needed to compose new poems and memorize existing ones. Because of it, an exceptionally gifted *aoidos*, working just prior to the corruption of the genre, could orally create long and finely structured poems like the *Iliad* and the *Odyssey* and those poems could then be transmitted accurately by the following generations of rhapsodes until complete, written texts were produced.

BASES OF THE EPIC

Oral heroic poetry, at its origin, usually deals with outstanding deeds of kings and warriors who lived in the heroic age of the nation. Since the primary function of this poetry is to educate rather than to record, however, the personages are necessarily transformed into ideal heroes and their acts into ideal heroic deeds that conform to mythological or ideological patterns. Some of these patterns are archetypes found all over the world, while others are peculiar to a specific nation or culture. Thus, in many epic traditions, heroes are born as a result of an illegitimate union of a maiden mother with a divine or supernatural being; they are exposed at birth, fed by an animal, and brought up by humble foster parents in a rustic milieu; they grow up with marvellous speed, fight a dragon—in their first combat—to rescue a maiden whom they marry, and die young in circumstances as fabulous as those that surrounded their birth.

Heroic patterns and deeds

In the traditions of Indo-European peoples a hero is often a twin, who acquires soon after his supernatural birth an invulnerability that has one defect, generally of his heel or of some other part of his foot, which ultimately causes his death. He is educated by a blacksmith, disguises himself as a woman at some time in his youth, and conquers a three-headed dragon, or some other kind of triple opponent, in his first battle. He then begets, by a foreign or supernatural woman, a child who, reared by his mother in her country, becomes a warrior as brave as his father. When this child meets his unknown father, the latter fails to recognize him, so that the father kills his own child after a long and fierce single combat. The hero, himself, usually dies after committing the third of three sins.

In Japan, to take another example, renowned members

of the warrior aristocracy of the past, who have acquired the status of popular heroes, are in many cases supplied in their legend with four exceptionally brave and faithful retainers called their *shitennō* or guardians of the four cardinal points; these form the closest entourage of their lord—who is usually depicted as excelling in command but not in physical strength—and defend him from dangers. The retainers reflect a mythological model, taken from Buddhism, of four *deva* kings, who guard the teaching of the Buddha against the attack of the devils.

A striking pattern for a number of epic traditions has been found in a so-called "tripartite ideology" or "trifunctional system" of the Indo-Europeans. The concept was based on the discovery of the remarkable philosophy of a prehistoric nation that survived as a system of thought in the historic Indo-European civilizations and even in the subconsciousness of the modern speakers of Indo-European tongues.

This philosophy sees in the universe three basic principles that are realized by three categories of people: priests, warriors, and producers of riches. In conformity with this philosophy, most Indo-European epics have as their central themes interaction among these three principles or functions which are: (1) religion and kingship; (2) physical strength; (3) fecundity, health, riches, beauty, and so forth. In the long Indian epic the *Mahābhārata*, for example, the central figures, the Pāṇḍava brothers, together with their father Paṇḍu, their two uncles Dhṛtarāṣṭra and Vidura, and their common wife, Draupadī, correspond to traditional deities presiding over the three functions of the Indo-European ideology.

Central themes of Indo-European epics

During the first part of their earthly career, the Pāṇḍava suffer constantly from the persistent enmity and jealousy of their cousins, Duryodhana and his 99 brothers, who, in reality, are incarnations of the demons Kālī and the Paulastya. The demons at first succeeded in snatching the Kingdom from the Pāṇḍava and in exiling them. The conflict ends in a devastating war, in which all the renowned heroes of the time take part. The Pāṇḍava survive the massacre, and establish on earth a peaceful and prosperous reign, in which Dhṛtarāṣṭra and Vidura also participate.

This whole story, it has been shown, is a transposition to the heroic level of an Indo-European myth about the incessant struggle between the gods and the demons since the beginning of the world. Eventually, it results in a bloody eschatological battle, in which the gods and the devils exterminate each other. The destruction of the former world order, however, prepares for a new and better world, exempt from evil influences, over which reign a few divine survivors of the catastrophe.

EARLY PATTERNS OF DEVELOPMENT

Development in the ancient Middle East. The earliest-known epic poetry is that of the Sumerians. Its origin has been traced to a preliterate Heroic Age, not later than 3000 BC, when the Sumerians had to fight, under the direction of a warlike aristocracy, for possession of this fertile Mesopotamian land. Among the extant literature of this highly gifted people are fragments of narrative poems recounting the heroic deeds of their early kings: Enmerkar, Lugalbanda, and Gilgamesh. By far the most important in the development of Mesopotamian literature are the five poems of the Gilgamesh cycle. This epic tells the odyssey of a king, Gilgamesh, part human and part divine, who seeks immortality. A god who dislikes his rule, fashions a wild man, Enkidu, to challenge him. Enkidu first lives among wild animals, then goes to the capital and engages in a trial of strength with Gilgamesh, who emerges victorious. The two, now friends, set out on various adventures, in one of which they kill a wild bull that the goddess of love had sent to destroy Gilgamesh because he spurned her marriage proposal. Enkidu dreams the gods have decided he must die for the death of the bull, and, upon awakening, he does fall ill and die. Gilgamesh searches for a survivor of the Babylonian flood to learn how to escape death. The survivor shows him where to find a plant that renews youth, but after Gilgamesh gets the plant it is snatched away by a serpent. Gilgamesh returns, saddened, to his capital.

The legend of Gilgamesh

The legend of Gilgamesh was taken over by the Babylonians, who developed it into a long and beautiful poem, one of the masterpieces of mankind.

Another Babylonian epic, composed around 2000 BC, is called in Akkadian *Enuma elish*, after its opening words, meaning "When on high." Its subject is not heroic but mythological. It recounts events from the beginning of the world to the establishment of the power of Marduk, the great god of Babylon. The outline of a Babylonian poem narrating the adventure of a hero named Adapa ("Man") can be reconstructed from four fragmentary accounts. It shares with the *Epic of Gilgamesh* the theme of man's loss of an opportunity for immortality.

Among clay tablets of the 14th century BC, covered with inscriptions in an old Phoenician cuneiform alphabet, from Ras Shamra (the site of ancient Ugarit), in northern Syria, there are important fragments of three narrative poems. One of these is mythological and recounts the career of the god Baal, which seems to coincide with the yearly cycle of vegetation on earth. As was usual with the death of gods in the ancient Mediterranean world, Baal's end brings about a drought that ceases only with his resurrection. Another fragment, about a hero named Aqhat, is perhaps a transposition of this myth of Baal to the human level. Just as the death of Baal is avenged on his slayer by Baal's sister Anath, so is the murder of Aqhat, which also causes a drought, revenged by his sister Paghat. Since the end of the poem is missing, however, it is not known whether Paghat, like Anath, succeeded in bringing her brother back to life.

The third fragment, the Ugaritic epic of Keret, has been interpreted as a Phoenician version of the Indo-European theme of the siege of an enemy city for the recovery of an abducted woman. This theme is also the subject of the Greek legend of the Trojan War and of the Indian epic *Rāmāyaṇa*. The fragmentary text does not reveal, however, whether the expedition of Keret, like that of the Achaean army against Troy, was meant to regain the hero's wife or to acquire for him a new bride.

The Greek epic. In its originative stage, especially, the Greek epic may have been strongly influenced by these Oriental traditions. The Greek world in the late Bronze Age was related to the Near East by so many close ties that it formed an integral part of the Levant. At Ugarit a large quarter of the city was occupied by Greek merchants, whose presence is also attested, among other places, at the gate of Mesopotamia, at Alalakh, in what is now Turkey. Thus, it is no surprise that, for example, the Greek myth about the succession of the divine kingship told in the *Theogony* of Hesiod and elsewhere is paralleled in a Hittite version of a Hurrian myth. In it, Anu, Kumarbi, and the storm god respectively, parallel Uranus, Cronos, and Zeus in the *Theogony*. The Hittites had continuous diplomatic relations with the Achaeans of Greece, whose princes went to the royal court at Hattusas to perfect their skill with the chariot. The Greeks, therefore, had ample opportunity to become familiar with Hittite myths.

Eastern influences on the Greek epic

The *Epic of Gilgamesh* was then well-known in the Levant, as is indicated by discoveries of copies of it throughout this wide area. Many parallels with the *Epic of Gilgamesh* have been pointed out in the *Odyssey*; the encounters of Odysseus with Circe and Calypso on their mythical isles, for instance, closely resemble the visit by Gilgamesh to a divine woman named Siduri, who keeps an inn in a marvellous garden of the sun god near the shores of ocean. Like the two Greek goddesses, Siduri tries to dissuade Gilgamesh from the pursuit of his journey by representing the pleasures of life, but the firm resolution of the hero obliges her finally to help him cross the waters of death. In the *Iliad*, Patroclus, who dies as a substitute for his king and dearest friend, Achilles, and then gives Achilles a description of the miserable condition of man after his death, bears striking similarities to the friend of Gilgamesh, Enkidu.

If these are indeed borrowings, it is all the more re-

markable that they are used in Homer to express a view of life and a heroic temper radically different from those of the Sumerian epic of Mesopotamia. Gilgamesh persists in his quest of immortality even when Siduri shows him the vanity of such an ambition, but Odysseus shuns a goddess's offer of everlasting life, preferring to bear his human condition to the end. The loss of a beloved friend does not make Achilles seek desperately to escape from death; instead he rushes into combat to revenge Patroclus, although he knows that he is condemning himself to an early death, and that the existence of a king in Hades will be incomparably less enviable than that of a slave on earth. The Mesopotamian mind never tires of expressing man's deep regret at not being immortal through stories about ancient heroes who, despite their superhuman strength and wisdom, and their intimacy with gods, failed to escape from death. A decisively different idea, however, is fundamental to the Greek heroic view of life. It has been demonstrated that the Greek view is derived from an Indo-European notion of justice —that each being has a fate (*moira*) assigned to him and marked clearly by boundaries that should never be crossed. Man's energy and courage should, accordingly, be spent not in exceeding the proper limits of his human condition but in bearing it with style, pride, and dignity, gaining as much fame as he can within the boundaries of his *moira*. If he is induced by Folly (Ate, personified as a goddess of mischief) to commit an excess (*hybris*) with regard to his *moira*, he will be punished without fail by the divine vengeance personified as Nemesis.

The Greek heroic view of life

At the beginning of the *Iliad*, a plague decimates the Achaean army because its commander in chief, Agamemnon, refuses to return a captive, Chryseis, to her father, a priest of Apollo who offers a generous ransom. By unjustly insulting Achilles, Agamemnon commits another excess that causes the defeat of his army. Achilles, in the meantime, lets Ate take possession of his mind and refuses, to the point of excess, to resume his fight. He thus brings about a great misfortune, the loss of his dearest companion, Patroclus. Patroclus, however, also contributes to his own death by his *hybris* in pushing his triumph too far, ignoring Achilles' order to come back as soon as he has repulsed the enemy far from the Greek ships. The death of Hector also results from his *hybris* in rejecting the counsel of Polydamas and maintaining his army on the plain after the return of Achilles to combat. After so many disasters caused by the mischievous action of Ate among men, the last book of the *Iliad* presents a noble picture of Priam and Achilles, who submit piously to the orders of Zeus, enduring with admirable courage and moderation their respective fates.

On the other hand, at the beginning of the *Odyssey*, Zeus evokes the ruin that Aegisthus will have to suffer for having acted "beyond his due share" by marrying Clytemnestra and murdering Agamemnon. This sets an antithesis to the story of the wise Odysseus, who, to accomplish his destiny as a mortal hero, never changes his purpose trying always to make the best of his countless misfortunes. He earns by this the favour of Athena and succeeds eventually in regaining Ithaca and punishing the wooers of Penelope for their *hybris* during his long absence. Present scholarship inclines to the view that such admirably well-structured poems as the *Iliad* and the *Odyssey* could have been created only by a single highly gifted poet whose name was Homer. This position contrasts with the extreme skepticism that marked all phases of Homeric criticism during the previous century. Yet the personality of Homer remains unknown and nothing certain is known about his life.

Hesiod's contribution

In comparison, information derived from his own works is fairly plentiful about the other great epic poet of Greece, Hesiod. He produced them presumably around 700 BC, while tilling a farm in Askra, a small village of Boeotia. The social and geographical background of his poems, called didactic because of their occasionally moral and instructive tone, differs from the aristocratic society of Ionian Asia Minor that Homer addressed. Despite their different style, subjects, and view of life, however, Hesiod's *Theogony* and the *Works and Days* illustrate the same basic conception of justice as the Homeric epic. The *Theogony* describes a long sequence of primordial events that resulted in the present world order, in which man's inescapable lot is assigned to him by Zeus. The *Works and Days* explains, through a series of three myths, why the lot of man is to work hard to produce riches. Man has to shut his ears to the goddess who causes wars and lawsuits, listening only to the goddess who urges him to toil more laboriously than his neighbour to become richer. Pain and suffering have become unavoidable since Pandora opened the fatal jar containing all the ills of mankind at Prometheus' house in conformity with the will of Zeus. Moreover, the age of the race of iron has arrived when the fate of human beings is not to pass their lives in perpetual banquets or warfare, as did the preceding races, but to suffer constantly the fatigue and misery of labour. As long as the goddesses Aidos (a personification of the sense of shame) and Nemesis (a personification of divine retribution) stay with mankind, however, helping people observe their *moira* without committing excesses, man can still gain riches, merits, and glory by the sweat of his brow. Only if he knows how to avoid all faults in doing his daily work will he not offend Justice (Dike), the sensitive virgin daughter of Zeus. This is why it is so vitally important for a farmer to know all the rules listed in the rest of the poem about seemingly trivial details of his work.

LATER VARIATIONS

The Latin epic. Latin epic poetry was initiated in the 3rd century BC by Livius Andronicus, who translated the *Odyssey* into the traditional metre of Saturnian verse. It was not until the 1st century BC, however, that Rome possessed a truly national epic in the unfinished *Aeneid* of Virgil (70–19 BC), who used Homer as his model. The story of Aeneas' journey, recounted in the first six books, is patterned after the *Odyssey*, with many imitative passages and even direct translations, while the description of the war in the last six books abounds with incidents modelled after those from the *Iliad*. More basically, however, Virgil made use of another model, Rome's own national legend about the war fought under Romulus against the Sabines. This legend preserves, in a historical disguise, an original Indo-European myth about a primitive conflict between the gods of sovereignty and war and the gods of fecundity, ending with the unification of the two divine races. In the development of this theme by Virgil, Aeneas and the Etruscans can be seen as representing the gods of sovereignty and war, and the Latins representing the gods of fecundity. Aeneas, who has brought the Trojan gods to Rome, is forced to fight with the help of the Etruscans against the Latins. It is the destiny of Aeneas to rule, and it is the fate of the Latins to share their land and women with the invaders and to accept Aeneas as their king. This resembles the unification of the warring races that climaxes the Indo-European myth.

The power exercised by the Indo-European ideological pattern on the Roman mind even under the empire is seen in the *Pharsalia* of Lucan (AD 39–65). In this historical epic, Cato, Caesar, and Pompey are depicted respectively as moral, warlike, and popular in a way that gives the story a clear trifunctional structure.

The persistence of the Indo-European pattern

Germanic epics. A typical Heroic Age occurred during the wanderings of the Germanic tribes from the 3rd to the 6th centuries AD. Out of this, too, came a rich oral tradition, from which developed in the Middle Ages many epic poems. One of the greatest of these is the Old English *Beowulf*, written down in the 8th century. Archetypal Indo-European themes also reappear in these epics. For example, the theme of the fatal fight between father and son is recounted in the German *Hildebrandslied*, of which a 67-line fragment is extant. Again, an heroic version of the Indo-European myth about the rescue of the Sun Maiden from her captivity by the Divine Twins, which also provided the basic plot of the Greek Trojan cycle and the Indian *Rāmāyaṇa*, is found in the German *Gudrun* (*c.* 1230).

Chansons de geste. The French chansons de geste are epic poems whose action takes place during the reign of

Charlemagne and his immediate successors. The *Chanson de Roland*, probably written down about the end of the 11th century, is by far the most refined of the group. The story of the poem had developed from a historical event, the annihilation of the rear guard of Charlemagne's army at Roncesvalles in the Pyrenees in 778 by Basque mountaineers. The Basques, however, are transformed in the epic to the Saracens, who to a later generation typified France's enemies in Spain. The other chansons de geste, none of which is comparable to *Roland* as a literary work, have been classified into three main cycles. The Cycle of William of Orange forms a biography of William (probably a historical Count William of Toulouse, who had, like the hero of the epic, a wife called Guibourg and a nephew, Vivien, and who became a monk in 806). Guibourg, the most faithful of wives, and the noble Vivien take prominent roles in the epic. The so-called Cycle of the Revolted Knights groups those poems that tell of revolts of feudal subjects against the emperor (Charlemagne or, more usually, his son, Louis). The Cycle of the King consists of the songs in which Charlemagne himself is a principal figure.

Arthurian Romance. The Arthurian Romance seems to have developed first in the British Isles, before being brought to the Continent by Bretons, who migrated to Brittany in the 6th and 7th centuries. The core of the legend about Arthur and his knights derives from lost Celtic mythology. Many of the incidents in the former parallel the deeds of such legendary Irish characters as Cú Chulainn, an Ulster warrior said to have been fathered by the god Lug, and Finn, hero of the Fenian cycle about a band of warriors defending Ireland, both of whom are gods transformed into human heroes. The earliest extant works on Arthurian themes are four poems of Chrétien de Troyes, written in French between 1155 and 1185: *Erec*, *Yvain*, *Le Chevalier de la Charette* (left unfinished by Chrétien and completed by Godefroy de Lagny), and an unfinished *Perceval*. In German, after 1188, Hartmann von Aue (who also wrote two legendary poems not belonging to the Arthurian cycle, *Gregorius* and *Poor Henry*) modelled his *Erec* and *Iwein* on those of Chrétien. The story of Perceval was given a full account by Wolfram von Eschenbach (*c.* 1170–1220) in his *Parzival* and in the unfinished *Titurel*. Another incomplete work of Wolfram, *Willehalm*, deals with the legend of William of Orange. *Tristan* of Gottfried von Strassburg is based directly on the older French version of Thomas of Britain (*c.* 1170–80). The romance proper, however, although it has similarities to the epic, differs in its lack of high purpose: fictions are told for their entertainment value rather than as models for national heroism. Developed in France in the Middle Ages, the romance is usually an adventure story with a strong love interest, intimately associated with the "courtly love" tradition of that time.

Distinctions between the epic and the romance

The epic in Japan. In Japan, there were in ancient times families of reciters (*katari-be*) whose duty was to hand down myths and legends by word of mouth and to narrate them during official ceremonies and banquets. After the introduction of Chinese letters, however, from the 4th century AD onward, these traditional tales were put in writing and the *katari-be* professional gradually died out. By the end of the 7th century, each clan of the ruling aristocracy seems to have possessed a written document that recounted the mythology and lengendary history of Japan in a form biassed in favour of the clan concerned. These family documents were collected at the command of the emperor Temmu (672–686) and were used as basic materials for the compilation of the first national chronicles of Japan, the *Koji-ki* (712) and the *Nihon shoki* (720). The myths and legends, contained in the earlier parts of these two books, derive, therefore, from the oral tradition of *katari-be*. Although no document preserves those narrations in their primitive form, it is generally assumed that they were originally in the form of poems. Many scholars believe that they were genuine epic poems produced during a period of incessant warfare around the 4th century. At that time mounted aristocratic warriors of the future imperial family struggled to extend its power over the larger part of Japan. Exploits of warriors, such as the emperor Jimmu or Prince Yamato-Takeru, in the earliest extant texts—the *Koji-ki* and *Nihon shoki* of the 8th century—probably derive from a heroic epic about the wars of conquest of the first emperors, whose legendary feats were transformed into those of a few idealized heroic figures.

The middle of the Heian period (794–1185) saw the emergence of a new class of warrior known as samurai. They attached a greater importance to fame than to life. The battles they fought became the subject of epic narratives that were recited by itinerant blind priests to the accompaniment of a lute-like instrument called a *biwa*.

Tales of the samurai

In the early part of the 13th century, tales about the wars of the preceding century, fought between the two strongest families of samurai, the Genji or Minamoto and the Heike or Taira, were compiled in three significant war chronicles. The *Hōgen monogatari* and the *Heiji monogatari* deal with two small wars, the Hōgen (1156) and Heiji (1159), in which the Genji and Heike warriors fought for opposing court factions. The structure of the two works is roughly the same. Each celebrates the extraordinary prowess of a young Genji warrior, Minamoto Tametomo in the *Hōgen monogatari* and Minamoto Yoshihira in the *Heiji monogatari;* each hero fights to the finish in exemplary manner not so much to win, for from the beginning each foresees the defeat of his own side, as for the sake of fame; and the consummate courage of the two heroes forms a striking contrast to the cowardice of court aristocrats. The bitterly fought Gempei War (1180–85), in which survivors of the Genji family challenged and defeated the Heike, is recounted in detail in the *Heike monogatari*, the greatest epic of Japanese literature. The sudden decline and ultimate extinction of the proud Heike, whose members had held the highest offices of the imperial court, illustrates the Buddhist philosophy of the transitory nature of all things; it invites the readers to seek deliverance from the world of sufferings through a faith that will take them to a land of eternal felicity at the moment of their death. The work is filled with tales of heroic actions of brave warriors. The most conspicuous is Minamoto Yoshitsune, one of the chief commanders of the Genji army: the legend of this man of military genius continued to develop in later literature, so that he has become the most popular hero of Japanese legend.

The later written epic. The vitality of the written epic is manifested by such masterworks as the Italian *La Divina commedia* of Dante (1265–1321) and the great Portuguese patriotic poem *Os Lusíadas* of Luiz de Camões (1524–1580), which celebrates the voyage of Vasco da Gama to India. In more recent times, novels and long narrative poems written by such major authors as Scott, Byron, Tennyson, William Morris, and Melville, were patterned, to some extent, on the epic. Their fidelity to the genre, however, is found primarily in their large scope and their roots in a national soil; their distance from the traditional oral epic tends to be considerable.

Among the epics written in modern times, the Finnish *Kalevala* (first ed. 1835; enlarged ed. 1849) occupies a very special position. This is because its author, the Finnish poet-scholar Elias Lönnrot (1802–84), who composed this masterpiece by combining short popular songs (*runot*) collected by himself among the Finns, had absorbed his material so well, and identified himself so completely with the *runo* singers. He thus came close to showing what the oral epic, which he could study only at its degenerative stage, might have been at its creative stage, on the lips of an exceptionally gifted singer.

BIBLIOGRAPHY. H.M. CHADWICK, *The Heroic Age* (1912), and H.M. and N.K. CHADWICK, *The Growth of Literature*, vol. 1, *The Ancient Literatures of Europe* (1932), are two classic works on European heroic poetries that are still valuable. A more comprehensive and up-to-date general survey is given in C.M. BOWRA, *Heroic Poetry* (1952); and J. DE VRIES, *Heroic Song and Heroic Legend* (Eng. trans. 1963). A.B. LORD, *The Singer of Tales* (1960), was written by an authority on the Yugoslav oral epic of the *guslari*. G.S. KIRK, *The Songs of Homer* (1962), also contains a brief and clear account of the Heroic Age and oral poetry. On Homer, see

also W. SCHADEWALDT, *Von Homers Welt und Werk*, 4th ed. (1965); and C.M. BOWRA, *Homer and His Forerunners* (1955). F.R. SCHRODER, *Germanische Heldendichtung* (1935), is a basic reference book for the study of the Germanic epic. J.B. PRITCHARD, *The Ancient Near East* (1958), gives summaries and translations of Akkadian and Ugaritic epics. A theory of the epics of the Indo-Europeans is presented by G. DUMEZIL in *Mythe et épopée*, vol. 1 (1968). For the use of mythical themes in the epics of the Indo-Europeans, see also D. WARD, *The Divine Twins: An Indo-European Myth in Germanic Tradition* (1968); and R. CARPENTER, *Folk Tale, Fiction and Saga in the Homeric Epics* (1946).

(A.Y.)

Epicureanism

In a strict sense, Epicureanism is the philosophy taught by Epicurus (341–270 BC); in a broad sense, it is a system of ethics embracing every conception or form of life that can be traced to the principles of his philosophy. In ancient polemics, as often since, the term was employed with an even more generic (and clearly erroneous) meaning as the equivalent of hedonism, the doctrine that pleasure or happiness is the chief good. In popular parlance, Epicureanism thus means devotion to pleasure, comfort, and high living, with a certain nicety of style.

NATURE OF EPICUREANISM

Atomism; pleasure; friendship

Several fundamental concepts characterize the philosophy of Epicurus. In physics, these are Atomism, a mechanical conception of causality, limited, however, by the idea of a spontaneous motion, or "swerve," of the atoms, which interrupts the necessary effect of a cause; the infinity of the universe and the equilibrium of all forces that circularly enclose its phenomena; the existence of gods conceived as beatified and immortal natures completely extraneous to happenings in the world. In ethics, the basic concepts are the identification of good with pleasure and of the supreme good and ultimate end with the absence of pain from the body and the soul—a limit beyond which pleasure does not grow but changes; the reduction of every human relation to the principle of utility, which finds its highest expression in friendship, in which it is at the same time surmounted; and, in accordance with this end, the limitation of all desire and the practice of the virtues, from which pleasure is inseparable, and a withdrawn and quiet life.

In principle, Epicurus' ethic of pleasure is the exact opposite of the Stoic's ethic of duty. The consequences, however, are the same: in the end, the Epicurean is forced to live with the same temperance and justice as the Stoic. Of utmost importance, however, is one point of divergence: the walls of the Stoic's city are those of the world, and its law is that of reason; the limits of the Epicurean's city are those of a garden, and the law is that of friendship. Though this garden can also reach the boundaries of earth, its centre is always a man.

HISTORY OF EPICUREANISM

Epicurus' predecessors were, in physics, Leucippus and Democritus and in ethics, Antiphon Sophista, Aristippus of Cyrene, and Eudoxus of Cnidus, a geometer and astronomer. Epicurus differed from all of these in his systematic spirit and in the unity that he tried to give to every part of philosophy. In this respect, he was greatly influenced by the philosophy and teachings of Aristotle—taking over the essentials of his doctrines and pursuing the problems that he posed.

Epicurus' own life and teachings. In 306 BC, Epicurus established his school at Athens in his garden, from which it came to be known as The Garden.

His works. In accordance with the goal that he assigned to philosophy, Epicurus' teaching had a dogmatic character, in substance if not in form. He called his treatises *dialogismoi*, or "conversations." Since the utility of the doctrines lay in their application, he summarized them in *stoicheia*, or "elementary propositions," to be memorized. In this respect, Epicurus was the inventor of the catechetical method. The number of works produced by Epicurus and his disciples reveals an impressive theoretical activity. But no less important was the practical action in living by the virtues taught by him and in honouring the obligations of reciprocal help in the name of friendship. In these endeavours, continuous assistance was rendered by Epicurus himself, who, even when old and ill, was occupied in writing letters of admonishment, guidance, and comfort—everywhere announcing his gospel of peace and, under the name of pleasure, inviting to love (see EPICURUS).

Doctrine of Epicurus. Philosophy was, for Epicurus, the art of living, and it aimed to assure happiness and to supply means to achieve it. As for science, he was concerned only with the practical end in view. If possible, he would have done without it. "If we were not troubled by our suspicions of the phenomena of the sky and about death," he wrote, "and also by our failure to grasp the limits of pain and desires, we should have no need of natural science." But this science requires a principle that guarantees its possibilities and its certainty and a method of constructing it. This principle and this method are the object of the "Canon," which Epicurus substituted for Logic. Since he made the "Canon" an integral introduction to the "Physics," however, his philosophy falls into two parts, the "Physics" and the "Ethics."

His "Canon"

The name canon, which means "rule," is derived from a special work entitled "On the Criterion, or Canon." It held that all sensations and representations are true and serve as criteria. The same holds for pleasure and pain, the basic feelings to which all others can be traced. Also true, and included among the criteria, are what may be called concepts (*prolēpsis*), which consist of "a recollection of what has often been presented from without. . . ." Man, therefore, must always cling to that "which was originally thought" in relation to every single "term" and which constitutes its background. Since the truth attested by each of the criteria is reflected in the *phainomena*, man must cling to these, employing them as "signs," and must "conjecture" whatever "does not appear." With the use of signs and conjecture, however, the level of judgment is reached, and thought is well advanced into that sphere in which error is possible, a state that begins as soon as single terms are tied into a proposition. Error, which consists of what "our judgment adds" to the evidence, can be of two types, one relative to what is not an object of experience, the other relative to what is such an object but for which the evidence is dubious. Each type has its own method of proof. Following the principles and methods of the "Canon," Epicurus arrived at an Atomism that, like that of the ancient naturalist Democritus, taught that the atoms, the void space in which they move, and the worlds are all infinite. But in contrast to Democritus, who had followed the deductive route of the intellect, considering the knowledge of the senses to be spurious, Epicurus, following an inductive route, assigned truth to sensation and reduced the intellect to it. On the basis of the totality of problems as Aristotle posed them in his "Physics," Epicurus modified entirely the mechanical theory of causes and of motion found in Democritus and added the concept of a natural necessity, which he called nature, and that of free causality, which alone could explain the freedom of motion of man and animals. For this purpose he distinguished three forms of motion in the atoms: a natural one of falling in a straight line, owing to their weight; a forced one due to impacts; and a free motion of declination, or swerving from a straight line. Secondly, he made finite the number of forms of the atoms in order to limit the number of sensible qualities, since each form begets a distinctive quality, and he taught a mathematical as well as a physical Atomism. Lest an infinity of sensible qualities be generated, however, by an infinity of aggregations (if not of atomic kinds), Epicurus developed, from just this concept of infinity, the law of universal equilibrium of all the forces, or "isonomy." Upon it, enclosing the events in a circle, he founded a theory of cyclic returns.

Natural necessity and free causality

As part of his "Physics," Epicurus' psychology held that the soul must be a body. It is made of very thin atoms of four different species—motile, quiescent, igneous, and ethereal—the last, thinnest and the most mobile of all, serving to explain sensitivity and thought. Thus consti-

tuted, the soul is, from another perspective, bipartite: in part distributed throughout the entire body and in part collected in the chest. The first part is the locus of sensations and of the physical affects of pain and pleasure; the second (entirely dissociated from the first) is the *psychē* par excellence—the seat of thought, emotions, and will. Thought is due not to the transmission of sense motion but to the perception of images constituted by films that continuously issue from all bodies and, retaining their form, arrive at the *psychē* through the pores. The full autonomy and freedom of the *psychē* is assured, as, with an act of apprehension, it seizes at every moment the images it needs, meanwhile remaining master of its own feelings.

The object of ethics is to determine the end and the means necessary to reach it. Taking his cue from experience, Epicurus looked to the animal kingdom for his answer. He concluded from this cue that the chief end is pleasure. He distinguished two kinds—a "kinetic" pleasure of sense and a "static" pleasure, consisting in the absence of pain—and taught that the pleasure of sense is good, though it is not good merely as motion but rather as a motion favourable to the nature of the receiving sense organ. In essence, pleasure is the equilibrium of the being with itself, existing wherever there is no pain.

The two kinds of pleasure

Epicurus concluded that "freedom from pain in the body and from trouble in the mind" is the ultimate aim of a happy life. The damages and the advantages following the realization of any desire must be measured in a calculus in which even pain must be faced with courage if the consequent pleasure will be of longer duration.

Having thus given order to his life, however, the wise man must also provide himself with security. This he achieves in two ways—by reducing his needs to a minimum and withdrawing, far from human competition and from the noise of the world, to "live hidden"; and by adding the private compact of friendship to the public compact from which laws arise. To be sure, friendship stems from utility; but, once born, it is desirable in itself. Epicurus then added that "for love of friendship one has even to put in jeopardy love itself"; for every existence, being alone, needs the other. "To eat and drink without a friend," he wrote, "is to devour like the lion and the wolf." Thus, the utility sublimates itself and changes into love. But as every love is intrepid, the wise man, "if his friend is put to torture, suffers as if he himself were there" and, if necessary, "will die for his friend." Thus, into the bloody world of his time, Epicurus could launch the cry: "Friendship runs dancing through the world bringing to us all the summons to wake and sing its praises."

If man's unhappiness stemmed only from his own vain desires and from worldly dangers, this wisdom, founded upon prudence alone, would suffice. But besides these sources of unhappiness there are two great fears, fear of death and fear of the gods. If science, however, is effective in revealing the bounds of desire and (as already seen) in quelling the fear of the gods, it can also allay the fear of death. Regarding the soul as a body within another body, science envisions it as dissolving when the body dissolves. Death, then, "is nothing to us, so long as we exist, death is not with us; but when death comes, then we do not exist." But death is feared not only for what may be awaiting man in the beyond but also for itself. "I am not afraid of being dead," said the comic Epicharmus of Cos: "I just do not want to die." The very idea of not existing instills a fear that Epicurus considered to be the cause of all the passions that pain the soul and disorder men's lives. Against it Epicurus argued that if pleasure is perfect within each instant and "infinite time contains no greater pleasure than limited time, if one measures by reason the limits of pleasure," then all desire of immortality is vain. Thus, Epicurus' most distinguished pupil, Metrodorus of Lampsacus, could exclaim, "*bebiōtai*" ("I have lived"), and this would be quite enough. He who has conquered the fear of death can also despise pain, which "if it is long lasting is light, and if it is intense is short" and brings death nearer. The wise man has only to replace the image of pain present in the flesh with that of blessings enjoyed, and he can be happy even "inside the bull of Phalaris." The most beautiful example was set by Epicurus at the moment of his death:

The two great fears

> A happy day is this on which I write to you. . . . The pains which I feel . . . could not be greater. But all of this is opposed by the happiness which the soul experiences, remembering our conversations of a bygone time.

The ultimate concentration of all his wisdom is the *Tetrapharmacon*, preserved by Philodemus: "The gods are not to be feared. Death is not a thing that one must fear. Good is easy to obtain. Evil is easy to tolerate."

The Epicurean school. Epicurus' successor in the direction of the Garden was Hermarchus of Mitylene, and he was succeeded in turn by Polystratus, who was the last survivor to have heard Epicurus. Superior to both, however, were Metrodorus and Colotes, against whom a small work by Plutarch was directed. Among the Epicureans of the 2nd century BC, mention must be made of Demetrius of Lacon, of whose works some fragments remain, and Apollodorus, who wrote more than 400 books. Much was also written by his disciple Zeno of Sidon, who was heard by Cicero in 79 BC in Athens. After Zeno, there were Phaedrus, also a teacher of Cicero, who was in Rome in 90 BC, and Patro, the head of the school until 51 BC. Already famous as an epigram writer was Philodemus of Gadara (born 110 BC). In the papyri of Herculaneum, comprising the effects of Philodemus' library, there are sizable remains of almost all of his numerous works. Epicureanism had already been introduced in Rome, in the 2nd century BC. The first person to spread its doctrines in Latin prose was a certain Amafinius. At the time of Cicero, Epicureanism was in fact the philosophy in vogue; and the number of Romans subscribing to it was, according to Cicero, very large. Among the greatest was Titus Lucretius Carus (*c.* 95–55 BC), who, in the poem *De rerum natura* ("On the Nature of Things"), left an almost complete and amazingly precise exposition of Epicurus' "Physics." The extent to which Epicurus was still popular in the 1st century after Christ is demonstrated by Seneca, who cited and defended him. To the 2nd century AD belongs Diogenes of Oenoanda, who carved Epicurus' works on a portico wall. In the same century should perhaps be mentioned Diogenianus, fragments of whose polemic against the Stoic Chrysippus are found in the church historian Eusebius. Also Epicurean, between the 4th and 5th centuries, was the epigrammatist Palladas.

Lucretius' *De rerum natura*

On account of its dogmatic character and its practical end, the philosophy of Epicurus was not subject to development, except in the polemic and in its application to themes that Epicurus either had treated briefly or had never dealt with at all. To be aware of this, it is sufficient to run through what remains of the representatives of his school and particularly of the works of Philodemus of Gadara. Epicurus' philosophy remained essentially unchanged. Once truth has been found, it requires no more discussion, particularly when it completely satisfies the end toward which man's nature tends. The main thing is to see this end; all of the rest comes by itself, and there is no longer anything to do but follow Epicurus, "liberator" and "saviour," and to memorize his "oracular words."

Epicureanism and egoism in modern philosophy. In the Middle Ages Epicurus was known through Cicero and the polemics of the fathers. To be an Epicurean at the time of Dante meant to be one who denied Providence and the immortality of the soul. In the 15th century, the notable humanist Lorenzo Valla—following brief hints by Petrarch—wrote, in the dialogue *De voluptate* (1431; "On Pleasure"), the first modern defense of the ethics of Epicurus, maintaining that the true good is pleasure and not virtue but concluding that the supreme pleasure is that which awaits man in heaven, which even the Bible calls *paradisum voluptatis.* In the 16th century, in terms of attitude and direction of thought, the first two great Epicureans were Michel de Montaigne in France and Francesco Guicciardini in Italy. Epicurean in everything, as man and as poet, was the early 16th-century classicist Ludovico Ariosto. But not until the 17th-century Provençal abbot Pierre Gassendi was the system

of Epicurus to rise again in its entirety—this time, however, by approaching truth through faith. Gassendi in 1649 wrote a commentary on a book by the 3rd-century (AD) biographer Diogenes Laërtius. This comment, called the *Syntagma philosophiae Epicuri* ("Treatise on Epicurean Philosophy"), was issued posthumously at The Hague ten years later. At the same time, in England, Thomas Hobbes, a friend of Gassendi, took up again the theory of pleasure and interpreted it in a dynamic sense, which was therefore closer to the doctrine of the ancient Cyrenaics. Starting from the premise that, in the natural state, "man is a wolf to man," he concluded that peace, without which there is no happiness, cannot be guaranteed by anything but force, and that this force must be relinquished, by common agreement, to the power of only one.

Gassendi's revived Epicureanism

During the 17th and 18th centuries, the European nation in which Epicureanism was most active was France, where its representatives were called libertines, among them moralists such as François, duc de La Rochefoucauld, and Charles de Saint-Évremonde; scientists such as Julien de La Mettrie, who believed that man could be explained as a machine, Claude-Adrien Helvetius, who reduced the ethic of the useful to a form of experimental science but who put public above private well-being, and Paul Henri Dietrich, baron d'Holbach, who gave particular importance to the physics of the atoms. The purely sensistic conception of knowledge had its most thoroughgoing theoretician in Étienne de Condillac. In England, Adam Smith, developing the ethical concepts of Hume (founded on sympathy), surmounted the egoism that is the basis of every act by using the principle of the impartial observer invoked to sympathize with one or another of the antagonists. After him, the jurist Jeremy Bentham, eliminating sympathy, reduced ethics to the pure calculus of the useful, which—in an entirely Epicurean formula—he defined as a "moral arithmetic." In the Epicurean stream lay also the Utilitarianism (*q.v.*) of the 19th century, of which the greatest representative was John Stuart Mill.

Pleasure as a lifestyle

Epicureanism in contemporary philosophy. In contemporary times, the interpretation of pleasure as a psychic principle of action was inititated by Gustov Fechner, the founder of psychophysics, and developed toward the end of the century by Freud on the psychoanalytic level of the unconscious.

Epicureanism and egocentric hedonism have few faithful representatives among 20th-century philosophers, though the viewpoint remains as a residue in some strains of popular thinking.

CRITICISM AND EVALUATION

In the first half of the 17th century, at a time when Pierre Gassendi was reviving atomistic Epicureanism, René Descartes, often called the founder of modern philosophy, offered arguments that tended to undercut Atomism. Reality is a plenum, he held, a complete fullness; there can be no such thing as a vacuous region, or the void of Atomism. Since matter is nothing but spatial extension, its only true properties are geometrical and dynamic. Because extension is everywhere, motion occurs not as a passage through emptiness, as Epicurus supposed, but as vortices, or "whirlpools," in which every motion sets up a broad area of movement extending indefinitely around itself—a view that has tended to be confirmed in contemporary gravitational and field theory.

Close to the heart of Epicureanism is the principle, which occurred also in Democritus, that denies that something can come from or be rooted in nothing. In a poem composed by an ancient monist, Parmenides of Elea (born *c.* 515 BC), this principle had been expressed in the two formulas: "Being cannot be Non-Being," and "Non-Being must be Non-Being." Though Epicurus had faithfully adhered to this principle almost throughout his system, he has been criticized for abandoning it at one point—in the swerves that he attributed to occasional atoms that take them aside from their normal paths. Epicurus abandoned the principle at this point in order to avoid espousing a physics that was inconsistent with the autonomy that he observed in the physical behaviour of men and animals. But to his Stoic critics, the swerves of the atoms were a scandal, since they implied that an event can occur without a cause. It has seldom been noted, however, that the swerve is merely a special case—a transposition into atomistic terms—of Aristotle's theory of accidents (*i.e.*, of properties that are not essential to the substances in which they occur), inasmuch as an accident, too, as Aristotle himself had stated (*Met.* I 3), is without a cause. Moreover, a similar view has been seriously advanced in modern times under the name of tychism by Charles Sanders Peirce, a philosopher of science.

To the Stoic charge that Epicurus lacked a doctrine of Providence (since he viewed the gods as being lazy), Epicurus answered that "mythical gods are preferable to the fate" posited by the Stoics. It has been suggested that he might equally well have added that the "immobile Prime Mover" of Aristotle's theology was hardly less lazy than Epicurus' gods. Epicurus' way of phrasing this issue, however, must not obscure the fact that the problem of Providence and its secularization has been a crucial one in the philosophy of history ever since ancient times.

Critique of egoistic hedonism

The effort of Epicurus to reduce the good to pleasure reflects the only criterion to which he would entrust himself, the "evidence of those passions immediately present," which give man the word of Nature. In the argument of psychological hedonism, here implied, the Epicurean holds that men as a matter of fact do take satisfaction in pleasure and decry pain, and he argues then to an egoistic ethical hedonism that identifies the (objective) good with pleasure. Most moralists, however, have felt that a thoroughgoing psychological hedonism cannot be defended; that desire is often, as a matter of fact, directed toward an object with no thought at all about the pleasure that it will bring; that a mother's impulse to save her young from danger is more fundamental than any pleasure involved (which usually comes only afterward); that the tendency of a child to imitate his parents can be, in fact, quite painful; and that, as a 19th-century Utilitarian, Henry Sidgwick, has argued, in what he called the "hedonistic paradox," one of the most ineffective ways to achieve pleasure is to deliberately seek it out.

Some scholars have even argued that an Epicurean egoistic hedonism, however foresighted it may be, must logically be self-defeating. If the view is universalized, the egoist must advocate the maximization of his enemy's pleasure as well as of his own, which can lead to actions painful to himself. In consequence, the entire branch of ethics that covers the advising or judging of other agents is banned from consideration, and it may be questioned whether such a view can comprise an ethic at all.

On the other hand, it has been argued that man is subject to antinomies, or contradictions, that no system can escape; there are dimensions in his nature that transcend the rational level. Thus, whatever its rational credentials may be, Epicureanism, as an attitude toward life that was theorized in its purest form by Epicurus, nonetheless remains one of the important forms that human behaviour has often assumed; and, at its best, it has achieved a type of asceticism that, even in retirement and solitude, does not negate company but welcomes it, finding the purest joys of life in the unique richness of human encounters.

BIBLIOGRAPHY

Editions of Epicurus and Lucretius (original languages): The first complete collection of the extant works and fragments of Epicurus is H. USENER (ed.), *Epicurea* (1887, reprinted 1966). A smaller selection, with Eng. trans. and commentary, is CYRIL BAILEY (ed.), *Epicurus: The Extant Remains* (1926), a very useful book that includes the "Vatican Fragments." All of the ethical fragments (and several other items) are published in CARLO DIANO (ed.), *Epicuri Ethica* (1946), with extensive Latin comments. CARLO DIANO (ed. and trans.), *Lettere di Epicuro e dei suoi* (1946), containing 14 letters of Epicurus and his friends taken from Pap. Herc. 1418. G. ARRIGHETTI (ed.), *Epicuro, Opere* (2nd ed., 1967), contains all of the works and fragments (including the *Peri Physeos*), with notes and an *index verborum;* to be used with caution. Lucretius can be read in the three volumes prepared, with introduction, translations, and comments, by CYRIL BAILEY, *Lucreti, De Rerum Natura* (1947).

English translations of Epicurus and Lucretius: WHITNEY J. OATES (ed.), *The Stoic and Epicurean Philosophers* (1940), contains translations by CYRIL BAILEY and by H.A.J. MUNRO. See also *The Philosophy of Epicurus: Letters, Doctrines, and Parallel Passages from Lucretius*, trans. by GEORGE K. STRODACH (1963); DIOGENES LAERTIUS, *Lives of Eminent Philosophers*, bk. 10 on Epicurus, trans. by R.D. HICKS (1925); and LUCRETIUS, *On the Nature of the Universe*, trans. by R.E. LATHAM (1964).

Fundamental studies: Valuable for its breadth and richness of detail, CYRIL BAILEY, *The Greek Atomists and Epicurus* (1928; new ed., 1964), is a fundamental work. For Epicurus' psychology and his relation to Aristotle, see CARLO DIANO, "La psicologia d'Epicuro e la teoria delle passioni," in *Giornale critico della filosofia italiana* (1939–42). A book that stresses Epicurus' anti-Platonism is N.W. DEWITT, *Epicurus and His Philosophy* (1954). Also important for its perceptive study of Epicurus' religiosity and ethics is A.J. FESTUGIERE, *Épicure et ses dieux* (1946; Eng. trans., *Epicurus and His Gods*, 1956). See also the article "Epikur," by W. SCHMID in the *Reallexikon für Antike und Christentum*, 5:681–819 (1961); L.L. WHYTE, *Essay on Atomism: From Democritus to 1960* (1961); and BENJAMIN FARRINGTON, *The Faith of Epicurus* (1967). G.D. HADZSITS, *Lucretius and His Influence* (1935), traces the influence of Epicurean ideas, especially since Roman times, in a broad perspective. For further bibliography, see PHILLIP DE LACY, "Some Recent Publications on Epicurus and Epicureanism (1937–1954)," in *Classical Weekly*, 48:169–177 (1955).

(C.D.)

Epicurus

The Greek philosopher Epicurus, creator of an ethical philosophy of simple pleasure, friendship, and retirement, founded schools of philosophy that survived directly from the 4th century BC until the 4th century AD. The principal sources for the life of Epicurus are a section of the works of Diogenes Laërtius, fragments from Philodemus, and six other notices. For a detailed account of his philosophy, see EPICUREANISM.

By courtesy of the Soprintendenza alle Antichita della Campania, Naples

Epicurus, bronze bust from a Greek original, c. 280–270 BC. In the Museo Archeologico Nazionale, Naples.

Early life and training. Epicurus was born in 341 BC on the island of Samos of Athenian parents who had gone there as military settlers. His father, a schoolteacher, was named Neocles, his mother Chairestrate; both were of the same village, the deme Gargettos. According to his own report, Epicurus began his study of philosophy at the age of 14. One account has him turning to philosophy when his schoolmaster could not explain the concept of chaos in Hesiod, an early Greek philosophical poet. His first master is said to have been the Platonist Pamphilus of Samos. Much more significant, however, is the report that Epicurus was for three years (327–324) a student in the Ionian city of Teos, where his teacher was Nausiphanes, a disciple of the naturalistic philosopher Democritus. It may have been from this source that Epicurus' atomistic theory came, which he used not as a means of studying physics but as the basis for a philosophical system that ultimately sought ethical ends.

Youthful interest in philosophy

At the age of 18, Epicurus went to Athens to perform the two years of military training required for Athenian citizenship. While there he may have heard Xenocrates, second in succession after Plato as head of his Academy, and Aristotle, who was then in Athens. One year later Epicurus rejoined his parents at Colophon, where they had gone as exiles when, at the close of the Lamian War, Athens lost Samos to the Macedonians. For the next 10 years, there is virtually no record. It seems probable that Epicurus travelled and studied, and it is reasonable to suppose that this was the period during which he developed his philosophical outlook and confirmed it in exchanges with the Platonists and Aristotelians. A letter written by him from Teos, addressed to his mother, was preserved by Diogenes of Oenoanda. At the age of 32, Epicurus began to teach, first at Mytilene and subsequently at Lampsacus, a period that lasted from 311/310 to 307/306.

In various places Epicurus met the disciples who were destined to follow him to Athens and to become of great significance as vehicles through whom the Epicurean school would achieve its mature development: at Mytilene, he met his first disciple, Hermarchus, who eventually succeeded him as head of the Athenian school; and at Lampsacus, he met Metrodorus and Polyaenus, whose death preceded the master's and whose sons Epicurus provided for in his will; Metrodorus' brother, Timocrates; Leonteus and his wife, Themista, who had been a hetaera (an independent courtesan); Colotes, whom Epicurus flattered with the pet name Colotarion; and Idomeneus and his wife, Batis, sister of Metrodorus.

Thus, apart from his two years in Athens, Epicurus spent the first 35 years of his life in Asia. This need not mean, however, that he developed an aversion to the literary circles in Athens. Instead, his Asiatic ties, which he continued to cultivate intensely all his life (including two or three actual journeys to Asia Minor) seem to have been reflected mainly in his choice of words and style and, more significantly, in the ecumenical scope of his philosophy.

Retirement in The Garden

The schools at Athens and elsewhere. When Epicurus and his followers came to Athens in 306, he bought a house and, in the garden, established a school, which came to be known as Ho Kepos (The Garden). At this time in Athens, cultural life was dominated by the Academy of Plato and the Lyceum of Aristotle, both of which had passed into the hands of successors. These schools attracted both the best theoretical students and those concerned with the application of philosophy to politics and public life. Therefore, any school that hoped to endure through this period had to enter into direct rivalry with the Academy and the Lyceum by establishing itself—as did the Stoa a few years later—in the city of Athens.

What Epicurus brought to Athens was more a way of life than a school or a community. Unlike both of the famous schools, it admitted women, and even one of Epicurus' slaves, named Mouse. It taught the avoidance of political activity and of public life, although, when one follower from a school outside Athens rose to political power and then fell, he was succoured by the school. Quite different from the usual connotations borne by the term epicurean today, life in the house and garden was simple. Water was the usual drink, although a half-pint daily ration of wine was allowed, and barley bread was eaten. During a famine Epicurus saved his students by doling out a few numbered beans daily. There was no communal property, as was the case in Pythagorean schools. Whereas the relationships of the members of the school were not platonic, in either the contemporary or any later sense, there are only the attacks of Stoic opponents to support any idea of sexual irregularity. Epi-

curus wrote clearly but in no highly organized way. There was much correspondence with students in Athens and at other schools, some letters being concerned with doctrinal matters but many seeming to be merely social and friendly.

On the day in his 72nd year (270 BC) that Epicurus died painfully of prostatitis, he dictated an affectionate and touching letter to Idomeneus—probably intended, in fact, for all of his friends in Lampsacus—which displayed the spirit in which he had remained true to his philosophy of repose and serenity even in the throes of pain. Epicurus' will left the house, garden, and some funds to trustees of the school. Remaining funds were left to honour Epicurus' deceased family and to celebrate his birthday annually and his memory monthly. His slaves were freed, and provision was made that the daughter of Metrodorus should be wed to someone in the Athenian school, with the approval of Hermarchus.

Writings. Diogenes Laërtius described Epicurus as a most prolific writer and preserved three of his letters and the *Kyriai doxai* ("Principal Doctrines"). The three letters are (1) *To Herodotus*, dealing with physics; (2) *To Pythocles* (probably a disciple's abridgement), on meteorology; and (3) *To Menoeceus*, on ethics and theology. The *Kyriai* consists of 40 short aphoristic statements. Another major source is the papyri from the Casa dei Papiri discovered at Herculaneum (1752–54), which include not only parts of his great work *Peri physeōs* ("On Nature"), originally in 37 books, but also numerous fragments of correspondence with his friends.

Assessment. Many of Epicurus' methods made him comparable to a religious figure. The breadth of his appeal in Rome during the 1st century BC is indicated by the fact that the poet-philosopher Lucretius based his work on Epicurus (Lucretius in fact held Epicurus in reverential awe); by the references to his thought by the statesman-moralist Cicero; and by the detailing by the biographer Plutarch of how Cassius soothed the mind of Brutus with his Epicurean ideas. Epicurus' atomistic theory was revived in the 17th century by Pierre Gassendi, a French philosopher-scientist.

BIBLIOGRAPHY. For sources of additional information on the life and philosophy of Epicurus, see the bibliography to the article EPICUREANISM.

(C.D.)

Epigraphy

Epigraphy (from the classical Greek *epigraphein*, "to write upon, incise," and *epigraphē*, "inscription") denotes a branch of scholarship devoted to the study of written matter recorded on hard or durable material. Because such media were exclusive or predominant in many of mankind's earliest civilizations, epigraphy is a prime tool in recovering much of the firsthand record of antiquity. It is thus an essential adjunct of the study of ancient man; it secures and delivers the primary data on which historical and philological disciplines alike depend for their understanding of the recorded past. In a narrower sense, epigraphy is the study of such documents as remain of the written self-expression of early cultures and as communication media in their own right, attesting to the development of visible sign systems and the art of writing as such. Finally, in later periods including our own, where perishable writing media predominate, epigraphy affords insights into the styles and purposes of monumental or otherwise exceptional techniques of written recording.

Uses of epigraphy

MATERIALS AND TECHNIQUES

The delimitation of epigraphy vis-à-vis contiguous and related areas of antiquarian scholarship meets with some ambiguity. In a wide sense, epigraphy concerns itself with the total firsthand transmission of the written remains of ancient civilizations (as opposed to *post-factum* copying) and the nature of the material (*e.g.*, stone, marble, metal, clay, terra-cotta, pottery, wood, wax tablets, papyrus, parchment) and the technique of recording (cutting, carving, engraving, casting, embossing, scratching, painting, drawing, etc.) have mere secondary relevance. Under this maximum definition certain subdisciplines may be included under the overall canopy of epigraphy: notably numismatics, which concerns itself with legends on coins and medals, and papyrology, the study of a special type of perishable record that is normally preserved only in the dry climate of Egypt and in adjacent desert regions. In the case of Egypt, papyrology tends to impinge upon wood and clay media as well, thus leaving mainly stone and metal objects as the concern of epigraphy proper.

In general, however, unless so subdivided, epigraphy encompasses inscriptions at large, be they on primary writing surfaces or on such assorted objects as vases, potsherds, gems, seals, stamps, weights, rings, lamps, and mirrors. A further related discipline is paleography, which concerns itself with the study of scribal hands and styles of writing and has significance for the dating of epigraphic as well as other written documents.

The nature of the materials and techniques used for inscriptions is closely tied to the external purpose of the record itself. Thus, inscriptions may be divided into monumental, archival, and incidental. Monumental inscriptions were intended for enduring display and were therefore, as a rule, executed in lasting material, such as stone or metal. Maximal exposure to mortal eyes need not have been the prime purpose of their originators—*e.g.*, the tomb chambers of Egyptian pharaohs, intended to be sealed forever, had their inner surfaces covered with monumental hieroglyphs; the great Bīsitūn inscription of King Darius I of Persia is on a high rock surface and legible only after precarious rock-climbing or from airborne conveyances. Under this classification may be included also micromonumental inscriptions found on such objects as coins, seals, and rings, meant to endure in their own right.

Monumental, archival, and incidental inscriptions

Archival inscriptions were essentially a feature of those early societies that kept records and that used such materials as have been preserved thanks to their intrinsic, accidental, or incidental durability. Many ancient Near Eastern cultures employed clay tablets for writing, which they fired to insure their soundness. Minoan and Mycenaean archivists in ancient Crete and Greece used perishable temporary clay records that were preserved by unintentional baking in the conflagrations that destroyed their storerooms. Papyrus records from Egypt have survived as a result of climatological chance—mainly low humidity. The official purposes of public display and of archival preservation were sometimes complementary, and therefore coincidental or overlapping matter has been preserved. In some cultures the techniques employed in monumental and archival writing tended to differ (notably in Egypt, where increasingly cursive hieratic or demotic script contrasted sharply with monumental hieroglyphic), and occasionally the language itself would be different (for example, in the Hittite Empire, where the clay tablets in cuneiform employed mainly straight Hittite or Akkadian, whereas the monumental "hieroglyphic" rock inscriptions and seals used a distinct language).

Incidental inscriptions may be defined as those not seriously meant for preservation. They include, for example, wall scrawlings of the graffiti type and casual records that were kept on cheap writing matter such as potsherds (*ostraka*) and scraps of papyrus. Many a city dump of ancient Egypt has yielded a rich harvest for the study of daily life.

INSCRIPTIONS AS HISTORICAL SOURCE MATERIAL

In studying the political, administrative, legislative, and dynastic records of extinct civilizations, the modern historian must bring to bear all the evidence at his disposal; and such evidence may vary sharply from one locality and period to another. Historiography in the modern sense—the analytical ordering and interpretation of past institutions and events—is an invention of ancient Greece, and even there it only gradually eschewed the fabulous. In many early societies (*e.g.*, Mesopotamia, Egypt), chronographic records were either annalistic or legendary in kind and sometimes apologistic or propa-

Evidence of extinct civilizations

gandistic in purport; and in others (*e.g.*, India) a cyclical world view prevailed and a lack of feeling for linear time depth precluded an ordered appreciation of the past, leaving only legendry—in such cases only synchronisms with time-anchored events elsewhere allow a precise dating.

Thus, the amount of predigested ancient information bearing on antecedent events may vary from sophisticated literary to scrupulous epigraphic, or may be wholly lacking or largely valueless. In the latter instances the historian is almost exclusively dependent for native information on primary documents, and such documents are in most cases inscriptional.

Ancient Mesopotamia. Surviving epigraphic matter from the 3rd and early 2nd millennia BC includes both historical and quasi-historical material. The Sumerian king list is a compilation of names, places, and wholly fabulous dates and exploits, apparently edited to show and promote time-hallowed oneness of kingship in the face of the splintered city-states of the period. The Sargon Chronicle is a piece of literary legendry concentrating on spectacular figures and feats of the past, whereas contemporary royal inscriptions, notably by Sargon I of Akkad and Gudea of Lagash, are historical documents in the proper sense.

Both kinds of texts are preserved also from the Babylonian and Assyrian periods, from the reign of Hammurabi (1792–1750 BC) to the 6th century BC. There are lists of date formulas and year names from Hammurabi's reign and from that of his son Samsuiluna; lists of Assyrian eponymous year names, based on those of dignitaries; the Babylonian king lists, running from Hammurabi through the Kassite era and the Assyrian domination of Babylon to the last flicker of Babylonian self-assertion in the early 6th century BC; the Assyrian king list from Khorsabad, which made good use of earlier compilations; and notably the so-called Synchronistic Chronicle, which juxtaposed the kings of Assyria and Babylonia in the same millennial sequence. Historical documents comprise, above all, the stately sequence of annals by the kings of Assyria, recorded on stone slabs, stelae, foundation markers of buildings, bronze gates, statues, and obelisks and in clay archives (prisms, cylinders, tablets). Starting in the Old Assyrian period, they were especially extensive in the reigns of Tiglath-pileser I (1115–1077 BC), Ashurnasirpal II (883–859 BC), Shalmaneser III (858–824 BC), Adad-nirari III (810–783 BC), Tiglath-pileser III (744–727 BC), Shalmaneser V (726–722 BC), Sargon II (721–705 BC), Sennacherib (704–681 BC), Esarhaddon (*c.* 680–669 BC), and Ashurbanipal (668–627 BC).

For all their swaggering bombast and flaunting of deliberate cruelty, the annals provide prime historical source material. The detail of the Assyrian conquest of Syria, Palestine, parts of Asia Minor, Cyprus, Arabia, and Egypt would be spotty indeed without recourse to these annals, for they show the centre of political power, unlike such provincial records as those from contemporary Egypt or the Old Testament.

Legal compilations and law codes

Legal compilations and law codes also have pride of place in the epigraphic record of ancient Mesopotamia. These form a unique succession, starting in the 3rd millennium BC with that of King Ur-Nammu of the Sumerian 3rd dynasty of Ur (*c.* 2100 BC), continuing with those of the Sumero-Akkadian king Lipit-Ishtar (in Sumerian) and King Bilalama of Eshnunna (in Akkadian, during the interval of the 3rd dynasty of Ur) and the rise of the Amorite dynasty of Hammurabi (*c.* 2000 BC), culminating in the great diorite stela of Hammurabi (*c.* 1750 BC), showing retardation and recrudescence in the Middle Assyrian laws that are found on clay tablets at Ashur (at the time of Tiglath-pileser I), and petering out in the fragmentary Neo-Babylonian laws dating from the 7th century BC.

The stela of Hammurabi must have been originally set up in some Babylonian population centre for the literate to read and know their rights. Some Elamite invader must have carried it off to Susa (perhaps *c.* 1200 BC), where it was found in 1901 and removed to the Louvre in Paris. The bulk of the stela contains the text of the code, partly erased on the obverse but restorable in some measure from clay-tablet versions of the same laws. The top depicts the king in a worshipful pose, receiving the laws from the sun god, Shamash. In reality Hammurabi —the sixth of 11 kings of the Old Babylonian or Amorite dynasty—was a practical codifier rather than a revelatory mediator of law. His code was an effort to fuse into a workable whole the ancient inheritance of Sumerian-based jurisprudence and the Semitic talion law (punishment according to the "eye-for-an-eye and tooth-for-a-tooth" principle) of the Akkadian superstratum. The result is not a model of economy or arrangement or logical organization, but the code of Hammurabi constitutes nevertheless the first great legal monument in the history of mankind. The later Assyrian laws show traces of further removal from the cradle of Sumerian civilization since they are both harsher and noticeably more primitive.

Ancient Egypt. Egypt attracted the special curiosity of the Greeks, and Herodotus (5th century BC) devoted an entire book to on-the-spot observations and fanciful tales about the land of the Nile. The lost *Aigyptiaka* (or *Aegyptiaca*) of Manetho (3rd century BC) contained the roster of 30 dynasties, which still underlies the chronology of ancient Egypt. Such classical writers as Strabo, Plutarch, and Pliny the Elder all dealt with various aspects of Egyptian antiquities.

Yet the fund of knowledge would be woefully skeletal and inaccurate without the explicit testimony of contemporary records from Egypt itself. The decipherment of the Egyptian writings gave the impetus to Egyptian epigraphy. The progress of excavations multiplied the corpora of texts, especially adding the papyrological dimension. In addition, cuneiform Akkadian on clay tablets was the international diplomatic medium of writing during the most brilliant phases of Egyptian history and is hence an integral part of the Egyptian epigraphic record.

Decipherment of Egyptian writings

The historically significant Egyptian epigraphic texts, apart from their external peculiarities, have likewise special traits relating to genres. There is little attempt at historiography and great fluctuation in bulk in the course of dynastic vicissitudes. They are mainly annalistic and thus firsthand accounts of pharaonic or other high-level deeds; but the peculiar features of stylization, stereotyping, and usurpation must frequently give the careful historian pause and sometimes debase the face value of the record. Written monuments became somewhat numerous during the 4th dynasty (*c.* 2613–2494 BC), that of the pyramid builders Khufu, Khafre, and Menkaure. Notable are the fragmentary annals of Snefru, which already alluded to dealings with Asia, the chronic goal of Egyptian territorial ambitions. Historic records persisted under the following two dynasties, with particular articulateness in the reign of Pepi I, third king of the 6th dynasty (*c.* 2345–*c.* 2181 BC), then subsided until a modest re-emergence in the (Theban) Middle Kingdom of the 12th dynasty (1991–1786 BC). Another silence shrouded the period of the Hyksos kings (*c.* 1720–1570 BC), broken only subsequently by such retrospective revulsion at the memory of barbarian domination as that in Queen Hatshepsut's temple inscription (*c.* 1475 BC) at Istabl Antar in Middle Egypt. The golden age of historical recording began in the 15th century BC with the central rulers of the 18th dynasty, notably Thutmose III and Amenhotep II and III. Thutmose's annals on the walls of the temple of Karnak describe 20 years of ceaseless military activity in Asia, some 16 campaigns in all, and are supplemented by stelae from Armant in Upper Egypt and Gebel Barkal near the Fourth Cataract, as well as by lists of conquered lands at Karnak. Similar material continued in the reigns of Amenhotep II and III, in the latter's case importantly supplemented by the cuneiform correspondence with foreign powers (Mitanni, Arzawa, etc.), which was subsequently stockpiled and archived by Ikhnaton in his transitory new capital, where it lay buried to await the modern excavators of Tell el-Amarna. Ikhnaton's religious preoccupations (he changed the offi-

Golden age of Egyptian historical recording

cial religion to the worship of the sun god Aton), and political apathy led to the loss of many of Egypt's Asian possessions. Records of Ikhnaton's short-lived son-in-law, Tutankhamen, at Thebes (*c.* 1350 BC), make recantation and restoration for the heresy. Tutankhamen's successor, the warlord-pharaoh Horemheb, left boastful accounts of foreign conquest that sound suspiciously grandiose in relation to plausible reality.

In the 19th dynasty (1320–1200 BC) Seti I went to war against the Syrians, Hittites, and Libyans, letting the world know about it on the walls of Karnak. But in this respect he was no match for his long-lived son, Ramses II, who usurped the monuments of others and covered unprecedented amounts of wall space with his own real or inflated exploits. (The Battle of Kadesh against the Hittites in 1299 BC, which ended in a stalemate, was given lavish coverage as a triumph on temple walls at Karnak, Abydos, and Abu Simbel.) In the 20th dynasty (1200–1085 BC) occurred incursions of the "sea peoples," and the records of Ramses III detailed both the crisis and the increasing accumulation of wealth and power in the religious establishment. Subsequently, Egyptian history receded from the world scene, with "Libyan" and "Ethiopic" dynasties and a brief Saite renaissance of the 26th dynasty (664–525 BC), already under the Assyrian and Babylonian shadow, soon to be replaced by the Persian. The firsthand political records declined accordingly, although they remain of significance for local history down to the Ptolemaic era, a dynasty that ruled Egypt beginning in 305 BC, founded by Ptolemy I Soter, a general under Alexander the Great.

No law codes have been found in Egypt, presumably because codification was not practiced. There are, however, royal administrative and legal decrees granting privileges and immunities and also records of legal proceedings, especially of the Theban tomb-robbery trials during the 20th dynasty.

Other ancient Near Eastern regions. Regions adjacent to the power centres of Mesopotamia, Egypt, and Iran were frequently mere political and administrative adjuncts, often obscure vassaldoms or adversaries without notable or attested written traditions. The Mitanni kingdom in northern Mesopotamia had some ephemeral big-power dealings with Egypt in the days of Amenhotep III, but its capital city is still lost in the sands, and thus its presently known epigraphic tradition is merely part of the correspondence in the Tell el-Amarna archives. The records of the Elamite kingdom with its capital at Susa were mostly ancillary to Mesopotamia in the 2nd millennium BC, and to Iran later on. The region of Syria and what later came to be called Palestine was in the 2nd millennium the object of an extended tug-of-war between Egypt and the Hittite kingdom.

The Hittite Empire. The Hittites were, in fact, the third great international power in the Near East during part of the 2nd millennium BC, and the epigraphic yield of their royal archives at Boğazköy in central Asia Minor matches or even surpasses in richness that of Mesopotamia and Egypt for the few centuries in question. The cuneiform records of the Hittites contain a tradition of unique royal political self-expression. These documents begin with the oldest known Hittite text, the inscription of the early ruler Anittas, detailing dynastic struggles of an obscure and possibly apocryphal past. From the founder of the Old Kingdom, the firmly historical Hattusilis I (Labarnas II), came an annalistic autobiography (excavated in 1957) and a "farewell address," or political testament, in Hittite as well as Akkadian versions. Subsequent events, including the capture of Babylon by Hattusilis' son, Mursilis I (*c.* 1590 BC), and the succeeding era of regicidal upheavals, are known from an edict of King Telipinus, who detailed them as he set about regulating the rights of royal succession. The subsequent founder of the Hittite Empire, Suppiluliumas I (*c.* 1350 BC), and his son Mursilis II left annals detailing their military and political deeds. Mursilis was a particularly prolific annalist and edited his father's annals as well. The great encounter with Ramses II at Kadesh in 1299 BC occurred in the reign of Mursilis' son, Muwatallis, and left an echo in the autobiography of his brother and successor, Hattusilis III. Hattusilis' autobiography is a tract of self-justification for a breach of the edict of Telipinus in deposing his nephew and predecessor Urhi-Teshub (Mursilis III).

The earliest Hittite texts

Other Hittite documents inveigh against treasonable behaviour ("Indictment of Madduwattas") or contain detailed instructions for military, civil, and court officialdom. Hittite queens had prerogatives of independent high-level initiative, and examples of their correspondence with foreign potentates supplement the archives of their husbands. The most remarkable external political documents are numerous state treaties, sometimes between equals but more often covenants specifying protectorate or vassaldom status for subordinate states on the fringes of the kingdom. Equally notable is the Hittite Law Code, relatively enlightened and mild in the face of its contemporary counterparts in Mesopotamia. Altogether, the inscriptional documents are practically the exclusive source material for knowledge of the Hittites; not even the existence or location of their empire was surmised prior to the discovery of their archives.

Syria-Palestine. After the collapse of the Hittite Empire (*c.* 1190 BC), significant records from Asia Minor ceased for many centuries, whereas local history in the Syro-Palestinian area was recorded in the inscriptions of petty dynasts increasingly under the shadow of Assyrian domination. The break with the past is evident in the writing systems (Hittite Hieroglyphs or West Semitic alphabet rather than cuneiform) and in the languages (Indo-European Anatolian, Canaanite, Aramaic). Into this category fall the stela of King Mesha of Moab (*c.* 830 BC) now in the Louvre, the Phoenician-language hieroglyphic inscription of Azitawadda of Adana (late 8th century BC), and those of the kings of Ya'diya-Sam'al. Contemporary cuneiform documents from the Urartu kingdom around Lake Van in eastern Anatolia are historically and culturally an offshoot of the history of 8th-century Assyria.

Ancient Iran. Epigraphically recorded history in ancient Persia began dramatically with the rise of the Achaemenid dynasty in the 6th century BC. Cyrus II the Great's conquest of Media, Lydia, and Babylonia, Cambyses' occupation of Egypt, and the incursions into Greece of the succeeding side branch of the family, beginning with Darius I, created in short order a world power destined for centre stage on the international scene for the following two centuries. The international character of the empire is reflected in the frequently trilingual royal inscriptions—with Akkadian and Elamite versions in traditional syllabic cuneiform, and the Old Persian text in its own simplified quasi-alphabetic system of wedge-shaped writing. The Achaemenids' time span ranges from Darius' great-grandfather, Ariaramnes, to their less glorious progeny, who were ultimately extinguished by Alexander the Great. The empire was centred in Persia; but a granite stela of Darius, found near the Suez Canal, recorded in Old Persian, Elamite, Akkadian, and hieroglyphic Egyptian the opening of a canal from the Red Sea to the Nile. The epigraphic material included rock surfaces, building walls, columns, doorways, cornices, statues, and doorknobs; bricks, plaques, plates, and tablets of clay, stone, gold, and silver; vases, weights, and seals. Almost all the longer texts were by Darius and Xerxes I; an important one of Xerxes was found on a stone tablet at Persepolis in 1967. The great Bīsitūn rock inscription of Darius runs to several hundred long lines in Old Persian alone, besides the Elamite and Akkadian versions. It is accompanied by 11 minor inscriptions, serving as keys to the sculpted scene of the panel, which shows Darius triumphant over the usurping impostor Gaumata and nine other rebels. The text is a self-statement of how Darius gained and consolidated his rule. It apparently had currency in the realm, apart from being tucked away on a sheer cliff wall, for a partial duplicate of the Akkadian version has been found on a dolerite (basalt) block from Babylon, and papyrus fragments from Elephantine have yielded scraps of an Aramaic edition. One peculiar documentary value of the text is

Rise of the Achaemenid dynasty

that the same period in Persian history is extensively covered in the Greek literary tradition by Herodotus, Ctesias, and others, and scholars can thus juxtapose Darius' own accounts with those of almost contemporary foreign historians. As an example, Darius stressed his role as saviour of the fatherland from the clutches of an upstart who pretended to be Smerdis, the brother of Darius' predecessor Cambyses. The latter had murdered Smerdis and was carrying on various outrages in Egypt when word came of the impostor's takeover back home. Darius stated that thereupon Cambyses "died his own death," meaning that it was a fatal matter without human interference, and that thus Darius' hands were clean in taking action against the impostor. According to Herodotus, Cambyses' legendary death involved a freak occurrence as he prepared to leave for home—an accidentally self-inflicted wound leading to gangrene. Both the meaning and intent of Darius' description are thus confirmed.

Juxtaposition of Darius' texts with Herodotus' historical writings

From later Arsacid (Parthian) and Sāsānid periods of Iranian history, there are likewise royal inscriptions that shed light on their respective eras down to the Islāmic conquest in the 8th century AD, and new specimens are still being discovered.

Ancient India. India's past became anchored in historical time and separable from legend only with the establishment of firm synchronisms with outside data. One such link is the Seleucid embassy of Megasthenes to the Maurya king Candragupta (Greek Sandrokottos) at Pāṭaliputra (Greek Palimbothra) in Magadha (modern day Bihar). The Maurya dynasty was continued in the early 3rd century BC by Candragupta's son Bindusāra (Amitrochates in the Greek sources) and had extended its power over much of the subcontinent. But then the Greek sources fall silent, and Indic literary tradition supplies only the usual web of timeless legendry. At this point, however, epigraphy makes a unique contribution in the form of the first authentic and datable historical documents from India, the edicts of Bindusāra's son and successor Aśoka. As a matter of epigraphic fact, Aśoka ruled all of northern India and a large portion of the south, from Taxila and beyond to Mysore and Kaliṅga (coast of Orissa and Andhra Pradesh). His 14 rock edicts and seven pillar edicts in numerous versions and copies, plus separate minor texts, are scattered over this expanse —in the Prākrit language of his time and in the Brāhmī script, except for some northwestern examples of the Aramaic-inspired Kharoṣṭhī writing. Even a Greek-Aramaic bilingual version was found in 1958 near Kandahār in Afghanistan. Aśoka's edicts are proclamations and ordinances in a Buddhistic spirit, designed to impart good order, morality, and moderation by the Emperor's personal enjoining and example. Particularly notable is the 13th rock edict, which bares the ruler's pangs of conscience over the conquest of Kaliṅga eight years after his coronation, his continuing sorrow over the cruelties committed, and his pledge to substitute the victory of Buddhist religious law (*dhamma*) for all earthly conquest.

Aśoka's edicts would rate a mere historiographic footnote for their inconsequential transitoriness, were it not that in the same breath Aśoka supplied the very synchronisms that are the main key to ancient Indian chronology. Among his western neighbours he mentioned Aṃtiyoge (Antiochus II Theos of Syria), Tulamaye (Ptolemy II Philadelphus of Egypt), Aṇtekine (Antigonus II Gonatas of Macedonia), Maka (Magas of Cyrene), and Alikasudaro (Alexander of Epirus or Alexander of Corinth). The dates of these contemporaries circumscribe the time of Aśoka's reign; combined with the earlier Greek synchronisms, they afford a firm foundation for the correlation of Indic and Mediterranean events. The edicts of Aśoka are thus a prime example of the value of inscriptions for historiographic dating and constitute a fixed record unparallelled in ancient Indian tradition. Later periods of Indian history, such as those of the Indo-Scythian and Gupta rulers, are also represented in epigraphic documents of some historical value.

The main key to Indian chronology

Ancient China. In China also, inscriptions are a means of separating chronological fact from historiographic legend. Nonepigraphic book composition on wood or bamboo strips had an early history in China, beginning in the later 2nd millennium BC; its scope was such that the Ch'in emperor Shih Huang Ti went down in history as a book burner in 213 BC. The San Tai, or three periods of early Chinese history (Hsia, *c.* 2205–1766 BC; Shang, *c.* 1766–1122 BC; Chou and Ch'in, *c.* 1122–206 BC), were long considered by Western scholars purely legendary down to the early Chou period, and the literary documents (such as the *Shu Ching* or "Classic of History") were dismissed as compilations consisting mostly of successive overlays of little historical value. But the historicity of written records from the later Shang era (*c.* 1400–1122 BC) is now apparent from the mass of inscribed archeological material found especially in northern Honan Province. These include, in particular, the so-called oracle bones (mostly tortoise shells and scapulae of animals), bearing incised records of royal divination. At the site of the Shang capital, Yin, were discovered inscribed vessels of bronze, bone, pottery, jade, and stone, probably ceremonial in nature and related to official ritual uses such as ancestor worship. The script is a mixture of pictograms, word signs, and phonograms. From the ensuing Chou era, bronze inscriptions of official provenance have likewise been found, especially records of royal largesse. Inscriptions from later periods form a steady but subsidiary source of information beside the larger, nonepigraphic written record.

Ancient Greece. The historically significant epigraphic record of classical Greece differs in many ways from most of those discussed above. Much of it is parallelled by a mature and independent tradition of professional literary historiography. Except for the preclassical Helladic (Mycenaean) period of the 2nd millennium BC (see below), there was no archival tradition, although the bulk of "monumental" records sometimes approximates the same purpose of massive preservation. There was no all-important power centre and no dominant rulership before Hellenistic and Roman times: thus the geographical scattering of records was extreme, although naturally with some focuses of emphasis such as Athens. Above all, there was continuity from the inception of literacy, with gradual but steady increase in bulk.

The special circumstances of the Greek epigraphic record

Epigraphically transmitted historiography in Greece is extremely scarce because the probing of past events has passed beyond the stage of dynastically centred and sheltered annalism; an example is the "Marmor Parium" (from the island of Paros and now at Oxford), which contains a chronographic rundown of traditional dates and events of Greek history. Rather than monolithic records of autocracy, there is history in the making by a plethora of tyrannical, oligarchic, or democratic microentities. Treaties of alliance and various other agreements between the multiple city-states—recorded on metal or stone and publicly displayed, or consecrated at such pan-Hellenic sanctuaries as Delphi or Olympia—form an important part of the epigraphic yield. Joint-citizenship covenants, decrees concerning the return of exiles, monetary agreements on coinage and debts, and tribute lists are typical examples. They are supplemented by the records of arbitration of interstate disputes, most often boundary matters, by third-party commissions. Thus, when around 240 BC a territorial disagreement arose between Epidaurus and Corinth, the Achaean League appointed a group of 151 Megarians as mediators, and their report survives. Further extensions of such "international" documentation are the *proxenia* decrees, which amount to letters of patent and resolutions of appreciation issued by one state to a citizen of another for service as *proxenos*, a kind of honorary consul looking after the interests of the other state's citizens. The extensive colonization efforts by the Greeks around the Mediterranean produced a further kind of political document—regulations governing conditions for emigration and return, citizenship rights of the colonists, and relations between the colony and the mother community. Not all historically meaningful international records are of the monumental type. Greek mercenaries of Pharaoh Psamtik II (ruled 594–589 BC) left their scrawlings on the legs of a colossal statue at Abu Simbel on the upper

Nile, proving by their names and dialect that they came from Rhodes and Ionia and were far abroad on foreign adventure.

Internal documents of the various Greek states include numerous records of decrees and ordinances, both administrative and legislative. Stereotyped Athenian ones are complemented by variant forms in other localities; most contain a preamble setting forth the date and the officialdom in charge, the circumstances occasioning the action, the decision itself, means and sanctions for its enforcement, and sometimes instructions for providing and affixing the very physical record that has been preserved. Sometimes they amount to formal laws, such as those directed against extravagances in funeral practices.

Financial data of the states were minutely and permanently inscribed on stone, and the accounts thus displayed recorded in detail the receipts, expenditures, and balances of public funds. Very specific reports cover projects of public construction, including both technical and budgetal details, allowing sometimes the integral modern reconstruction of the buildings from the reports alone. The records of the Erechtheum and the Parthenon at Athens are well preserved, as are inventories of military expenditures, especially those of the Athenian navy. Knowledge of the ephebic system at Athens, a paramilitary youth organization, is in the main based on epigraphic material.

The laws of Gortyn

The only law code in the Greek epigraphic tradition is the laws of Gortyn in central Crete, inscribed on the slabs of a circular wall which, if completely preserved, would have been nearly 100 feet (30 metres) in diameter. The 12 columns of text, each on four layers of stone and some five feet (1.5 metres) high, are about 30 feet (nine metres) in sideways length and contain more than 600 lines of text, being the longest Greek epigraphic monument; parts of some columns of further text survive, the so-called Second Code. The probable date of this inscription is the first half of the 5th century BC. The code deals with such matters as disputed ownership of slaves, rape and adultery, rights of a wife upon divorce or death of husband, disposition of children born after divorce, inheritance, sale and mortgaging of property, ransom, children of mixed marriages, and adoption. While self-contained, it evidently does not represent the entirety of laws; curiously, it stresses those areas of civil law (inheritance, adoption) that are notably lacking in the Hittite Law Code. The uniqueness of this code in the Greek world points up the relative isolation and marginality of the Cretan tradition, with tendencies to codification more reminiscent of Anatolia and the Near East generally.

Ancient Rome. While partly overlapping Greek inscriptions in time and type, those of Rome nevertheless present distinct peculiarities. There is a high measure of standardization in kind and style, despite lingering local traditions in more remote areas. Extensive and excessive use was made of initials and abbreviations, to the point of serious impediments to comprehension; lists of such abbreviations are standard adjuncts to modern handbooks on Latin epigraphy. Stone and bronze were standard material, but there was more use made of bricks, tiles, and terra-cotta, and practices of stamping and signing such matter are of help in identification and dating.

Literary and epigraphic records of early republican Rome are scant and fragmentary. Latin was at the time still largely confined to Rome proper, with Oscan, Etruscan, and colonial Greek spoken and written in much of Italy. With the arrival of extended political power there was little early literacy to fall back on, and historiographic attempts at retrospection ended in epicized myth and legendry (*e.g.*, in Livy). The Greeks on the southern coastal fringe had little truck with the hinterlands of early times. The Etruscan impact on Rome is evident, but shortcomings in discovering epigraphic records of Etruscan city sites (as opposed to necropolises) and in understanding the Etruscan language, limit the historical data derivable from Etruscology. The potential for such illumination is seen from the discovery of gold tablets at Pyrgi in 1964 that contain a dedication in Etruscan and Phoenician by the Etruscan king of Caere, Thefarie Velianas, to the syncretized goddess Uni (Juno) Astarte. Datable to around 500 BC, the text shows Etruscans ruling in the outskirts of Rome, with enough Phoenician or Punic (Carthaginian) maritime presence to warrant symbiotic and syncretistic bilingualism. The vital historical import of such attestations, pieced together with later Greek and Roman historiographic data, is patently manifest.

Texts from republican Rome

No historically important epigraphic Latin text from republican Rome antedates the 2nd century BC. The marble Columna Rostrata—found in Rome in 1565 and now at the Palazzo dei Conservatori on the Capitoline Hill—records a naval victory of Duilius (consul in 260 BC) over the Carthaginians; but the inscription, replete with fake archaism, dates from a restoration effort in early imperial days. The *fasti consulares* and similar lists afford a summary sequence of consulates, magistratures, and triumphs. The one truly significant epigraphic historical text is the "Res gestae divi Augusti," an autobiographical record of Augustus' rule, which was exhibited in many places but is best known as the Monumentum Ancyranum, from the bilingual (Latin and Greek) version carved on the walls of the Temple of Rome and Augustus at Ankara (Turkey).

By the time the epigraphic record became abundant, Rome's domination was secure and the political documentation was one of imperial outflow and of local sycophancy. Treaties of republican Rome with foreign powers survive merely in the works of literary historians. Among "internal" documents from republican days are several epigraphic texts of significance: the Senatusconsultum de Bacchanalibus, on a bronze tablet found in 1640 in Bruttium (the "toe" of Italy) and now in Vienna, is a consular edict on Senate authority, regulating Dionysiac outbursts in Italy in 186 BC; pieces of the laws Lex Acilia Repetundarum (123 BC) and Lex Agraria (111 BC) were found in the 16th century on opposite sides of what was once a large bronze tablet; the local laws of the town of Bantia (on the borderlands of Lucania and Apulia in southern Italy) are inscribed on a fragmentary bronze tablet found in 1790 (now in Naples), with a Latin-language text on one side and the longest known Oscan inscription on the other, both datable to the late 2nd century BC; parts of the Lex Cornelia de Viginti Quaestoribus (81 BC) are preserved on a large bronze tablet found at Rome; Julius Caesar's Lex Julia Municipalis of 45 BC was found near Heraclea in Lucania. On the whole, however, the transmission of Roman law, from the earliest fragments to the mature codifications, is nonepigraphic. In later times the flood of administrative decrees increases with the growth of centralized autocracy. Typically Roman epigraphic material of imperial date comprises further building inscriptions, military records, and honorific texts.

The oldest Turkic monuments

The Turkic peoples. The oldest monuments of Turkic languages—inscribed on stones, and datable to the early 8th century AD—were discovered in the late 19th century in southern Siberia around the Yenisey River and in northern Mongolia near the capital of Urga (modern Ulaanbaatar). Deciphered in 1893 by the Danish scholar Vilhelm Thomsen, they provide valuable insights into the history of Central Asia around the 7th century AD. These records of the Turk dynasty (Chinese T'u-chüeh) comprise especially texts found at Kosho-Tsaidam on the Orhon (Orkhon) Gol (river), including also Chinese text. These texts throw light on the nomadic culture of the tribal empire controlled by the Turk dynasty, including shamanism, calendar, customs, and social structure, with strong Chinese influence detectable in the latter. After the decline of the Turk people (*c.* 745), their successors, the Uighurs, perpetuated for a time the same kind of monumental dynastic epigraphy, the writing system of which is an offshoot of the Aramaic alphabet, presumably mediated by the Iranian-speaking Sogdians of Central Asia. Gradually, however, new scripts took over (especially the so-called Uighur alphabet, of Syriac origin, which was further transmitted to the Mongols and the Manchus) and inscriptional monuments gave way to manuscript records

such as those found in Chinese Turkistan (Turfan) in the late 19th century (along with texts in Sanskrit, Sogdian, Tocharian, and other Indo-European idioms), attesting to a coexistence of Buddhist, Manichaean, and Nestorian Christian religious communities. The later Turkish peoples, including the Anatolian Seljuqs and Ottomans, had an Islāmic book-tradition, to which the inscriptional record is merely incidental.

Northern Europe. The advent of writing was slow north of the Alps; it came either from direct expansionary exportation by Greek coastal colonies and the Roman Empire, as in Gaul and the Iberian Peninsula, or indirect inspiration from the same quarter, as in writing in the Irish and British ogham alphabet and the Germanic runes.

Celtic inscriptions. Celt-Iberian inscriptions from Spain and Celtic ones from Gaul and Ireland are scarce, mostly brief, and notably devoid of usable historical information, apart from their mere monumental existence and linguistic and onomastic (of or pertaining to names) content. Occasional items such as the fragmentary Gaulish Calendar of Coligny afford insights into local cultural practices, apart from an overwhelming trend to romanization.

Runes. The runic alphabet—a Germanic alphabet, originally of 24 letters, also called futhark—and its offshoots (the Scandinavian, especially Danish, 16-letter variety from the 9th century AD; and Anglo-Saxon versions, from the 3rd to the 10th centuries AD, also called futhorc) are probably of "North Etruscan" or "Sub-Alpine" Italic inspiration, datable to around 200 BC. The "North Italic" letters of the Germanic text *harixasti teiva*, "to the god Harigast," on a helmet from Negau (southern Austria) are probably from that time of transmission. Runic inscriptions from the era of migrations, ranging from eastern France through Germany up to Denmark and eastward via Poland to Romania, are supplemented by the later, richer yield from England and Scandinavia. Native Anglo-Saxon runic epigraphy, mostly in Northumbria, Mercia, and Kent, petered out around the 10th century, whereas the Scandinavian tradition (including its enclaves on British soil) endured for several more centuries. Sweden has some 3,000 runic monuments; Norway and Denmark, perhaps 400 each; while Iceland has remarkably few, apparently in inverse proportion to the literary flowering in that colonial outpost. The Vikings left their runic calling cards in far-flung places, including those in the Greek port of Piraeus, on the Black Sea coast, in Varangian Russia, in Scotland, Ireland, and the Isle of Man, and the Orkneys, Hebrides, and Shetland islands; Greenland has its share, but alleged specimens from North America, notably the Kensington stone from Minnesota, telling of the westward trek of an exploration party from Vinland, are invariably crude latter-day forgeries.

Runic monuments

The purposes of runic inscriptions were usually either dedicatory or commemorative, sometimes magic, and frequently sepulchral. The longest, that from Rök in Sweden (725 runes), seems to contain a catalog of epic deeds, possibly those of the Ostrogoth king Theoderic. The prime historical value of runic epigraphs is usually what and where they are, rather than what they depict or record.

INSCRIPTIONS AS SOCIAL AND CULTURAL RECORDS

In the preceding section, inscriptions were evaluated as sources for the presence and migrations of peoples, the existence and chronology of political states, their dynastic histories, foreign relations, internal governance, legal institutions, and official acts. In this section, epigraphy is surveyed for information about how past civilizations lived; their religious beliefs and practices; their business, financial, legal, and social relations; and what shape their aspirations assumed in terms of verbal creativity. The subdivision of civilizations surveyed differs somewhat arbitrarily from the earlier section by the omission of certain areas and the inclusion of Crete and Mycenaean Greece. In fact, Indic and Chinese epigraphic matter discussed above could just as well have fitted the "religious" slot, but its royal character and chronological importance for official history dictated otherwise. Conversely, the Cretan and Mycenaean tablets are purely economic inventories, but they might possibly have been included with history above for the very important historical fact that they prove the ruling presence of Greeks at Knossos during the 2nd millennium BC. The varying degree of importance of epigraphic material in various cultures persists: in Mesopotamia and the ancient Near East its dominance was nearly total; in Egypt it combined with the papyrological dimension; in Crete it was merely a flash in a prehistoric darkness; while in ancient Greece and Rome, it was a supplementary concomitant of the non-epigraphic literary tradition.

Ancient Mesopotamia. Ample specimens of Akkadian-language clay-tablet epistolography have been found at several sites, notably Tell el-Amarna in Egypt and Tell al-Ḥarīrī on the middle Euphrates (the ancient Mari of *c.* 1700 BC). The Amarna letters, about 400 of them, were composed in corrupt Akkadian by Canaanite scribes in Syria and Palestine and were largely official in character. The Mari letters, some 5,000 in number, are more illustrative of normal day-to-day written communication in a Mesopotamian milieu proper.

The Amarna and Mari letters

Legal inscriptions. Another aspect of everyday life in ancient Mesopotamia is amply illustrated by thousands of clay tablets of a practical legal nature, as distinct from the formal laws. These archivally preserved records from various periods use Sumerian, Old Akkadian, Babylonian, and Assyrian alike. They detail law suits, court decisions, marriage contracts, divorce proceedings and settlements, sale adoptions (fictitious acts circumventing prohibitions against land sales outside the family), loan agreements, tax receipts, and much else. Commercial inventories, such as those of the Old Assyrian merchant colony at *Karum Kanes* in central Asia Minor (20th century BC), complete the picture.

Cultic and literary texts. Due to the religious sanction of law, legal records were often stockpiled in temple archives. These latter are also the source of more directly cultic texts, such as descriptions of rituals, which come under such headings as "Temple Program for the New Year's Festivals at Babylon," "Ritual to be Followed by the Kalū (priest) when Covering the Temple Kettle-Drum," "Ritual for the Repair of a Temple," and "Program of the Pageant of the Statue of the God Anu at Uruk." Prayers, lamentations, and hymns in both Sumerian and Akkadian are extant, addressed to deities such as the goddess Ishtar, the moon god Sin, the sun god Shamash, the great triad Anu, Enlil, and Ea, and the Babylonian patron god Marduk. The Sumerian "Lament for the Destruction of Ur" bemoans the city's fall to Elamites and Subarians. Often the king himself is the spokesman in the text. Wisdom literature, such as proverbs and fables (*e.g.*, "Dispute between the Date Palm and the Tamarisk"), poetic meditations, oracles, divination records, omens, and prophecies are further examples of Mesopotamian genres that only epigraphy has preserved.

Sumerian and Akkadian narrative literature is likewise of wholly inscriptional transmission. It contains man's earliest preserved literary creations in the Sumerian sequence, especially the texts from tablets found at Nippur. These include the "Paradise myth" of the god Enki and the goddess Ninhursag in the pure, clean, and bright land of Dilmun; the story of Dumuzi and Enkimdu (the petulant shepherd god versus the peace-loving farmer god, inversely reminiscent of the Cain–Abel antagonism in Genesis but not culminating in murder); "The Deluge" with its Noah-hero Ziusudra; "Inanna's Descent to the Nether World," which prefigures the later Akkadian "Ishtar's Descent"; and the lays of Gilgamesh, which show the Sumerian traditions that were later partly organized and transformed into the Akkadian epic. The latter include "Gilgamesh and Agga of Kish," a story of confrontation between early Sumerian city-states; "Gilgamesh and the Land of the Living"; and "The Death of Gilgamesh," with its haunting parallelistic refrain "he lies, he rises not." The Akkadian *Epic of Gil-*

The Gilgamesh epic

gamesh is divided into 12 tablets, the longest of which is more than 300 lines; this "Flood Tablet" (the 11th) is virtually intact and comes, like almost all Assyrian-language Gilgamesh texts, from the library of Ashurbanipal at Nineveh (7th century BC). From the 2nd millennium there are fragments of a Hittite version from Boğazköy, as well as minor traces of a Hurrian translation. Old Babylonian correspondences to tablets 1–3 and 10 are found on a tablet from Sippar (*c.* 1800 BC). The 12th tablet is a literal translation from Sumerian, whereas the rest amounts to a self-contained Akkadian epic original, based on Sumerian motifs but with a thrust of its own. The most complete reconstruction involves a combination of Assyrian, Old Babylonian, and Hittite versions.

The other famous Mesopotamian epic, *Enuma elish*, "When on high," details the story of cosmic creation and of how Marduk became the great god of Babylon; it had more immediate cultic attachments because its recitation formed part of the New Year festival.

Further Akkadian literary creation is attested in the epic of Atrahasis, a tale of mankind's punishment through pestilence and flood, preserved in fragmentary Old Babylonian and Assyrian versions. The story of Adapa, found in parts in the Tell el-Amarna archives and the library of Ashurbanipal, is similar to Gilgamesh's quest for immortality. The myth of Zu deals with the theft of the tables of fate and the usurpation of almightiness by the bird god Zu. The legend of Etana, a namesake of the shepherd-king who ascended to heaven in the mythical postdiluvian Sumerian dynasty of Kish, recounts in its Old Babylonian and Assyrian recensions the heavenly flight of Etana on the wings of an eagle in order to acquire the magic birth plant that would cure his childlessness. Death-oriented themes appear in the tale of Ishtar's descent, in the story of Nergal and Ereshkigal, and in various netherworld texts associated with the Tammuz myth and liturgy.

The accidental preservation of Egyptian texts

Ancient Egypt. The multitude of incidental day-to-day written matter on potsherds and papyrus, preserved by climatological quirks in Egypt, affords insights into the normal living patterns of ordinary people during the pharaonic, the Hellenistic (Ptolemaic), and the Roman periods. More formal records of similar provenance may be described as legal in the widest sense, comprising such objects as land deeds, cadastral inventories (tax surveys), wills, adoption decrees, and trial transcripts. Thus many technicalities of landholding in the Nile Delta during certain ancient periods survive, unwarranted by any normal expectancies of epigraphic preservation elsewhere; only the happenstance of the Creto-Mycenaean records (see below *Crete and Mycenaean Greece*) has accidentally transmitted something analogous.

Ritualistic texts. Ritualistic texts are equally abundant, in both monumental and papyrological transmission. Notable among offertory liturgy was the mortuary service to the dead, acted out as a kind of ritual make-believe in which the offering was referred to as the eye of Horus (the sacrificer) being given up on behalf of his father, Osiris (the departed). Ritual drama in general is amply attested in Egyptian religious practice; day-to-day ceremonial in the cult of gods or god-kings (pharaohs) was recorded in the minutest detail. Magic texts include incantations and charms against such dangerous creatures as snakes, scorpions, crocodiles, and lions. Even the buried dead were deemed in need of such protection, and hence the preserved samples are largely of mortuary provenance. An important category of epigraphs is constituted by curses, imprecations, and threats of divine vengeance against enemies, evildoers, and tomb violators in particular. Dream interpretation and oracular inquiry are represented on both monuments and papyri.

Moral and didactic writings. Of a more secular nature and on the verge of true literature are moral and didactic writings, particularly during the early Middle Kingdom (began *c.* 2133 BC), when a profound social and spiritual crisis seems to have gripped Egypt. Of such kind are "The Admonitions of Ipuwer" (a denunciation of current sin and evil in Hebrew "prophetic" manner), the "Colloquy of a Prospective Suicide with His Own Soul," an especially copious compendium of conventional wisdom, frequently formulated as instructions from a royal or noble father to his son. Some parts offer remarkable parallels to the Hebrew book of Proverbs.

Mythical tales. On the same borderline to real literature stand mythical tales, including in particular the myths of creation by the god Atum of Heliopolis and the Memphite patron deity Ptah, whose eminence is reminiscent of the cultic fortunes of Marduk. The Memphite cosmogony, of 1st dynasty provenance, is an abstract one, with Ptah's creation conceived with his "heart" (mind) and "tongue" (word) rather than the usual physical detail. Other mythic tales concern the perils from the netherworld demon Apepi faced by the sun god Re during his nightly journey in his solar boat; the rivalries between Horus and Seth for the succession of Osiris (a patent allegory of the tensions between Lower and Upper Egypt); mortuary texts particularly concerned with the Osiris myth and the prospect of judgment (*e.g.*, the "Protestations of Innocence"), attested both monumentally and on papyri (the Book of the Dead); and pieces designed to bolster the supremacy of specific sites and deities, particularly Amon-Re at Thebes.

Poetry and prose. Out of the latter tendency emerged hymnic poetry during the 18th and 19th dynasties, triggered by an imperial trend toward unicentrism in government and universal syncretism in religion. The great hymns to Amon-Re were matched during Ikhnaton's "Amarna revolution" by his great Aton-hymn, uncannily reminiscent of the 104th Psalm of the Old Testament. Victory hymns and prayers are further examples of poetic genre, as are purely secular pieces, including several collections of love songs from *c.* 1300–1100 BC, songs of farmworkers and herdsmen, and feast songs for entertainment.

In nonreligious literary prose there is preserved the story of Si-nuhe, an account of the life of a Middle Kingdom official, his prosperous peregrination in Asia but continued yearning for the homeland, relieved at last by a chance to join the pharaonic court. Papyri and ostraca tell this story in a long succession of attestations through most of the 2nd millennium BC. Of later date (*c.* 1100 BC) is the "picaresque" narrative of the hieratic-script "Moscow papyrus," detailing the ill-financed mission of a certain Wen-Amon from Karnak to Byblos to buy timber for a ceremonial barge of Amon.

The Boğazköy archives

The Hittite Empire. The Boğazköy archives are the unique central storehouse of Hittite records for the duration of that empire (only minor additions have been found elsewhere, such as letters at Tell el-Amarna and Alalakh). The past of other cultures was known from external and nonepigraphic sources before the explorational and excavational discoveries of inscriptions; with the Hittites, however, the record is in one place and *in toto*.

Much may be learned of everyday life and social relations from legal texts, including the laws themselves. Such matters as family relations, property rights, land tenure, and commodity prices are dealt with in great and sometimes pedantic or amusing detail. More can be learned from gift deeds of land, cadastral lists, and trial records, the latter containing largely depositions of respondents and witnesses in civil suits. Unique in their kind are treatises on the training of racehorses. Astrological omen texts, divination records (hepatoscopy, auspicy, lottery oracles), and dream interpretations abound, largely patterned on Mesopotamian models; rituals of all kinds take up a disproportionate amount of space. Temple foundations and procedures, festival ceremonies, epidemics, impotence, family quarrels, countermagic, royal welfare, birth, insomnia, and the manipulation of deities are among the topics of Hittite ritual; in a class by itself is a funeral ritual for royalty, reminiscent of Homeric practices. Not only the Hittite language but Hattic, Hurrian, Luwian, and Palaic are used in rituals. Another large category is made up of hymns and prayers, most notably those of King Mursilis II, in particular his "plague prayers," in which he confesses sins and begs the storm god for mercy. Other notable examples are the prayer of Muwatallis to the sun god, the vows of Queen

Puduhepa to the sun goddess, and the prayer of Kantuzilis with its intimations of mortality. Native Anatolian mythical texts, such as those of the slaying of the dragon Illuyankas and of the disappearance of the god Telipinus, were part of ritual recitation. Other mythical matter points up the crossroads character of Hittite civilization, being made up of Akkadian, Hurrian, and Canaanite themes. Most of the Hittite epigraphic corpus has some connection with religion.

Other countries of the ancient Near East. Away from the big power centres some quite important sites remain to be identified, such as the Mitannian capital of Wassukkani. Others (*e.g.*, the fortress-town of Carchemish on the upper Euphrates, for some time a Hittite dependency) have yielded notable data, especially royal inscriptions in Hittite Hieroglyphs. In particular the French excavations at Ras Shamra on the Syrian coast since 1929 have uncovered the inscriptional and other remains of the small but strategic city-state of Ugarit, which flourished in the 15th–13th centuries BC. Its own vowelless cuneiform alphabet, a radical departure from the Mesopotamian syllabary prototype, denotes an archaic Semitic dialect closely akin to Canaanite. The written documentation of this crossroads community, in a variety of scripts and languages, provides information on the history of the area where Hittite and Egyptian power politics collided. But its most significant epigraphic products have been poetic epics of mythical and heroic scope, which go far to elucidate religious traditions otherwise known mainly from Old Testament bias. The great cycle of Baal and Anath brings to life the pantheon to which belonged also El, Asherah (the Astarte of the Phoenicians), Kothar ("Deft," the craftsman god), and Yamm (the sea god). The tale of Aqhat is also on the borderline of myth, while that of King Keret (or Kret) takes place in a saga world somewhat reminiscent of the Greek Homeric tradition.

The city-state of Ugarit

Scattered dedicatory inscriptions and papyrus texts shed some light on the lives and practices of the Aramaean-speaking populations during the 1st millennium BC. From Palestine there are the Hebrew ostraca of Samaria, datable to the reign of Jeroboam II of Israel (8th century BC), which record names, families, and administrative and religious practices. Of equal significance are the ostraca of Lachish in southern Palestine, which probably immediately preceded the Chaldean onslaught of 589 BC. Phoenician texts are scattered around the Mediterranean, and bear witness to an extensive and protracted maritime supremacy.

The Old Persian Achaemenid inscriptions (see above *Ancient Iran*) are of some value in assessing the state religion of the time, which seems to have been a rather anemic form of official Zoroastrianism. Later monuments from the wider Iranian area help map its complicated religious history, such as the great inscription of the Kuṣāṇa king Kaniṣka, found at Surkh-Kotal in Afghanistan in 1957 and attesting to the Iranian language and Mithraic cults of ancient Bactria in the 2nd century AD. The Sāsānian religious tradition of the 3rd–7th centuries, with its rigidified formal Zoroastrianism, is mainly of nonepigraphic attestation.

Crete and Mycenaean Greece. The decipherment of the latest and most copiously attested of the Minoan linear scripts, the so-called Linear B, by British cryptologist Michael Ventris in 1953, is a major example of the dramatic impact epigraphic discovery can have on the most varied antiquarian disciplines. It supplied incontrovertible proof that the Mycenaeans on the Greek mainland during the 2nd millennium BC were Greek in language and likewise that Knossos in Crete was a Greek-speaking stronghold at the time of its final destruction (*c.* 1400 BC). The records from Knossos, Pylos in Messenia, and Mycenae in the Argolid are exclusively perishable inventories on clay tablets, kept in royal palaces or emporia and apparently meant to be discarded or reused annually; thus they record data from the very last year of any given establishment, just before its final destruction in the fire that also accidentally ensured their preservation by baking the clay.

The Linear B script

The contents are brief, concise, practical, and quantifying in character. Nevertheless, they reveal an important amount of detail about a civilization that is otherwise accessible only via the testimony of archaeology, iconography, and the poetic memory of early classical Greece across several "dark" centuries. Apart from the philologically invaluable recording of preclassical linguistic forms and proper names, there is the cryptologic evidence of the syllabic writing system, which may yet help bring about a cogent decipherment and linguistic identification of the earlier forms of Cretan writing.

Something of bureaucratic accounting methods can be learned from the system of the tablets. Contents yield lists of personnel; of livestock and agricultural produce; of textiles, vessels, and furniture; of metals and military matériel (swords, corselets, chariots, etc.); as well as records of landholding and of temporal and cultic tribute. The cadastral inventories are significant for the light they shed on the system of land distribution as compared with classical Greece and especially with contemporary practices in adjacent culture areas, such as those recorded in the Hittite Law Code. The religious tributes are revelatory in proving what bold scholarship had merely surmised—that the bulk of the classical Greek deities existed in name and kind already in Mycenaean days. Something of the social order and the system of government can also be incidentally learned.

Classical Greece. Inscriptions in their variety and profusion are an important means of gauging everyday life in classical Greece, especially such more formal aspects as were deserving of recording. This qualification, however, does not prevent some obscene pederastic rock-carvings on the island of Thera from being the earliest epigraphic texts from the classical period (*c.* 8th century BC), in a form of the alphabet still strongly marked by its Phoenician source.

Earliest classical epigraphic texts

Business, financial, and legal texts. In a formal vein there is ample documentation of a business, financial, and legal type, such as regulations for building projects, contracts for medical services (*e.g.*, the Cypriot bronze tablet of Idalium, in a peculiar syllabaric script reminiscent of the Cretan and Cypro-Minoan types, used to write both Greek and the uninterpreted Eteocypriot language), collection agreements for defaulted notes, manumission decrees, records of bank deposits (frequently in a temple), and reports of land commissions (especially the Heraclean Tables from southern Italy).

Religious texts. Even in these formal documents it is impossible to keep out reference to religious establishments. Religion was, in fact, big business in classical Greece, and overlap in the records is frequent. Temples were often of the order of corporations, with extensive holdings of real estate and banking operations. The epigraphs of their transactions are coupled with minute records of dedicatory offerings, priest lists, administrative regulations, and ritualistic and liturgical instructions. So much of social life was centred around institutionalized religion that most of its documentation may be included under the latter heading. Dedicatory texts in the widest sense are common, often in verse, such as that by Nikandre of Naxos on a 7th-century-BC statue of Artemis at Delos, a kind of propitiatory offering to the maiden goddess on the occasion of marriage. Similar hexametric or elegiac texts are found as epitaphs, especially for those who perished in war or at sea.

Death-related inscriptions. Tomb inscriptions generally are the most numerous of all but tend to be laconic in style and lacking even in purely statistical data. Those from ancillary areas of the Greek-speaking world, such as the non-Greek ones from Lycia, Lydia, and Phrygia in Asia Minor, are sometimes more revealing to the extent that they are interpreted. Thus the Lycian stela of Xanthus, a tomb monument from the 5th century BC, runs to many hundreds of words, including a dozen lines of Greek. The peculiar Lycian system of matrilinear descent is clearly evident in the texts. The Lydian ones include a marble stela from the necropolis of Sardis with a Lydian-Aramaic bilingual epitaph that typically calls down the curse of Artemis of Ephesus on potential viola-

Tomb monuments

tors. Such imprecations are also standard in Phrygian sepulchral epigraphs. Some kinds of death-related inscriptions were intentionally buried, such as the curses that consigned the object to the infernal avengers. Another type is seen in the Orphic tablets—texts on thin gold found interred with the remains of devotees of the Orphic salvation religion near Sybaris in southern Italy, near Rome, and at Eleutherna in Crete. They are apparently meant as instructions to the deceased—exhortations to the soul on its progress in the beyond, enjoining the shunning of the spring of Lethe ("Oblivion") and the drinking from that of Mnemosyne ("Memory"), thus securing the end of transmigration of the soul and the attainment of perpetual higher consciousness.

Epigraphy and Greek literature. The Orphic texts provide the popular parallel to the Platonic myth of Er in the 10th book of *The Republic*, where Plato somewhat recast Lethe and Mnemosyne into Ameles ("Unmindfulness") and Anamnesis ("Recollection"). Thus, epigraphy not merely supplements the literary record of Greek civilization but also complements it in important aspects. If papyrology is included, important additions to the preserved portion of classical Greek literature have come from Egypt, especially in the papyri from Oxyrhynchus; these include such otherwise lost treatises as Aristotle's *Constitution of Athens* and entire dramatic works (*e.g.*, the *Dyscolus* of Menander).

Ancient Rome. The frequency of Roman inscriptions increased dramatically in direct proportion to the rise of Roman power; but that same rise brought centralization, stereotyping, and a certain sterility of the more formal parts of the epigraphic record. Much of the bulk is official, unidirectional, and from the top outward. The Greek-speaking parts of the empire continued in many of their ancient ways, and other annexed regions (such as Spain or Gaul) developed their own locally coloured practices. The spread of a variety of exotic religious cults all over the empire, among them Mithraism and Christianity, added a kind of epigraphic underground initially devoid of official sanction and largely unmatched by alternative avenues of preserved written transmission. Popular epigraphy, including such matter as graffiti at Pompeii and other Vulgar Latin inscriptions, provides further counterpoise to the official stereotypes.

From early republican days the Roman written record is very spotty. The earliest text of any length, the Forum inscription from the early 5th century BC, seems to refer to an augural rite (referred to by Cicero and Festus as the *juge auspicium*) which enjoins the immediate unyoking of beasts of burden that produce excrement. A 4th-century vase (the "Quirinal vase") is apparently a wedding present with an inscribed guarantee by the bride's guardian to use his good offices to further marital concord. Later texts of religious import include the "Song of the Arval Brethren" on a marble tablet found in the Vatican in 1778 (a chant largely reduced to gibberish by age-old ritualistic repetition), records of various other cultic "colleges" (*augures; fetiales,* or priestly diplomatic representatives; *salii,* or priests of Mars and Quirinus et al.), and outside of the Latin-speech area especially the seven Umbrian-language bronze tables found at Iguvium (modern Gubbio) in 1444, recording in more than 4,000 words the ritualistic details of a brotherhood (the Fratres Atiedii) that flourished in republican days. Official religion is further seen in many dedications, such as one found on Caelian Hill in Rome in the 18th century, which records the offering of a temple and statue to Hercules by Lucius Mummius in 146 BC in fulfillment of a vow he had made during his conquest of Greece and sack of Corinth. "Unofficial" cults, especially those of Mithra and Jupiter Dolichenus, can be traced and mapped throughout the empire chiefly with the help of dedicatory epigraphs. Sepulchral inscriptions are recorded in overwhelming numbers, not only in Latin but notably also in Etruscan; among the latter is the mummy wrap, a lengthy funeral liturgy inscribed on an apparently Etruscan mummy entombed in Egypt (now in Zagreb). The earliest notable Roman examples are the epitaphs in Saturnian verse on the sarcophagi of the Scipios from the 3rd century BC.

There are many texts of deep sentiment and poetic feeling. Yet the bulk itself is the main source of information because it affords what may be called demographic statistics, including population distribution, occupations, public health, and longevity. The frequent inclusion of the *cursus honorum*, or career summary of the departed, has similar informational value. Tombs were usually placed under the divine tutelage of the powers of the beyond. Imprecations against violation were common, as were buried curse tablets in Latin and in Oscan (the *defixiones*). Many of the unofficial inscriptions exhibit substandard degrees of literacy and colloquial features, being thus an important adjunct to the study of the totality of classical Latin and its subsequent developments.

THE USE OF INSCRIPTIONS

The dating of historical events. Inscriptions are important specimens for chronology because they are often physical objects contemporary in execution with their contents. The dating of the inscription itself frequently yields a trustworthy chronology of its message: a victory stela records something freshly deserving of celebration; an epitaph implies a recent death. Exceptions do exist, which record more or less remote events at a conscious historical remove; archival specimens, for example, and secondhand copies generally lack the contemporaneity of other inscriptions. On the whole, however, external dating is crucial and may be achieved in several ways. Excavated monuments can be chronologized by their archeological context, including stratigraphic analysis and radiocarbon dating of any adjacent remains of organic matter. The shape of the monument may permit stylistic and iconographic determination. The type and variety of script used, and especially the style of writing, often allow paleographic dating. Thus, the relative age of Hittite texts can be determined by spotting the typical "Old Hittite ductus" of the more ancient period, and the various "scribal hands" of the Linear B tablets have been differentiated with extreme subtlety. Sometimes a radical reform, such as the official adoption by Athens of the Ionic alphabet in 403 BC (replacing the local Attic variety), provides a chronological watershed. Internal evidence of the inscription may yield its own kind of dating, either by synchronism with otherwise known facts or events or in true calendaric fashion. The year is frequently indicated by a king's reign or the tenure of a magistrate. Such stable counting of time as, for example, from Rome's legendary founding in 753 BC, or from creation (5509 BC, according to Christian dating) is rare in inscriptions; it became more of a historian's device during the classical and postclassical periods. In smaller communities, however, especially in Asia Minor, analogous local departures were used (legendary or historical foundation dates or other epoch-making events), with confusing results for latter-day chronologists.

Contemporaneity of epigraphy

The use of inscriptions for the dating of historical events is most pervasive when the historical tradition itself is "timeless," as in ancient India; the entire Indian chronology comes to be anchored around the Aśokan inscriptions. Inscriptions also permit a check on the veracity of ancient historians such as Herodotus (dubbed both "father of history" and "father of lies"), as in the case of the Bīsitūn inscription of Darius. Equally dramatically, the Linear B tablets prove at one stroke that the Greeks were ensconced at Knossos in the 2nd millennium BC and that the bulk of the "Olympian" religion was already theirs at the time. Most significant of all, nothing would be known of the great Hittite Empire during the 2nd millennium BC, were it not for the discovery of its inscriptional archives.

The decipherment of ancient languages. Inscriptions as written records are usable only in proportion to their intelligibility. Important epigraphic corpora remain virtually undeciphered; *e.g.*, the "Indus script" from Mohenjo-daro and Harappā (3rd millennium BC), the Carian texts from Asia Minor and by Carian mercenaries in Egypt (1st millennium BC), and the pictographic Mayan "hieroglyphs" from Central America (*c.* AD 500–1500). Sometimes the writing system is intelligible, as in the

Undeciphered epigraphic corpora

case of Etruscan, but understanding remains deficient because the language is otherwise unknown and bilingual keys are lacking or inadequate. Chances for success are best if there are sufficiently extensive bilingual or multilingual copies, of which at least one language is previously understood. Such presence made possible the decipherment of ancient Egyptian (the Rosetta Stone of 196 BC, with hieroglyphic, demotic, and Greek versions). Once the affinity of an underlying language to known idioms is established (*e.g.*, Old Egyptian to Coptic, Old Persian to Avestan and Sanskrit, Akkadian to Hebrew), interpretation can proceed apace. The recovery of Hittite was not a true decipherment because the script was a relatively common variety of syllabic cuneiform. The interpretation was helped by the nature of the writing on the one hand (including intelligible ideograms, while an alphabet yields no such clues), and by the presence of Akkadian-Hittite bilinguals on the other; the soon-recognized Indo-European affinities of the Hittite language afforded further help. The Hittite Hieroglyphs were partly deciphered by painstaking internal analysis based on the correct assumption of an underlying dialect akin to Hittite; a bilingual with Phoenician text brought much welcome confirmation. The decipherment of Linear B was a sheer triumph of methodical cryptology, again based on the correct hunch that the hidden language was Greek.

In sum, the decipherment of ancient scripts and the recovery of lost languages are practically identical with the interpretation of previously unintelligible ancient inscriptions, because texts involved are almost exclusively epigraphic. Even when the language is otherwise preserved (as with classical Greek and Latin), inscriptions yield essential additional data for its history, dialects, and social diversification.

THE HISTORY OF EPIGRAPHY

Greek and Latin inscriptions. Inscriptions have commonly elicited the curiosity of posterity, and such ancient Greek historians as Thucydides and Polybius already made scholarly use of them. Sporadic systematic interest in Greek and Latin inscriptions is attested in later ages; *e.g.*, Cola di Rienzo in the 14th century made a collection, and Cyriacus of Ancona (Ciriaco de' Pizzicolli) in the 15th century was a renowned recorder of ancient written monuments on his mercantile travels to Greece, Anatolia, and Egypt. Cyriacus' material formed the nucleus of various compilations in the succeeding centuries, normally on a geographic basis. A rival typological method of publication was launched by Martin Smetius in Leiden in the late 16th century and was followed in the early 17th by a monumental collection of Janus Gruter and Joseph Justus Scaliger. After copious additions of material during the 18th and early 19th centuries, the *Corpus Inscriptionum Graecarum* was launched by August Böckh in 1815 under the aegis of the Berlin Academy and was completed in four volumes (1828–59). The material had by then again outrun the publication, and it was resolved in 1868 to re-edit completely all Attic inscriptions. After the completion of the new three-volume *Corpus Inscriptionum Atticarum*, U. von Wilamowitz-Moellendorff took charge in 1902 and settled upon a new corpus called *Inscriptiones Graecae*, which encompassed all Greek inscriptions from European Greece (including Magna Graecia in Italy) and Cyprus. Those of Anatolia were left to the *Tituli Asiae Minoris* of the Vienna Academy, which began with the Lycian-language inscriptions from Lycia in 1901 and continued with the Greek and Latin ones from Lycia in 1920–44. The rest of Greek Anatolia was combed somewhat by the multivolume American series, *Monumenta Asiae Minoris Antiqua* (since 1928). *Inscriptiones Graecae*, framed in 14 volumes, has turned partly into a kind of overall umbrella for diverse coverage; volumes 6, 8, 10, much of 11, parts of 12, and 13 were never completed, being pre-empted by such other large publications as *Inschriften von Olympia*, *Fouilles de Delphes*, V. Latyshev's *Inscriptiones Antiquae Orae Septentrionalis Ponti Euxini Graecae et Latinae* and G. Mikhaylov's *Inscriptiones Graecae in Bulgaria Repertae*, *Inscriptions de Délos*, and M. Guarducci's *Inscriptiones Creticae*. The continuous influx of new material is inventoried and presented in such annual outlets as *Supplementum Epigraphicum Graecum* (since 1923) and "Bulletin épigraphique" in the *Revue des études grecques*. Large recent undertakings either supplement or complement the *Inscriptiones Graecae*, such as *The Athenian Tribute Lists* (from the American excavations in the Agora) and *Inscriptions grecques et latines de la Syrie* (six volumes, 1929–67). There are, in addition, numerous more localized or specialized collections.

Epigraphic collections

The *Corpus Inscriptionum Latinarum*, founded by Th. Mommsen in Berlin in 1862, comprises 16 volumes fortified with copious supplements; it covers the classical Latin world and is notable for its homogeneity and systematization. Current finds and studies are presented in *L'année épigraphique* (since 1888). Etruscan inscriptions are less perfectly published in the antiquated *Corpus Inscriptionum Etruscarum;* many important later finds are included in transcription in M. Pallottino's *Testimonia Linguae Etruscae* and in M. Fowler and R.G. Wolfe, *Materials for the Study of the Etruscan Language* (1955; a computerized corpus). Of further relevance to the Roman world are collections of the type of E. Diehl, *Inscriptiones Latinae Christianae Veteres;* M.J. Vermaseren, *Corpus Inscriptionum et Monumentorum Religionis Mithriacae;* and *Inscriptiones Italiae.*

Other inscriptions. R.A.S. Macalister's *Corpus Inscriptionum Insularum Celticarum* (1945–49) gathers the oghamic and other early texts from Ireland and elsewhere. The runic inscriptions are inventoried in a variety of compilations. The *Corpus Inscriptionum Semiticarum*, in Paris (since 1881), covers in separate volumes Phoenician, Aramaic, and other speech areas. Urartean texts were collected by C.F. Lehmann-Haupt in *Corpus Inscriptionum Chaldicarum* (1928–35); the earlier found Hieroglyphic Hittite texts, by L. Messerschmidt in the antiquated *Corpus Inscriptionum Hettiticarum*. The *Corpus Inscriptionum Iranicarum*, intended to gather the epigraphs of Persia proper (Achaemenid, Seleucid, Parthian, Sāsānid) and of eastern Iran and Central Asia, began in London in 1955. The *Corpus Inscriptionum Indicarum* has published four volumes since the 1870s, comprising the Aśoka, Indo-Scythian, Gupta, and Kalacuri-Cedi periods, supplemented by the series *Epigraphia Indica* and *South Inaian Inscriptions.*

No equally closed corpora exist for cuneiform and Egyptian documents. The Hittite texts come closest to such organization, issued mainly in two great ongoing series, *Keilschrifttexte aus Boghazköi* (German Oriental Society) and *Keilschrifturkunden aus Boghazköi* (Berlin Academy). But mostly the publication is haphazard—as part of excavation records (*Keilschrifttexte aus Assur*), depending on storage sites (*Cuneiform Texts* in the British Museum), or on the basis of genres (F. Thureau-Dangin, *Rituels accadiens*). The same pattern holds true of the Egyptian records.

Much epigraphic material is published in excavation reports and in many archeological and antiquarian periodicals; *e.g.*, *Hesperia*, *Journal of Hellenic Studies*, *Bulletin de correspondance hellénique*, *Archeologia classica*, *Studi Etruschi*, and *Syria*. Specifically epigraphic serial publications include *Epigraphica* (Milan, since 1939), *Kadmos* (Berlin, since 1962), *Zeitschrift für Papyrologie und Epigraphik* (Bonn, since 1967), and *Chiron* (Munich, since 1971).

BIBLIOGRAPHY. J.B. PRITCHARD (ed.), *Ancient Near Eastern Texts Relating to the Old Testament*, 3rd ed. (1969), gives translations of inscriptional material, especially from Egypt, Mesopotamia, Palestine, and Asia Minor. J. FRIEDRICH, *Entzifferung verschollener Schriften und Sprachen* (1954; Eng. trans., *Extinct Languages*, 1957), deals especially with the decipherment of inscriptions, while his *Kleinasiatische Sprachdenkmäler* (1932), transliterates a variety of epichoric texts from Asia Minor. R.G. KENT, *Old Persian*, 2nd ed. rev. (1953), is a comprehensive work including grammar, texts, and lexicon. W. HINZ, *Altiranische Funde und Forschungen* (1969), includes a number of important recent inscriptional finds from ancient Iran. J. BLOCH, *Les Inscriptions d'Aśoka* (1950), gives texts and translations of all Aśokan epigraphs; whereas

R.B. PANDEY, *Historical and Literary Inscriptions* (1962), provides a transliterated selection of later Indic epigraphy as well. L. DEROY, *Initiation à l'épigraphie mycénienne* (1962), provides a comprehensive survey of the study of Mycenaean tablets. E.S. ROBERTS, *An Introduction to Greek Epigraphy*, 2 vol. (1887–1905), remains the classic compendium of its kind; while more recent, shorter surveys of its subject are found in G. KLAFFENBACH, *Griechische Epigraphik* (1957); and A.G. WOODHEAD, *The Study of Greek Inscriptions* (1959). A more elaborate recent treatment is that of M. GUARDUCCI, *Epigrafia greca*, 2 vol. (1967–69). The Cretan laws have been sumptuously edited and commented in R.F. WILLETTS, "The Law Code of Gortyn," *Kadmos*, suppl. 1 (1967). The old standard works on Latin inscriptions, J.E. SANDYS, *Latin Epigraphy*, 2nd ed. rev. (1927); and R.L.V. CAGNAT, *Cours d'épigraphie latine*, 3rd ed. (1898), are updated only in such summary surveys as R. BLOCH, *L'épigraphie latine*, 3rd ed. (1964). The elaborate edition of the principal Umbrian inscription by J.W. POULTNEY, *The Bronze Tables of Iguvium* (1959), gives text, translation, grammar, extensive commentary, and facsimiles of the tables themselves. R.W.V. ELLIOTT, *Runes; An Introduction* (1959), is especially thorough on the side of British runes; while W. KRAUSE, *Runen* (1970), is a compact but comprehensive survey of the entire field of runology.

(J.Pl.)

Epistemology

Epistemology (Greek *epistēmē*, "knowledge") is here understood as the study of the nature and validity of knowledge. Typically, epistemologists examine the degrees of certainty and probability and the differences between knowing (with certainty) and believing (without being certain). Their concern is primarily with human knowledge. It is true that, on observing the behaviour of animals, birds, and insects, they note certain points of comparison with human behaviour that may be held to illuminate and justify the view that animals can be said to know; and they may conceive of divine knowledge, too, as comparable to human knowledge but greatly transcending it—as perfect omniscience free from all the imperfections of human knowledge. But both inquiries depend upon a prior acquaintance with human knowledge.

It is not the intention of this article to present a chronological study of theory of knowledge throughout the ages. Though, in the course of the argument, the main teachings of the leaders in this field will be discussed, they will not be presented chronologically; the interest lies in concept and theory rather than in history.

Nor is it the purpose of the article to consider what is usually termed the philosophy of science (though this is what "epistemology" sometimes connotes in Continental philosophy). In all inquiries, including the sciences, there are common epistemological problems; but there are also specific problems pertaining to each of the different sciences, for instance, problems about methodology. These are not the concern of this article but are the task of the specialists in each particular field—in biology, physics, geology, and so on—or of the philosopher of science, whose prerogative it is to provide a catalog of methodologies and to discuss the specific problems of the sciences (see SCIENCE, PHILOSOPHY OF). The present article deals, instead, with what in Europe is sometimes called gnosiology.

The following subsection of the article will be followed by a subsection on the relations between epistemology and certain other disciplines, such as psychology and logic. In the first main section a brief sketch will be given of the points at issue in *contemporary* epistemological theories. The argument will be confined to theories; that is to say, individual authors will not be discussed. It will present the state of play in this field of inquiry at the present moment. In the suceeding sections, the main topics of epistemological theory set forward by past and present philosophers will be dealt with. In the sixth section some concluding reflections will be given.

The main topics are thus outlined as follows:

I. Function, scope, and relations of epistemology
  Epistemology as a discipline
  Relation of epistemology to other disciplines
    Relation to metaphysics
    Relation to psychology
    Relation to philosophy of mind
    Relation to logic
    Relation to other studies
II. Diversities of contemporary epistemological theory
  Sense perception
    Sense-datum theory
    Phenomenalism
    Perceptual assurance of physical objects
  Levels of assurance
    Quest for certainty
    Belief
  Mental activity
    Thinking and language
    Inference
    Reason and the a priori
III. Rationalist orientation
  Rationalism prior to Kant
    Classical Rationalism
    Rationalist elements in Scholasticism
    17th-century Rationalism
  Critical Rationalism of Kant
  Post-Kantian Rationalism
    Rationalism of absolute Idealism
    Realist Rationalism
  Attempts to transcend reason
IV. Empiricist orientation
  Ancient Greek and medieval Empiricism
    Classical philosophers
    Medieval philosophers
  Modern and recent Empiricism
    Early forerunners
    Classical British Empiricists
    Assessments and recent expressions of Empiricism
V. Conceptual thinking
  Concepts
  Language
  Universals
VI. Concluding reflections
  Conflict of Empiricism with Rationalism
  Blend of Empiricism with Rationalism

I. Function, scope, and relations of epistemology

EPISTEMOLOGY AS A DISCIPLINE

The need for knowledge about knowledge

If "the proper study of mankind is man," then epistemology or theory of knowledge is a worthy part of that study, for it deals with an aspect of human life that is of fundamental importance. To act wisely in the world, it is necessary to know that world and understand it; this is not to say, however, that, if a man knows, he will inevitably use the knowledge wisely. It is true that the past several centuries have witnessed rapid advances in knowledge that have made possible higher standards of life; but it is equally true that this very knowledge has frequently been misused, increasing man's misery. Moreover, the new knowledge has sometimes provoked a sense of insecurity, even a certain loss of morale. The universe has been shown to be incomparably vaster and man's role in it less significant than had been assumed. Advances in astronomy, nuclear physics, chemistry, and biology have in many respects been disturbing. Yet, paradoxically, men are proud of these achievements. If the world seems now less cozy than it used to be, if its very vastness is forbidding, it is man himself who has discovered the truth about the world—if it is the truth. Indeed, man is also uneasy about his liability to error. Men assert confidently what they take to be the truth, only to find later that their assertions frequently have been false—hence man's great need not only to know the world around him but also to know himself better and, in particular, the need to understand the character and reliability of his own cognitive powers. The epistemologist seeks to provide an urgently needed knowledge about knowledge.

RELATION OF EPISTEMOLOGY TO OTHER DISCIPLINES

The disciplines most closely related to epistemology are metaphysics, psychology, philosophy of mind, and logic.

Relation to metaphysics. Traditionally, metaphysics (the study of the nature of Being) was held to have three parts—ontology, cosmology, and epistemology—and was ranked as the highest of the university secular studies. Today, because of the influence of Positivist and Empiricist teachings, the tendency is to deny it any place whatsoever in the curriculum, on the grounds that what it at-

tempts to study transcends human comprehension and is not susceptible of philosophical analysis. But epistemology, though it is a part of metaphysics, has not suffered the same fate. Recently there have been signs of a possible partial restoration of metaphysics, and because the studies are concerned with the problem of the general structure of the world, the findings of metaphysicians may become a matter of interest to epistemologists.

The present tendency to reject metaphysics

Relation to psychology. Fifty years ago, psychology and epistemology were closely related branches of philosophy, and, in fact, insofar as philosophy was concerned with the *genesis* of knowledge, there was no distinction between them. The part concerned with the *validation* of knowledge was linked with logic; but logic was not as sharply separated from psychology as it is today. A century ago the Empiricist John Stuart Mill spoke of logic as "not a science distinct from . . . psychology. . . . Its theoretic grounds are wholly borrowed from psychology." In the early years of the 20th century, however, psychology broke away from philosophy, rejected its methods, and became an empirical discipline. Epistemologists, on the other hand, continued to use, with greater precision, the methods of analysis. They subjected their key concepts—know, believe, see, infer, and true—to close analysis, with useful results. The trend of psychology and epistemology away from each other ignores the fact that there are points at which they could still help each other: the epistemologist's analysis of concepts could help psychologists in the study of cognition. Conversely, the epistemologist would benefit from examining the work of psychologists on such things as the beginnings of knowledge of number, of movement, and of reality in young children, and it would be advantageous if various epistemological theories of sense perception were to show an awareness of, for example, psychological studies of perceptual systems and the role of accumulated past experience in conditioning present perception.

Psychology's break with philosophy

Relation to philosophy of mind. Philosophy of mind is a comparatively new discipline whose boundaries at present are loosely drawn. In the widest interpretation it would presumably include epistemology as one branch of its total subject matter, for an inquiry into the whole of mental life would cover the study of the theoretical or cognitive aspect of mind. In fact, philosophy of mind tends to leave the study of the nature and validity of knowledge to the epistemologist and concentrates on other issues: first, on the nature of mind and its relation to body; second, on the distinction between the self and other selves; and third, on the description of the various aspects of the mind's life. In his study of the cognitive aspect the philosopher of mind leans on the work of the epistemologist and finds it necessary to familiarize himself with the findings of epistemology and psychology. But the philosopher of mind is necessarily concerned with practice as well as with theory—*i.e.*, with willing and acting—and he is concerned too with man's emotional life. The epistemologist, in turn, finds it clearly of importance for him to come to terms with some of the problems considered by the philosopher of mind. Thus, his theory of knowledge is likely to be influenced by the attitude that he adopts toward the mind–body problem; he is concerned with the central notion of a knowing self; and he finds that the descriptions provided by philosophy of mind of mental states, mental acts, and mental dispositions are of considerable value.

Relation to logic. It is sometimes said that epistemology is an admixture of many disciplines, of, for instance, metaphysics, psychology, and (since its method is analytic) logic. The fact is that the aims of epistemology are very different from those of logic. The concern of logic is with the formal structure of valid reasoning; it sets down the principles of valid inference. Epistemology, on the other hand, deals with the nature of cognition of all kinds: with man's awareness of the world in sense perception, his insights into possible hypotheses, his guesses—which are in no sense the concern of logicians. Though epistemology makes use of logical teaching on inference and particularly on deduction, it is also true that there is a difference between the epistemologist's attitude and that of the logician: the logician seeks the form of deduction in general and formulates it: *p* implies *q; p*, therefore, *q*. The epistemologist accepts his formulation but then asks what addition is necessary to describe the inference in full. If, for instance, there is present an intuition that *p implies q*, this is a point that is irrelevant to the logician but of great concern to the epistemologist. Epistemology cannot be identified with logic even in respect to that part of it that is most analytical in character but logical techniques can be applied to the problems of epistemology (see LOGIC, APPLIED: *Epistemic logic*).

Difference in aims and scope

Relation to other studies. Three other studies are related to epistemology, though not as closely as those already considered: (1) *Morals:* it is sometimes argued that a person ought to accept certain beliefs, even when he is not completely convinced of their truth, because of the good consequences of the acceptance, and it can also be argued that a moral issue is present in any act of knowing, that knowledge is an ethical concept, that knowledge that a certain proposition is true involves the duty to accept or believe the proposition. (2) *Sociology:* the problem of the influence of society and its prevailing culture upon the beliefs of an individual or of the way social relationships affect the power of a man to see the truth for himself occurs not only at the level of ordinary life; when scientists work in teams, for example, the question arises whether pressure is exerted to bring an individual scientist into line with the rest of the team in his acceptance or rejection of scientific theories (are there, for example, group ideologies in science as there are in politics?). Sociologists often provide epistemologists with interesting information on such topics as whether philosophical theories are communal or individual. (3) *Philosophy of religion:* Interrelationships here have to do with such questions as whether all human knowledge is illumination; whether there is special revelation; whether men literally "see all things in God"; whether it is essential that any knowledge that men may gain of God is God's self-revelation to them; and how revelation fits into the rest of man's knowledge (see RELIGION, PHILOSOPHY OF: *Basic themes and problems in the philosophy of religion: The problem of God, the Absolute, or the supreme value: The knowledge of God*).

Morals, sociology, philosophy of religion

II. Diversities of contemporary epistemological theory

It is impossible to provide a single definition to cover all of the many uses of the words know, knowing, and knowledge. The use that most concerns the epistemologist is knowing that something is the case. Such knowing may be certain knowledge or it may be only probable knowledge. Knowing that something is the case is propositional knowledge: it is knowing that a proposition or statement is true or probable. Such knowledge presupposes that language exists. But "knowing that" is only one kind of knowing; and no epistemologist would say that knowing is knowing only insofar as something is expressed in a verbal statement—it is not mere assertion. Further, even though a man asserts that he knows something, it need not always follow that he does know it, for he may be stating something he really knows to be untrue, or he may be in error. There has been considerable contemporary concern about the so-called performative uses of "I know that . . ."; thus a man may use sentences beginning with this utterance to express, for instance, conviction or to emphasize a point ("I *know* that he will come") or to express agreement, or even to plead with someone ("I know that you will help"). It is thus wrong to identify knowing with saying that one knows.

The meanings of knowing

Epistemologists draw a distinction between "knowing that" and "knowing how." (One could similarly distinguish between "knowing that" and "knowing why," "knowing where," "knowing who," and so on.) "Knowing how" is not a knowing that something is the case. "I know how to speak German" points to a skill or capacity. "Knowing *how*" may also include some "knowing *that*" —knowing, for instance, that, in German, the verb tends to come at the end of a sentence. Of the two, "knowing that" is of greater interest to the epistemolo-

gist because the statement, if significant, has truth value. Among true–false statements, however, there is a further distinction in the ways in which a statement may be true. "Lead is heavy," for example, differs from "If something has a given quality, then it either has that quality, or it does not have it": the first is a factual statement made on the basis of sensory experience; the second is a logical truth and not a matter of experience. The first is contingent—lead might not have been heavy but in fact it is; the second is necessary. Acceptance of the finality of this distinction makes it possible to distinguish, in turn, between two kinds of knowings—one factual (of the external world and of the world of mind or spirit), the second logical. Finally, there is a distinction between immediate (direct) and mediate (indirect) knowledge. One form of the distinction comes down from the ancients, that between uninferred and inferred knowledge; a more modern form is that between knowledge by acquaintance and knowledge by description; a third is the distinction between knowing a sense-datum or sensum or phenomenon directly and knowing the thing itself, the reality behind it, indirectly.

Factual and logical; immediate and mediate

SENSE PERCEPTION

Sense-datum theory. The study of sense perception begins with consideration of a person's factual knowledge of the world and is concerned not merely with that world but with man's consciousness of it as well. The world is conceived of as a world of material objects that exist in space and time and continue in being over a period of time. Such objects are solid and are able to affect other objects and to be affected by them. Among the objects that they affect is man's body; and he, through his sense organs in particular, registers those affections. The outcome, it is claimed, is that he becomes aware of the physical world. Physiologists and psychologists describe these sensory processes without pretending to a full knowledge of them.

Anomalies of perception

The above account, however, needs considerably more detailed analysis, and with analysis comes dissension. In the first place, it is naïve to suppose that a table before a man directly affects his eyes and so enables him to see it; it is rather the light rays deflected from the table that affect his eyes. If this fact be admitted, it may still be argued that the presence of the table is nevertheless the differentiating factor causing him to see a brown, rectangular patch when he now looks in its direction. But to say this may also be naïve, for in another light its colour may appear yellow and from another angle its shape may appear rhomboid. It may be said that these are abnormalities and that the man knows how to allow for them. But the contention involves the assumption that the "normal" colour is the real colour. Such an assumption requires justification.

Reflection on these contradictions led to the so-called sense-datum theory. According to this theory, the yellow object must be included among the objects seen, so must the florin-sized moon that is seen. Furthermore, if a man sees a mirage, he sees it, and it, too, is a visible object. Objects immediately seen or felt are said to be sense-data that exist, though they are to be distinguished from "real" objects in the physical world. The moon that a man sees is different (*e.g.*, in size) from the "real" moon; yet, it is supposed to be related to it; the problem is to define how the two are related, to determine whether a person can gain any knowledge about the "real" moon by considering the characteristics of the visible moon.

Sense-data, sensa, or sensibilia exist, in this theory, as particular objects, and in this sense the dualism of sensum and real object is complete. Sense-data, some say, can even exist unsensed. They are "there." On the whole, however, theorists tend to be neutral on the nature of their existence; sense-data are fleeting private objects rather than public ones, and, since they are linked with specific physical objects, they may in some way represent them. But this part of the doctrine is invariably vague, and it does not necessarily follow that the sense-data must be products of the affection of sense organs and brain by external physical things; it is at least conceivable that they are brought about by man's own action or, again, by God's action. Therefore, the fact that a person experiences sense-data does not permit him to assert that physical objects exist.

Problem of the existence of material objects

This conclusion, however, is unwelcome to sense-data theorists, some of whom have tried to counter it by arguing for a causal theory of perception on the following lines: because sense-data exist and have come into being, it is natural to infer that they must have been caused, but the most plausible answer as to how they are caused is that they are the product of the action of physical objects upon sense organs. Yet, the causal argument, which may be strong in other contexts, is weak in this context. For in the dualist sense-datum theory (1) the person does not know directly that physical objects exist, and (2) he does know that sense-data could conceivably have been wrought by other means. It is consequently difficult in this theory to justify belief in physical objects and doubly difficult to characterize the physical objects linked with specific sense-data.

Three lines of action in response to this predicament appear possible. First, the epistemologist can attempt a thorough re-examination of the ways in which men come to know the existence of physical objects. Second, he can deny the independent existence of physical objects and attempt to develop a fresh Idealist, or mind-centred, theory of sense perception. Third, he can adopt a phenomenalist theory—an alternative that was in fact attempted by the Empiricists David Hume and John Stuart Mill and elaborated in the mid-20th century.

Phenomenalism. The phenomenalist begins by asserting that any attempt to infer the existence of physical objects upon experiencing sense-data is wholly unnecessary. Both (1) the naïve theory that a person sees at least the surface of physical objects directly and (2) the causal argument have been discredited. He is sure only of the phenomena—of what appears. But this is quite enough, the phenomenalist suggests, to enable a person to *talk* significantly of physical objects; and, if so, nothing more is needed. He can certainly talk sensibly of this table without first showing that physical objects exist independently of himself. For here is a brown, rectangular phenomenon in front of him: if he puts his hand on it, it feels smooth; if he presses upon it, he feels a resistance; if he tries to lift it, it feels heavy. He experiences, too, the phenomena of space and of time; for he has to move forward in space to experience the tactual, and he has an experience of earlier and later. When he says "There is a table here," all he means is: "If I look in the right direction, if there is sufficient light, etc., I shall see a brown, rectangular figure"; "If I stretch out my hand, I shall feel a smooth surface"; and so on. In other words, without any loss of meaning, he can always reduce a physical-object sentence into sentences about sense-data.

Criticisms and rejoinders of phenomenalism

Though phenomenalism still has its defenders, it has met with considerable criticism. (1) It is argued that in stating the theory the phenomenalist already presupposes the existence of physical objects. In saying "If I put out my hand, I feel a smooth surface," he presupposes his own existence as a physical body. If the phenomenalist replies that he sees his hand as he sees the table, phenomenally, it may be countered that, when he says that he puts out his hand, this *I* is, at least, a physical object bringing about a certain change in a situation and that such causal changes cannot be accounted for in pure phenomenalist terms. The phenomenalist, however, might argue, in turn, that the account of causation that is presently orthodox is itself one that is framed in terms of a constant and regular conjunction of phenomenally observed sequences. The question then remains whether the orthodox account is true. (2) The series of phenomenalist statements into which the phenomenalist reduces a physical-object statement is an infinite one because there are, for instance, an infinite number of locations from which an object may appear. The phenomenalist reduction must, therefore, consist of more statements than any human being could ever complete. The phenomenalist might reply that sufficient phenomena might be observed to make an approximate reduction possible and so to

enable him to speak significantly of the physical object. But the question would still remain as to the point at which it could be said that he has observed sufficient of these related phenomena to justify speaking of one and the same physical object. (3) The physical-object statement, "There is a table here," is a categorical statement (of sheer fact); but the phenomenalist statements into which it is reduced are hypothetical: "If I look in this direction, I shall see a brown, rectangular patch." No categorical statement, however, can be reduced without loss of meaning into hypothetical statements. The physical-object statement categorically asserts the existence of a physical object. The phenomenalist might reply that the sentence "I now see a brown patch" is a categorical statement that is nonetheless phenomenalist. Yet, if this be granted, the sentence is still not a physical-object sentence, for, even if (improbably) it were held to be an affirmation of the existence of the *I* perceiving, it would still not necessarily be the categorical affirmation of a physical object. Clearly, not one of the phenomenalist's reductive statements is a categorical physical-object statement; thus the reduction must always be inadequate.

It has sometimes been maintained that the only difference between phenomenalist and realist is one of language, that, although the realist chooses to speak in the physical-object language and the phenomenalist in sense-data language, both are speaking of the same thing. Yet, it would appear that, if the criticism just considered is sound, then what is asserted categorically in the physical-object statement, viz., the existence of the physical object, quite apart from any linguistic considerations, can never be asserted in phenomenalist statements.

The need for a radical change of theory

Perceptual assurance of physical objects. The problems of sense perception were studied closely and energetically by epistemologists (by the British philosophers C.D. Broad and H.H. Price, in particular) in the decades between the world wars. Their approach was from the Empiricist angle, and, in spite of protests from Realists, the sense-datum theory tended to dominate the scene until adherents of the theory found it difficult to silence the critics who held that the theory ruled out complete assurance that physical objects exist. Discussion of sense perception thus reached an impasse and in the postwar period little advance has been possible; hence, philosophers have tended to turn to other problems. The one development has been that of phenomenalism, but, as has been seen, it too has failed to silence the critics.

It seemed necessary, therefore, to make a radical change in the approach to these problems. The Idealist solution, that only mind-dependent objects exist, appears to many philosophers to conflict with man's assurance of the existence of independent physical objects, the very same assurance that beclouds the sense-datum theory and phenomenalism. Some epistemologists are therefore tentatively approaching the problem from the study of this assurance itself. The acquisition of sense-data, they argue, is not the beginning out of which comes assurance of the existence of physical bodies; on the contrary, it presupposes that assurance. Sense experience offers further knowledge of that of which the existence is already known. The mistake that Empiricists have made is to identify sense perception with the acquisition of sensa, such as yellow, square, or heavy, and then to explain knowledge of the physical by compounding and relating such data. This is to enter into the perceptual experience when it is at an advanced stage, but to assume that this is its beginning, and to solidify by abstraction the elements that one then finds. For a human being to survive as an organism he has to interact with objects around him, and from the first, therefore, it has been a condition of his life that he should be aware of the world of objects around him. He possesses a nervous system that makes interaction possible and enables him to seek actively for information. Sensations are accompaniments of the operations of such nervous systems; they are wrongly conceived as objects in themselves. Some have felt that a false emphasis on sensations has blinded the philosopher to the real situation and has caused a crop of problems that have confused him.

A new Empiricism may thus be in the making that finds confirmation and backing in new teachings in psychology and physiology. In this view, no veil hangs between man and the world, or certainly none formed from sense-data. This is an Empiricism without skepticism. Nonetheless, certain difficulties remain, the chief among them being perhaps the part that seeing colours and hearing sounds play in clarifying the qualities pertaining to objects in the world. Is, for example, an object lying before the observer really brown and rectangular; are there some qualities that belong to things essentially, whereas others are only contingent; and can such problems ever be solved by philosophical reflection upon sense experience?

The newer non-skeptical Empiricism

LEVELS OF ASSURANCE

Quest for certainty. Of central importance to epistemology is the alleged assurance of the existence of a physical world. How this assurance is to be graded is a critical problem; a person may be certain of the existence of an external world containing physical objects but not be certain that the world is exactly as it appears in sense perception. On the contrary, such a person may think it necessary to carry out further tests and may check to determine, if he can, the qualities and relations of objects. In certain circumstances, he may even be doubtful about the very existence of an object that appears to be present; he may suspect an optical illusion or even a hallucination. But he is still certain that the external world and physical objects within it exist. He would claim certainty, too, for much else—for instance, that he himself exists and that other persons exist, that two and two are four, and that, if a thing has a particular quality, it cannot at the same time not have it. These a person takes to be certainties. Nonetheless, people frequently fail to achieve such certainty, and there are philosophers who hold that men never attain complete certainty. Complete certainty must then be further considered. The traditional Rationalist asserts that man gains complete certainty in the exercise of the faculty of reason; reason can provide him with absolute truth, at least in those areas in which it is free to function. It does not and cannot mislead, as the senses do; it in no way rests on sense perception. It is positive and contains no hypothetical element; it is not itself inferential, though it provides the principles basic to inference. It is intuitive and direct, providing indubitable and infallible knowledge—*i.e.*, knowledge of what is necessary.

Rationalist approach to certainty

There are philosophers who continue to defend a Rationalism of this full-blooded kind, though many, if not most, philosophers are hesitant about its claims and would reply that, admittedly, men are completely certain of many things, that such knowledge is contrasted with belief, which may carry a high degree of assurance but is never completely certain, but that, though a person may be certain of something and not merely opine it, it is still possible that he may be in error. If this is so, then he does not know infallibly by reason, and being sure is not enough. It has been asserted that there are three necessary conditions of knowledge: (1) a man, say Arthur, is sure that a certain proposition *p* (*e.g.*, "London is in England") is true; (2) he is able to justify his assurance by producing adequate evidence for *p;* and (3) *p* is, in fact, true. Yet the cogency of the theory thus formulated has been questioned. It is not clear what justifying a man's assurance (2) signifies. Apparently it means causing others, say William and Henry and Mary, to be sure by producing evidence for them to consider. This only means, however, that others are now sure, too. But if one doubts Arthur's assurance he may also doubt that of William, Henry, and Mary. So again, though it is obvious that, if Arthur really knows that something is true, it is true, Arthur may nonetheless think that he knows something even though it is false. Moreover, William, Henry, and Mary may all think they know it too, and yet it may be false. This means that conditions (2) and (3), though genuine conditions of knowledge along with (1), cannot possibly be used as tests whether Arthur is knowing. So there is no adequate test whether Arthur is knowing, and yet it is admitted that men can be sure and at the same time be in error. Such a conclusion points to skepticism about the Ration-

alist's ideal. One should be content with belief, the skeptics argue, and not claim certainty. With the necessary modifications, many philosophers could adapt the formulation used in the case of knowing something and apply it to believing it instead. They could then hold, as many philosophers do, that belief occurs only if (1) someone believes that something is the case; (2) what is believed to be the case is evident for the person believing it; (3) what is believed to be the case is true.

Infallible versus certain knowledge

The question then arises, however, whether the rejection of the possibility of completely certain knowledge is justified. The distinction between infallible knowledge and fallible certain knowledge is of vital importance in this context. To say that men are never certain is to go against the facts. The fact is that men are daily quite certain of where they live, the size of their paycheck, that they are now tired, and so on; what does lie beyond them is infallible knowledge of absolute truth. But to deny the possibility of infallible knowledge is not to deny that of certain knowledge. It would appear, therefore, that the schools that speak today of certainty being beyond the reach of men and of all alleged knowledge being only belief are in error and that their error rests on their failure to distinguish between complete certainty, of which man *is* capable, and knowledge of absolute truth, beyond every possibility of error, of which, apparently, he is not capable.

Belief. In some of its uses, the word belief has an emotive overtone, as when a man speaks of belief in God, signifying thereby an element of trust in and esteem for the object. But most often the word is used without evaluative overtones. To believe is to accept something as being the case; it is a species of knowing that is not certain but probable. There are degrees of probable knowledge, ascending from the lowest probability to the highest. Further, belief may be either dispositional or occurrent—dispositional, as when a man is called kind although not now performing a kind act, and occurrent, as when he is said to be angry only while actually in that state of mind. A person gets used to certain recurrences and comes to believe dispositionally that they will go on recurring—that his food will nourish him, that clouds presage rain, and that fire causes heat. Some epistemologists argue that all believing is dispositional. And, for those who also deny the possibility of certain knowledge, it would seem that all cognition is dispositional. But this would seem to be an extreme view. Although much believing is properly dispositional, not all of it need be. If, for instance, an attorney in examining a case looks at all the evidence, weighs the possibilities carefully, and finally concludes that Smith is in all likelihood guilty, he cannot be said to have acquired this belief dispositionally; for in his case there is an act of judging, and the belief is thus an occurrent.

Sources and adequacy of evidence

The sources of the evidence on which a man may base belief require consideration. The first of these is, clearly, sense perception, the actual seeing, touching, and hearing; the second is the memory of such experiences; and the third is inference from the information gained initially by the senses. In addition, an observer may be conscious of his own feelings of pleasure, pain, joy, sorrow, grief, envy, love, and hatred. From these sources a body of information may be acquired to which a person can turn when seeking evidence and by which he is so disposed that certain statements are acceptable while others arouse suspicion. Another important source of evidence is "hearsay knowledge," all that a person has been or is being told by others. His information would be very slight indeed if it were confined to what he has gained independently of others; for the greater part of it comes from communication with others—through talking, writing, reading, picturing, and so on.

The next question debated is one as to the adequacy of evidence for or against a belief. Evidence is usually taken to be adequate (1) when it enables a man to settle the point at issue, to make up his mind, and (2) when it provides him with the truth. It has been suggested that there is no real difference here, for only a knowledge of the truth could settle the issue. But men often have to make up their minds on mere probabilities. Moreover, if knowledge is always fallible, then to say that the evidence is adequate would inevitably be to speak relatively; it would be adequate for the person now seeking to determine its adequacy. He would want to re-examine the evidence and to check the sensory experiences, the memory, and the inferring. He would reflectively study the dispositions and tendencies, both those that seem almost mechanical—for instance, beliefs that (to quote David Hume) are "lively ideas associated with present impressions"—and those that are the outcome of wide attentive experience of recurring phenomena stored over many years. He would ask whether too much was being taken for granted in such dispositional beliefs and whether they could always be trusted. In the same way, he would check the hearsay knowledge; the time and circumstances in which the informants reported and the reliability of the informants are important. Having carried out this re-examination of the evidence, he would seek for fresh evidence. The total result of his labours might then be that he would still find the evidence inadequate or adequate enough to base a probable belief upon it or entirely satisfactory. In such circumstances the test of adequacy would be subjective throughout; on the other hand, it would now be a test satisfying a reasonable man. Whether this would give it objectivity depends presumably on the account one gives of the reasonable. At this stage it would be wrong to assume that the adequacy is ever other than subjective.

There is also the further debatable point whether a belief for which the evidence has been shown to be adequate, a justified belief, is identical with knowledge in the strict sense. It can now be seen that the evidence is adequate only in a relative and not an absolute sense; nonetheless, it justifies belief in the opinion of reasonable men. A person might be convinced and certain, so that justified belief would here be one with knowledge in the strict sense and not be mere opinion. On the other hand, such knowledge would not be infallible and would not be knowledge of absolute truth. Clearly, three states are to be distinguished: (1) probable knowledge; (2) certain knowledge, which is not knowledge of absolute truth; and (3) infallible knowledge of absolute truth. Adequate evidence gives man (2), and, in this sense, fully justified belief is certain knowledge; but it is not knowledge of absolute truth.

Meaning of probable

The meaning of the term probable requires brief consideration. The probable is sometimes defined as what is believed—*i.e.*, it is subjectively determined. Efforts may be expended to make it less subjective by bringing forward supporting objective statements; but the probable, in this sense of the word, is never wholly objective. In other senses, however, the probable is objective throughout. When the mathematician speaks of what is probable—for instance, the probability of an event recurring in a sequence of events—he seeks objective certainty, and his probability is the ratio of the number of favourable cases to the total number (both favourable and unfavourable) of equally possible cases and can be worked out mathematically. In another sense, the mathematically determined probability of an event can turn out to be improbable; if the occurrence has, for example, only a 4 percent probability, one should speak of the occurrence as being improbable. Finally, it is reasonable to ask whether, if it is said that a statement is probable, this means that the speaker is certain that it is probable or merely thinks it is probable that it is probable.

MENTAL ACTIVITY

Thinking and language. What distinguishes contemporary epistemology from earlier studies in the field is its profound interest in linguistic problems. On the one hand, the demand for precision in the use of language has been greatly stressed. Some writers have spoken of the need for a special, technically exact, philosophic language; others have held that the whole task of philosophy is to make completely plain what is being said at any moment. Doubts have been expressed about the possibility of a specialist philosophic language, and attempts to handle philosophic problems in the language of formal logic so

far have not always been successful. It has further been doubted whether analysis, in the sense of analyzing precisely what is being said, is the whole of philosophy. Probably no one would deny, however, that as high a degree of linguistic precision as is possible is a requirement in all philosophic discussion and that the epistemologist, for instance, has been helped by closer analyses of such words as know, believe, and see.

His interest in linguistic problems has been heightened, too, by his realization of the part that language plays in human thinking. It has even been argued that there can be no thinking without the use of language. It would seem difficult to deny, however, that men who are capable of thinking verbally need not always do so: intelligent thinking in a football player on the field is not necessarily verbal thinking. Nonetheless, it remains true that human thinking is, for the most part, verbal. The question that then arises—and it is central to epistemology—is that of the influence of language on thinking. At issue is not so much the different influences that different natural languages—English, Chinese, Iroquoian—have on the thought of the respective peoples but the differences in thought that the use of language in general makes possible. Of great importance is the fact that it makes possible statements (in which a quality of a subject is predicated or a member is placed in a class or a class in another class) or compound statements (involving conjunction, disjunction, or implication). Though one sometimes speaks of animals as "deducing" something, yet logical deduction, which adds power to thought, appears to be possible only when a person has learned to speak in statements and to think verbally.

Nature of the concept and of the universal

Difficult problems that arise in discussing thinking are those of the nature of the concept and of the universal (or general term—see below *V. Conceptual thinking: Concepts* and *Universals*). There is fairly general agreement that a concept is not, as traditionally assumed, an inner accusative (or object)—although some contemporary accounts of "intentions" as objects of thought seem to come close to saying this. In general, however, the teaching that the concept is any kind of an object is rejected; philosophers prefer to speak, instead, of having concepts, and the fact that a person has a concept is revealed by what he says and does. In this view, the concept is a mental capacity and is thus dispositional; *i.e.*, present even while not overtly expressed. The concept of liberty that a person has, for instance, will be clear from what he says. But even this interpretation of what a concept is can hardly be adequate. The concept of liberty for the English speaker is rather the accretion of both linguistic and nonlinguistic dispositions around the word liberty, and to understand it fully would be to make plain the character of these accretions.

The problem of universals is more metaphysical because it involves consideration of what it is that lies behind the successful use of general words like man or whiteness (as opposed to "Adolf Hitler" or "this sheet"). Nominalists reply that nothing is behind it but man's powers of grouping, classifying, and naming; conceptualists claim that man's ability to use general words successfully involves dispositional beliefs about what is and what occurs, as well as linguistic dispositions; realists stress the importance of recognizing recurrences in nature. (For a discussion of these rival schools, see below *V. Conceptual thinking: Universals.*) The final theory of universals must assure man freedom to frame universals as principles of thought, but also, since thought must be applicable to the real world, knowledge of universals must in some cases rest upon a recognition of recurrences in nature.

Problem of induction

Inference. Textbooks group induction with deduction as species of logical inference. Induction, sometimes described as a passage from the observation of particular cases to general statements, is claimed to be the method of many of the natural sciences, normally providing probable conclusions only. Further, it is the method in daily life of finding rough generalizations or working hypotheses. But induction is somewhat hesitatingly grouped with deduction as a method of logical demonstration, for philosophers have been hard pressed to establish the validity of inductive procedures. On the face of it, induction is a passage from some cases to all, from all the observed cases (all observed swans are white) to all cases (all swans are white); and this is an extrapolation that cannot be logically justified. It is generally claimed, however, that the element of induction in inductive reasoning may be bolstered by items of deduction and also by analogical reasonings. Attempts have been made to use statistical methods, based on the mathematical theory of probability, to strengthen the induction. Again, it has been suggested that induction should not be thought of solely as a passage from particular observations to a generalization, that this step should rather be followed by a secondary induction, viz., the testing of such a generalization by its coherence, or lack of coherence, with the system of generalizations already established in the relevant field of inquiry.

But these proposals have not satisfied the critics, some of whom insist roundly that induction is a nonlogical type of demonstration. Indeed, it has been argued that the passage from observations to generalization is no demonstration whatsoever, certainly not one susceptible of logical analysis; "induction" in this view becomes simply an inspired guess, a hunch. Once a generalization or hypothesis is made, it can then be tested logically (*e.g.*, the discovery of one negative instance will refute it). Others reject this view and argue that the thinker, in proceeding inductively, does not merely leap at a generalization but also makes a rational estimate of the probability of the generalization, and in so doing, he proceeds logically.

Deduction from premises

Of all mediate knowledge the surest is established by deduction, a passage from premises known to be true to conclusions following validly from these premises (see SYLLOGISTIC). The epistemologist, however, is concerned with the nature of inference, seen at its best in deduction. Inference has been conceived as a technique of sentence transformation providing an efficient means of advancing in thought. On a different view, inference begins with a thought content that develops of itself in thought. For both of these views, inference is in a certain sense passive. On a third view, however, inference is an activity, viz., perceiving or intuiting an implication. The premises necessarily imply the conclusion; no one can choose what he is about to intuit, but the intuiting itself is free and the mind active.

Further, there are problems about acquiring premises. It is essential that the reasoner should have established their truth, for a valid inference from true premises can never yield a false conclusion. But these truths, as has been seen, may be either truths of fact or truths of reason. Whether this difference can be ignored or necessitates two theories of inference depends upon a person's view of reason.

Reason and the a priori. Finally, consideration must be given to contemporary Rationalism. Its expression differs in detail from the older traditional doctrine, but what is essential to it today, as in the past, is the assertion that account must be taken of man's faculty of reason as a source of information that cannot be derived from any other source. Reason, operating in its purity, provides knowledge of foundational concepts, such as being, unity, substance, and cause, and of principles, such as noncontradiction (that a statement and its negation cannot both be true), which are conditions necessary for the possibility of knowledge. It provides, that is to say, the pure a priori features of human knowledge, as contrasted with the a posteriori features derived from experience. Traditionally, thus pure a priori knowledge is infallible.

The Rationalist today can adduce considerable evidence in favour of his thesis, an instance being the manner in which man's thinking is dominated by such principles as that of noncontradiction. He does not claim that all knowledge is acquired by reason (he claims less in this respect than did the traditional school); but he insists on a certain minimum of a priori without which human knowledge, he holds, cannot be understood. Nonetheless, the Rationalist thesis confronts acute problems, one of which is whether there are truths that only reason, a special exclusive faculty, can know. Analytic truths are

Analytic, categorial, and transcendent truths

said to be of this nature; so, too, are categorial truths about the structure of the world, and, in addition, there are transcendent truths. In respect to the first, the so-called analytic truths, it must be admitted that a statement can be true solely in virtue of the meaning of its terms (*e.g.*, "A bachelor is unmarried"). Once the terms are understood, the truth of the statement must be granted. Here, it is said, is a truth apprehended by reason. Yet the question can be asked whether it is in fact essential to suppose a special faculty to be at work here. These statements are analytic; they are necessary and, in that sense, a priori. Yet it may be asked whether this discerning or intellectual perceiving—that, since the terms mean what they do mean, the statement must be accepted as true—is really a special faculty operative only in this very limited field or whether it may not be the discerning that provides assurance throughout man's cognitive experiences.

As for the body of basic categorical knowledge that Rationalists believe conditions the development of further knowledge, considerable doubt is expressed today about their description of such knowledge—*e.g.*, of substance and of cause. The fact is that the bases of man's conceptual systems can and do change, that the principles usually held to be basic to thought in general, such as those of noncontradiction, identity (that everything is what it is), and excluded middle (that no third statement can be admitted between two contradictories), are still being reformulated and amended by logicians. It is not clear how this rethinking could happen, if reason once and for all had planted these absolute truths upon man's mind. Finally, it is argued that knowledge of the transcendent, the nonempirical, which, too, is taken to be a priori, may well have been derived from the experienced, even though it itself is never directly experienced. Thus, Rationalism has had to face considerable criticism. Some of the criticisms have pointed toward the view that human knowledge rests throughout on convention, but conventionalism itself has come up against assurances that man has in his experience about things and occurrences that are independent of him; and such knowledge can in no way be regarded as conventional.

Though it is inevitable that many matters that are relevant to contemporary epistemology have been omitted from this brief sketch, some of the gaps will be filled in later. The purpose of the foregoing outline has merely been to provide an insight into the main confrontations.

III. Rationalist orientation

In the next three sections the main doctrines of epistemology are set out primarily on a notional rather than a chronological basis. Epistemological theories can be divided roughly according to whether, in the gaining of knowledge, they stress the role of (1) reason, (2) sense experience, or (3) concept and language.

RATIONALISM PRIOR TO KANT

Pre-Socratic Rationalism

Classical Rationalism. Though ancient Greek Rationalism had its first explicit expression in Plato's doctrine of Forms or Ideas, there had been earlier suggestions of Rationalism. The Greek philosophers had consciously set before themselves the task of understanding the world, primarily the physical world, and this they interpreted to mean understanding the laws that hold within that world and its permanent structure and order. This was to assume, however, that it was an ordered and rational world—an assumption that is seen clearly in many of the extant fragments from early Greek writers. Thus, the philosopher and mathematician Pythagoras of Samos (flourished late 6th century BC) thought that "all things are numbers"; and the monist Parmenides, a generation after him is held by many scholars to have been a Rationalist. Certainly, a little later Anaxagoras argued in Athens that *nous* (reason or intellect) was the original cause of movement in the universe. He described *nous* as "the thinnest of all things," as infinite, completely pure and unmixed, and as absolute in power and in knowledge of all things. Both Plato in the *Phaedo* and Aristotle in the *Metaphysics* expressed disappointment that Anaxagoras made so little use of the notion of *nous* in his cosmology.

Theory of Forms in Plato. In his theory of Forms or Ideas, Plato, no doubt influenced by Socrates, put forward the first fully explicit Rationalist epistemology. Possibly the most striking description of Form and of the knowledge of it is found in the *Symposium*, in which it is said that a man who has observed many beautiful objects and reflected about them may, of a sudden, glimpse supreme Beauty, "eternal, unproduced, indestructible"—not beauty relative to a time, a place, a thing, or a person but absolute Beauty. "All other things are beautiful through a participation of it." Along with the Form of Beauty go the Forms of the Good and of the Just discussed in *The Republic*. Other forms mentioned by Plato are those of unity, plurality, equality, likeness, unlikeness, motion, and rest. Later, he adds the Forms of things, for instance, of man and all living creatures, of fire and water, possibly even of hair, clay, and dirt.

Nonsensory origin of the Forms of Plato

The Forms are not known by the senses, and he explains in the *Phaedo* how the senses may show things to be more or less equal, though equality itself ("the just equal") is not seen or felt but is known independently of sense. Forms, Plato holds, may be recalled from knowledge that the soul possessed prior to its union with its present body or may be directly apprehended. In either case the knowledge emanates from a divine element in the soul that knows the transcendent Form and eternal object and thereby reveals its own eternity. The same account of Forms is found in *The Republic*, in which a contrast is made between opinion and knowledge. Opinion, resting on sense experience, informs a person of the phenomenal, but knowledge reveals the unchanging real; the former is fallible but the latter infallible.

Thus, the outcome of Plato's discussion of knowledge was, first, a conviction that, whereas sensory experience is superficial and deceptive, the soul nevertheless is able to know—and to know infallibly—the immutable, the essential, the Forms that exist separated from the phenomenal. Secondly, a universal science that is eternally true is possible.

In turning to Plato's later dialogues, it is surprising that so little attention is paid in them to the Forms. The most important epistemological study among them, the *Theatetus* (to be considered later in *IV. Empiricist orientation: Classical philosophers*), is an examination of sensory knowledge and of belief that manages to carry this out without once referring explicitly to the Forms. The *Philebus* and *Timaeus* have only certain comparatively brief references to them. The *Sophist* also fails to provide a full discussion of Forms; and what is most interesting is that, in it, Plato argues that "change, life, soul, and understanding" are real along with the Forms. The one dialogue of this latter period that deals at length with the Forms is the *Parmenides*, in which the ancient monist Parmenides takes the young Socrates to task and is critical of his theory of Forms. The purpose of the dialogue is puzzling; though it is sometimes thought that Plato here repudiates his earlier doctrine of Forms, this is not the generally accepted view. Instead, the dialogue may well express some of Plato's doubts about the adequacy of his earlier views on a crucial point—the relationship between the Forms and the particulars that participate in the Forms. Possibly it may express, too, his doubts about the ontological status of the Forms and of the relations that the Forms have to one another.

What is important is Plato's statement of Rationalism. The senses do not reveal the true structure of the world, he holds, but reason reveals it; consequently rational knowledge must be clearly distinguished from sensory awareness. These are quite distinct faculties and their objects are probably distinct as well. Knowledge of the objects that reason reveals is presupposed in all other knowledge. The theory of Forms thus aimed at providing a Rationalist foundation for any subsequent epistemology. Further, it provided ontology (the study of Being) with the permanent and immutable (in contrast to the mere flux of the followers of the Pre-Socratic philosopher Heracleitus) and morality with absolute standards (in contrast to the subjectivity of the Sophists, private teachers of success).

Categories, nature, and reason in Aristotle. Aristotle qualified Plato's Rationalism but did not reject it. He was critical of the Platonic view that knowledge is a universal system deducible from a principle, such as that of the Form of the Good. However excellent such a deductive universal science may appear in theory, Aristotle saw that in practice it transcends human capabilities; when human knowledge is considered as it is, it is clearly not Plato's universal science, and his epistemology is erroneous because the theory of Forms on which it rests is erroneous. In his *Ethics* and *Metaphysics,* Aristotle attacked the concept of a Form. Plato, he held, thinks of it as a separate existence, an individual. But, if the Form man is an individual existent, it presumably has something in common with Socrates; but this common element points to a further Form, the third man, and this begins an infinite regress (a fourth man, a fifth man, . . .).

Criticism of Plato's transcendent Forms

Aristotle held, instead, that the things of the world are individual existences having qualities. Though each individual is a substance separate from every other, individuals may have qualities in common. Yet, if two substances have a common quality, this in no way justifies the assertion that a Form of that quality exists. In Aristotle's opinion this theory is false and misleading, for the quality—admittedly common and admittedly as real as the things sharing it (though not particular, as they are) —exists as a quality of both and possibly of many other things. It does not exist as a self-subsistent entity, and there is no Good in the Platonic sense, though many things are good.

Nor, finally, did Plato succeed in making clear the relations between the Forms. Aristotle rejected the view that there is a genus, such as Being or Good, which contains all other Forms as species within it or which has as species substance, quality, quantity, and relation. Aristotle held that the latter are what he called categories and listed. What exists may be a substance, a quality, or a quantity, and what has being may exist in any of these senses. But there is, for Aristotle, nothing that exists that is all of these things, that is to say, Being; a category is not a species of the genus Being.

In these respects, then, Aristotle rejected certain aspects of Plato's Rationalism, but he nonetheless retained Rationalism itself and reaffirmed it, holding that the world is rationally ordered (though there is a place for chance in it). God is its unmoved mover. Because God, however, is wholly self-contemplative, it is nature rather than God that controls the growth of the world, a nature that "does nothing in vain" but always works for an end.

There are difficulties in this account of the function of nature. It is not clear, for example, that nature proceeds rationally; Aristotle thought that conscious reason is not found at any level of life lower than man. In the *Ethics* a high value is set on reason, and its use in contemplation is said to be man's deepest happiness. It is "the best thing in us, for even if it be small in bulk, in power and worth it surpasses everything." Aristotle followed Plato in ascribing the adjective divine to reason. His fullest account of it is in *De anima* (*On the Soul*), in which his psychology is set forth. A man's soul, he wrote, is to his body as axeness is to axe; it is that which realizes the body's potentiality. Like other living things, men learn how to find nourishment; they reproduce themselves; they see, imagine, desire, and seek. But man, too, can reflect and think, can affirm that a predicate pertains to a subject or deduce one statement from another. Human beings thus possess intelligence and the power of reason (*nous*). Insofar as it concerns itself with material provided by sensory experience and imagination, he held, reason itself partakes of the passivity of these faculties.

Reason as divine: "the best thing in us"

Man's rationality, however, reflects also the presence of active, as opposed to passive, reason. Active reason does not proceed mediately, through judgment and deduction, which are fallible; it is direct knowledge. Being wholly free from the affection of the material and in no way dispositional, active reason owes nothing to the development into which it momentarily enters and which it illuminates; it remains, instead, what it is essentially, immortal and eternal. It would appear from this (the passage is difficult to interpret) that Aristotle—in spite of his emphatic rejection of the theory that knowledge of transcendent Forms is essential for the attainment of truth—nonetheless concluded that man gains the final truth only when an element enters into the knowing process that has not appeared previously in that process; that which is no development of it but something itself wholly transcendent, eternal, and infallible. This is active reason. By it man knows such basic principles as that of noncontradiction.

Later Greek and Roman Rationalism. There are Rationalist elements in extant fragments from the Stoics, Epicureans, and Skeptics, the most notable writers of which were Zeno of Citium, Epicurus, Chrysippus, Poseidonius, Carneades, and Aenesidemus. All of these (with the possible exception of Poseidonius) were Materialists; yet the Stoics, in particular, freely introduced Rationalist elements (which could not always be made consistent with their Materialism) into their philosophies. Even the Skeptics, paradoxically, employed rational arguments against Rationalism.

Finally, reference must also be made to Plotinus, the foremost Neoplatonist philosopher, who was a Rationalist in the respect that he paid to *nous.* In Plotinus' view, *nous* is the wisdom in the universe and the first emanation of the One, which is supreme. Through *nous* the world is illumined throughout all stages of its hierarchy. These stages consist in the One, *nous,* soul, soul-in-body, matter. Man, who is soul-in-body, can strive by rational means to understand what is above him; he may grasp the rationality of the world. Yet he cannot take the final step by reason alone; God is the One; God is not an object standing over against a knower, and men cannot be such knowers. The final step is not reasoning to God nor even intuiting him; it is union with him. Plotinus' philosophy thus ends in a mysticism that rests on Platonic Rationalism.

Plotinus' *nous*

Rationalist elements in Scholasticism. The Scholastics, as the philosophers of medieval Christendom were called, are characterized by their outstanding confidence in reasoning and are sometimes criticized for having overemphasized the importance of sheer deductive argument and of subtle analysis in philosophy. Yet their confidence in reasoning and argument does not of itself prove them to have been Rationalists, except in a superficial sense. As has been seen, Plato was a Rationalist not in stressing the importance of deductive argument but in his doctrine of eternal Forms apprehended by the infallible grasp of *nous* (rational intuition); and Aristotle, who rejected the Forms as separate, eternal entities, nevertheless accepted and retained the notion of a supreme intuitive faculty essential to knowledge.

In the centuries that followed, there was considerable discussion of the celebrated account of active and passive *nous* in Aristotle's *De anima;* in the teaching (in the 11th and 12th centuries) of the great Arab-language philosophers Avicenna and Averroës, for instance, some part of the separatedness of the Platonic Forms was brushed off onto Aristotle's *nous.* As stated above, Plotinus thought of *nous* as separate and, while not deifying it, made it the first emanation from the One. The Arabs, too—and, after them, Siger of Brabant, of the secular faculty at Paris—tended to speak of *intellectus* not as a faculty in the human mind but as a separate entity, in which men participate, so that it is one in all men. In the ensuing (13th) century, however, St. Thomas Aquinas rejected this doctrine forcibly. To accept it, he argued, is to make impossible a coherent account of the intellectual lives of men in view of the intellectual differences between them; it is to deny man's individuality and so his moral responsibility; and it is to impugn his status as "master of his own operations." Though Aquinas rejected this influential doctrine, he in no way rejected Aristotle's Rationalism in so far as this (along with Plato's) consists in regarding the intellect as the supreme faculty in knowing.

The works of Aquinas best reveal the character of Scholastic Rationalism. The soul, he held, is "the first principle of life in those things which live" and shows itself in two aspects, viz., knowledge and movement. In man's case the

Thomas Aquinas' view of the soul

soul organizes the life of the body, and soul-in-body possesses not only sensitive and nutritive powers but also intellectual powers. Intellect, being one of the soul's powers, is not the essence of the individual human soul, but neither is it the same in all men. The intellectual principle is multiplied according to the number of bodies. Man is the lowest of intellectual beings, as contrasted with angels and God, and his intellect is described as "potentiality with regard to things intelligible." But it is not mere potentiality, for Aquinas accepted Aristotle's view that the human soul has, too, its active *nous:* human intellect is able to participate in the superior intellect of God and so man is illumined and things are made intelligible. Intellect, wherever it is found, is concerned with the universal and general. In perfect intelligences, these universals and the relations between them are all grasped directly, but man "advances from one thing understood to another" through discursive reasoning. Nonetheless, both the intuition and the reasoning are intellectual, though reasoning has a certain measure of imperfection as compared with intuiting. Men, however, are not wholly deprived of intuitive knowledge since they can begin with principles, such as that the whole is greater than the part, which come to man *naturaliter*—*i.e.*, "everyone knows them."

Not that these principles are known innately; but, since man is intelligent, they are from the first present "in potential." Man needs the occasion to activate what is there. His mind "is at first like a clean tablet on which nothing is written": but then he experiences the sensible; and, in consequence, what is all along present potentially in his intelligent life is now actualized, and he gains understanding. "With us to understand is to be passive." Bodily sensations provide the occasion, and the person, who is soul as well as body—*i.e.*, intellectual as well as sensitive—is illumined. Aquinas did not adopt the Platonic view that sensation and intellect are alien to one another in man's case. On the contrary, it is through sensory experience that the human intellect comes to understand, and it is to the soul's advantage that it is united with the body. Man's knowledge begins with sense experience, and it can do so because the sensory phantasma (or presentations of the mind) can be illuminated by active intellect, which then by its power abstracts intelligible species or concepts from them. The senses determine the particular, but intellect then receives the universal. Having received it, however, the human intellect—because of its lowly grade—is not powerful enough to gain immediate and full understanding, but it gains enough to proceed mediately by discursive reasoning.

So Aquinas demonstrates the necessity of sensory experience as providing particular sense images and of intellect as illuminating the sensory so as to abstract the universal. Just as in Aristotle, there are thus two aspects to human intellect: (1) the actualizing of the potential, which is the work of intellect as active, and (2) the passive reception of understanding, passive intellect. Aquinas made sure, however, that his own theory of knowledge, in spite of its recognition of the necessity of sensory experience, is no sensualism but is Rationalist in its recognition that knowledge is possible only on the possession of rational powers. Of all the Scholastics, Aquinas was thus the most like Aristotle in his epistemology, and he shared in his Rationalism.

The work of Duns Scotus and William of Ockham

The fortunes of Rationalism in Scholastic thought after the death of Aquinas can be traced in the philosophies of two scholars. At the turn of the 14th century, Duns Scotus, the great British Scholastic, sought to make clearer the character of a truly rational knowledge and maintained that arguments from effects to causes (*e.g.*, from the existence of the world to the existence of God) never provide conclusive proof. Man knows of God's existence through revelation; he can have no a priori proof—the only kind that, in Scotus' view, would be a true rational demonstration. William of Ockham, trained as an Oxford Franciscan, developed the same theme in the mid-14th century. Like Scotus he had no doubts about revelation and the faith, but he carried skepticism about Aristotle's and Aquinas' Rationalism well beyond Scotus. He asked whether the intelligible species are anything other than the philosopher's own mental creations. He even suggested that *intellectus,* whether active or passive, might be nothing more than an abstraction thought to be a real entity. Entities, said Ockham, should not be unnecessarily multiplied.

17th-century Rationalism. Descartes, Spinoza, and Leibniz—sometimes known as "the continental Rationalists"—mark the climax of precritical Rationalism.

Cartesian Rationalism. René Descartes is held to have been the father of modern Rationalism. His Rationalism lies in two features of his teaching, presented in his *Discours de la méthode* (1637; *Discourse on Method,* 1950): first, in what he says about method and, second, in his account of clear and distinct knowledge. Having expressed dissatisfaction with his education, he decided to subject everything that he had hitherto accepted as true to a methodical doubt in order to build a surer science of God, man, and nature. His main interest in the *Discourse,* it is clear, was in the latter, the science of nature. (His particular purpose was "to deduce rules in medicine of greater certainty than those at present in use.") Descartes, himself an expert mathematician, attributed the success of mathematicians in his day to the method that they employed. Mathematicians begin with the simplest truths, the foundational axioms, and pass step-by-step to knowledge of the more complex. He proposed to use "all that is best both in geometrical analysis and in algebra" to solve the problems of natural science. Each of these problems was to be divided into the simplest possible parts; and, beginning with these simples, by chains of easy reasonings, satisfactory conclusions could be reached. This meant, he admitted, assuming a specific order to exist in the natural world, although one cannot be sure that it does exist there. The aim is to project mathematical order into the physical world so as to enable the inquirer to understand it rationally. In every case the investigator should also check each link in the chain by careful enumerations and reviews. This procedure is rational in a specific way, for Descartes sought not merely any order, as all Rationalists do, but an order that is strictly mathematical.

Descartes' reliance on the mathematical method

Of fundamental importance to his whole epistemology is another rule:

> To accept nothing as true which I do not know to be such clearly (*évidemment*) that is, carefully to avoid precipitancy and prejudice and to comprise nothing more in my judgments than what presents itself so clearly and so distinctly to my mind, that I have no ground whatsoever for holding it in doubt.

What is clear is "present and manifest to an attentive mind," and what is distinct is "precise and different from all other objects." The mind grasps the clear and distinct as infallibly true, so that an intuition of this order is the irrefutable answer to Skepticism. Descartes held that "*intuitus*" is

Descartes' notion of clear and distinct knowledge

> not the fluctuating testimony of the sense, nor the misleading judgment of a wrongly combining imagination, but the pure intellectual cognising (*conceptum*) of which an unclouded and attentive mind is capable; a cognising so ready and so distinct that we are wholly freed from doubt about that which we thus intellectually apprehend.

He who seeks for truth must not hope to find it in sense perception—he must "detach his mind" from the senses—nor again in blundering imagination but in a pure intellectual apprehension of infallible truth resistant to doubt. Here, then, at the opening of modern philosophy, is a definitive assertion of man's power to know indubitably by pure intellect, entirely by the light of reason.

Later, Descartes would seem to qualify the absoluteness and finality of *intuitus,* for he appeared to appeal to God's veracity as the final guarantee of the truth of what is known clearly and distinctly. Inasmuch as his argument—both to his own existence from his consciousness that he thinks (*Cogito, ergo sum*) and subsequently to God's existence—rests on intuition, it has been held that he is here involved in a vicious circle. His defenders, however, have held that no such circle is involved, that Descartes was pointing to the source of all man's powers in

God and at the same time to his own failures, particularly his failures of memory. The failure, in some views, applies to the occasions on which he fails to intuit clearly and distinctly; for, whenever he does so intuit, in spite of his general fallibility, he attains infallible knowledge. There is, further, a relation between *intuitus* and deduction Descartes speaks of both as part of the intellectual faculty of knowledge, which knows and never fails to know. Yet, Descartes has to admit that deducton is not as certain as intuition, for it is not simple. Intuition, is simple as sheer apprehension and also in its object. It can only apprehend "the simple nature," for the latter alone is clear and distinct. Deduction, on the other hand, is complex. Yet, it too, being intellectual, can give man the truth, as long as he does not *will* to err.

But the problem, then, is how to relate deduction to the intuitive apprehension of simple truth—which is what, in Descartes's view, knowledge essentially is. Deduction is mediate knowledge of the complex and can be knowing in the strict sense, Descartes argues, because each implication in the deduction is intuited; the implication itself must be "a simple nature." Yet, it cannot be the sheer simple of Descartes's sheer intuition. It must be simple in the sense of a "simple relation." But this raises two questions: (1) whether this relation is simple in the sense required and (2) whether, if deduction is intellectual, in this narrow sense, the Cartesian deduction would not have to be confined to the reasoning of pure mathematics and whether, in particular, inferences about the natural world—which are neither simple nor clear and distinct—could ever be *deductio*. This is to ask, in other words, whether deductive knowledge is always knowledge of the logically possible rather than of what actually is. It would appear that there remain difficulties in the account of deduction that Descartes provided.

Subsequent continental Rationalism. If this were a history of Rationalism, it would be necessary to add here extensive accounts of the 17th-century philosophies of Benedict de Spinoza and Gottfried Leibniz.

Spinoza's use of the mathematical approach

Spinoza, like Descartes, emphasized the importance of proceeding in a strictly orderly manner. The order to be followed is that of mathematics, as reflected in the title of his famous work *Ethica in Ordine Geometrico Demonstrata* (completed 1675).

All true knowledge thus shares the necessity of mathematics; each step and every new conclusion must be demonstrated with logical necessity. But this is to assume a metaphysic: the world known must be a system that is rational and necessary, and, if it is such a system, it must be one. Spinoza thus speaks of it as God, or Substance. Men know it under two aspects—thought and extension—that are not to be regarded, however, as different entities (as in Descartes's account of mind and body) but as two attributes of the one entity. A change in the body is a change in the mind; if the body is affected by another body, there is a parallel change in the mind, and this is the explanation of sense perception. The mind has an "idea." If it happens to a person's mind as the result of a body affecting his body, the idea is not an adequate thought of his body or of the other body; it is confused. But people do have adequate ideas—*i.e.*, ideas such that, if one thinks any one of them, he will see that it has all of the intrinsic marks of a true idea. This is the stage of *ratio* as contrasted with sense perception. The object is known as self-evident, clear and distinct (as Descartes said), but as something, too, that cannot be methodically doubted. "Whoever has a true idea, at the same time knows that he has a true idea, nor can he doubt the truth of the matter." Or, "for the certainty of truth no other sign is necessary than to have a true idea." True ideas, too, are the same for all men; a common body of science that all men accept can be worked out. But Spinoza, having reached this point, does not find it complete as an account of human knowledge, for he thinks that beyond *ratio* is an intuitive knowledge of God, of the one Substance. It is in this intuition that man finds pure intellect, and out of this insight comes an "intellectual love of God" that is eternal.

Like Descartes and Spinoza, Leibniz wished to see the method of the mathematician applied to philosophical problems. For this purpose he demanded a new symbolic logic that would ease the burden of philosophical speculation and at the same time ensure accuracy. Given such a logic, philosophers could settle their differences with the ease of an accountant, merely by calculating. Leibniz realized that such a calculus would only be possible if the world were one system, but this did not mean that it would have to be one Substance as Spinoza thought. The world, Leibniz held, is a composite of a vast number of simples; each simple, or "monad," is a substance having its own qualities. Every monad is different from every other monad; it "has no windows by which anything may enter or go out." Any knowledge that it has, consequently, is innate and comes from within. It is the subject of its own predicates and is constantly changing; for it expresses (without observing) the changes of all other things, and this is its "perception." Though this is a kind of knowledge, it is confused and indistinct.

Leibniz' symbolic logic

There are, however, monads—and men are monads of this kind—that know eternal and necessary truths (their opposites being impossible). They have this knowledge because they are reflective and contemplate themselves. They are thus led to such truths as the principle of noncontradiction. Here, the knowledge is clear, distinct, adequate, and intuited and is the highest that a man can attain. But Leibniz saw that man is also aware of contingent happenings, or truths of fact, about which he makes statements that are true without being necessary. They do not occur without a reason, however, for there is a sufficient reason for them in God's will, insofar as God desires to bring about the best of an infinity of possible universes. Necessary truths, on the other hand, are not dependent on God's will but depend solely on his understanding. Because God chose this universe as the best, the whole of it is in harmony with itself. Insofar as man is rational and apprehends the necessary and eternal, he is a being capable of understanding some part of the universe and so of entering into a relationship with God.

CRITICAL RATIONALISM OF KANT

In the 18th century, Immanuel Kant, the founder of critical philosophy, was a Rationalist but a Rationalist with a difference. He had been trained in the school of Leibniz and his systematizer, Christian Wolff, and accepted their philosophic standpoint until middle life, when he became critical of what he termed their "dogmatic Rationalism." The *Kritik der reinen Vernunft* (1781; *Critique of Pure Reason*, 1929), Kant's chief work, was a sustained and penetrating examination of the claims that Rationalists had made for human reason with the rejection of many of them. Nonetheless, Kant remained a Rationalist.

Kant's criticism of dogmatic Rationalism

Kant rejected the view of Descartes, Spinoza, and Leibniz that the philosopher should adopt the methods of mathematics. To suppose that the problems of philosophy can be dealt with by these methods is, he thought, to deceive oneself; for, whereas the mathematician constructs his own concepts, defines for himself what a triangle is, sets down axioms, and draws chains of inferences with exactness, the philosopher deals with concepts of metaphysics, such as space, time, substance, and causality, or of morals or aesthetics and does not construct his concepts but has to wait on experience and then seek concepts that appear to explain it. Inevitably, therefore, the method of philosophy is different from that of mathematics. In the second place, Kant takes the doctrine of representative perception seriously. Sensory experience provides a phenomenal knowledge of things. Consequently, no universal statement about what is experienced is immediately applicable to the real world; it is "a representation of a representation." In the third place, in such a situation, any talk of calculating the universe seems naïve. Kant denies that rational analysis, of itself, can provide new knowledge of what exists, and it is idle to expect analytical demonstrations to establish the nature of what lies beyond immediate experience. They cannot, for instance, prove or disprove the being of God. Kant expressly criticized Leibniz for "intellectualizing appearances." Leibniz, he said, had spoken of sensations as con-

fused, feeble intellections not yet clarified into adequate ideas and had failed to realize that sensibility and understanding are "two sources of representations which, while quite different, can supply objectively valid judgments of things only *in conjunction* with each other."

Kant's total rejection of dogmatic Rationalism is the explanation of the curious arrangement of his *Critique*. At the outset he had divided the work into what he intended to be two main parts, viz., (1) the transcendental doctrine of elements and (2) the transcendental doctrine of method. In the actual composition, however, the latter turned out to be little more than an appendix dealing with the uses of reason and reporting certain positive uses of it on the practical side. In respect to the theoretical side, however, Kant was content to repeat the warning given in the transcendental dialectic that, whereas pure reason is adequate in mathematics, it cannot provide, independently of experience, knowledge of the real world.

Yet, in spite of what has just been said, Kant remained a Rationalist. More often than not his thought was conditioned by his Rationalist training, and his language was the language of the Leibniz–Wolff school. This complicates his argument, especially at those points at which he was moving away from this school. Certainly, this continued use of the language and procedures of the "dogmatists" would not of itself make Kant a Rationalist; but his comment on Leibniz, referred to above, is more revealing in the contrast it makes between Leibniz and John Locke, a British Empiricist. Leibniz had intellectualized sensibility just as Kant thought that Locke had sensualized intellectual concepts. Leibniz and the dogmatic Rationalists were wrong, Kant believed, in supposing that knowledge is gained by intellect or reason only and that sensation is a low, passive form of intellect, unclear and confused; they did not understand that human knowledge must involve the operation of two distinct faculties, sensibility and intellect. Nevertheless, as against what he took to be Locke's standpoint, Kant fully endorsed the Rationalist criticism. If no account of human knowledge that ignores empirical elements is adequate, equally no account of knowledge in purely empirical terms is adequate; for a rational element is present in all knowledge.

It would appear that, although experience is admittedly a source of information, elements that are not empirically given are already present in the simplest experiences, that a man's more elementary experiences of himself and of things around him, for instance, involve space and time; but space and time are not given. So, too, in the case of the more precise knowledge of the natural sciences, features are present that, in Kant's view, cannot be derived from sensibility; insofar as knowledge is scientific, its statements are true necessarily, and, being necessary, they are also universal and eternal. Kant assumed, without ever proving it, that space and time are not given empirically, and he further assumed that experience can never provide the universal and necessary. These are basic assumptions of Cartesian Rationalism, and Kant accepted them without question. It then follows that there are elements in all knowledge that are not empirical. Kant's main purpose in the *Critique* was to reveal these elements, and, insofar as this is the case, the *Critique* is predominantly a Rationalist statement, even if it rejects some part of the earlier Rationalism.

The synthetic a priori and the categories

Kant agreed that a priori elements enter into all human knowledge. But, whereas dogmatic Rationalism held that all a priori knowledge is analytic, Kant held that the most important part of it is synthetic—*i.e.*, it provides genuine additional knowledge. Kant did not deny the possibility of an a priori knowledge that is purely analytic in the Leibnizian sense, but he regarded a man's knowledge of himself and of the natural world as synthetic rather than analytic and yet as involving features not given in experience—that is, as a priori. Whatever is given in sensation becomes part of a system of unified experience; the occurrence of the system presupposes a "unity of apperception" transcending sense. Such apperception should not be regarded as the work of memory alone, and it is no mere disposition; it involves the judgment-making and statement-making function—that is, the logical function. It is a system thought of, for instance, in terms of substance, a thing that is, or in terms of quantity, quality, or relation.

> All the manifold, therefore, so far as it is given in a single empirical intuition is *determined* in respect of one of the logical foundations of judgment and is thereby brought into one consciousness. Now the *categories* are just these functions of judgment, in so far as they are employed in determination of the manifold of a given intuition. Consequently, the manifold in a given intuition is necessarily subject to the categories.

Kant's account of the categories is worked out at great length in the transcendental deduction of the categories. Deducing in this context means justifying, and what he seeks to justify is that the process of knowing, bringing the given manifold into one consciousness, is regulated by pure concepts of the understanding, not given in experience. Sometimes the suggestion is made that these categories are innate, sometimes that men possess them although they are never conscious of acquiring them; but always it is insisted that they are not derived from the senses. They are the contribution of intellect or reason or understanding to the totality that is human knowledge. The number of the categories is fixed, and Kant gives a list of them and in complicated arguments attempts to prove the necessity of each category.

Transcendental Idealism

Kant's stress on knowledge as a coherent system pertaining to one consciousness leads to Idealism; the absolute Idealism that developed out of this aspect of his philosophy will be considered below. But, when Kant himself referred to Idealism in the *Critique of Pure Reason*, he was speaking of something else, viz., of the ontology of immediate objects of perception.

In two important passages, one in the first edition only and the other in the second only, he reflected on the difficulties arising from the accounts given of sense perception by Descartes and the British Empiricists. Sense perception is of appearances only, which, it is supposed, are representations of physical objects; the physical objects represented are not directly seen. But whether or not they exist is a question, it may be argued, that cannot be answered. One is tempted to hold that nothing can be said about the existence or nonexistence of such objects. This is one form of Idealism, however, that Kant rejected. Nonetheless, he did not reject Idealism. He proposed, rather, to give significance to talk about a physical object by thinking of it not as a "thing-in-itself" beyond the veil of sense but as phenomenal. It is what appears, appearing always in accordance with empirical laws and with the principles of space and time. These are not laws and principles of things-in-themselves but of the empirically known, of "the actual." Such an Idealism, which Kant termed transcendental, he thought was consistent with his view that space and time and the categories are a priori, for the transcendentals can apply to the actual physical world only if that world is phenomenal. They could not apply to things-in-themselves.

Subsequent reflection on this problem, however, led him to introduce a new theory in the second edition of the *Critique*. It is not only true that man does experience physical objects (though again not as things-in-themselves) in sense perception, but it is also true that such an experience is necessary if he is ever to be conscious of his own existence. For this is a consciousness of a being in time and presupposes whatever is necessary for experience of such a being. What is presupposed is perception of something permanent, and he found such an experience in his perception of a physical object or thing outside himself. Thus, he cannot know that he exists if he does not also already know that physical things having a measure of permanence exist.

POST-KANTIAN RATIONALISM

Rationalism of absolute Idealism. Absolute Idealism was first set forth explicitly and fully by Hegel, whose first published work, entitled *Differenz des Fichte'schen und Schelling'schen Systems der Philosophie* (1801), reveals the sources of his Idealism. Johann Fichte, an ethi-

cal Idealist, had developed Kant's notion of the ego, the conscious, free, creative moral agent, and applied it beyond the realm of morality to metaphysics and epistemology. The universe itself is to be explained in terms of absolute creative ego, and consciousness is to be understood in the same terms. Knowing is creating rather than discovering. Friedrich von Schelling, an aesthetic Idealist, on the other hand, had concentrated on the unity and systematic identity of consciousness that he thought was implicit in Kant's doctrines, particularly the doctrine of the unity of apperception. Schelling argued that it involves an identity of being, so that subject and object, the knower and nature, that which is other than him, are all one. Hegel built his system on these foundations.

The system of G.W.F. Hegel

To Hegel, the fundamental reality is not matter but mind, of which matter is one manifestation. Absolute mind is immanent, dynamic, and developing, the development being orderly and tending to ever-increasing complexity. The order is dialectical. The three-fold dialectical argument begins with a thesis, opposite which stands an antithesis; and out of both emerges a synthesis. It is not merely a method of argument or thought but is itself the method of the development of the universe—and this fact, indeed, explains its great usefulness in thought. Thus, the *Encyklopädie* (1817), the fullest statement of Hegel's philosophy, is set out in three parts: (1) the science of the idea in and for itself—*i.e.*, logic; (2) the science of the idea in its otherness—*i.e.*, the philosophy of nature; and (3) the science of the idea returning to itself out of its otherness—the philosophy of mind.

Hegel's is a logic very different from that of the textbooks of his day. The latter dealt with fixed, stereotyped, and rigid methods of reasoning, so abstract as to be in the extreme case almost irrelevant to being. A new logic, he claimed, is needed, not of reasoning but of reason. "The battle of reason is the struggle to break up the rigidity to which the understanding has reduced everything." Hence, much of the "Logic" of the *Encyklopädie* (its first part) is a discussion of epistemological problems, leading up to a final part that is a study of the idea itself—as it is rational but also as it has values, is free, and is the good. The second part, the "Philosophy of Nature," proceeds by the same sort of dialectic through the study of narrow mechanics and physics to organic nature; and the third part, the "Philosophy of the Spirit," begins with subjective mind, the sensitive nutritive organic consciousness operating in the world of nature, and proceeds through the objective mind of the person in society to absolute mind revealed in the deepest consciousness of the world in art, religion, and philosophy. These are themes that Hegel also elaborated in his other works, for instance, in the *Phänomenologie des Geistes* (1807; *Phenomenology of Mind*, 1910). It is seen that, in a negative sense, Hegel was opposing the formalism of the past while setting out positively a theory of development and evolution, both metaphysical and epistemological, which deeply influenced the thought of the 19th century. His faith, in the last resort, is a faith in reason; his evolutionary theory in no way rests on change. Reality works out its own development rationally. The foundation of his philosophy (and that of his followers) rested on his celebrated statement, "What is rational is real, and what is real is rational."

Absolute Idealism was a powerful force in the German universities in the mid-19th century, although (apart from Hegel) the movement yielded no outstanding thinkers; possibly the most important was Ludwig Feuerbach, a humanizer of theology, who gave a strongly naturalistic turn to the Hegelian teaching and provided some of the foundations of the Marxist philosophy. Outside Germany, absolute Idealism flourished at the turn of the 20th century in Britain in the writings of F.H. Bradley and others, in America in those of Josiah Royce, and, somewhat later, in Italy in Benedetto Croce.

The system of F.H. Bradley

Bradley did not regard his philosophy as Hegelian; he was critical of Hegel and was doubtful, indeed skeptical, about many points in his philosophy. Nevertheless, the influence of Hegel is apparent everywhere in his writings. Metaphysics, logic, psychology, and epistemology are, in his view, all one inquiry. Human knowledge is what it is because the universe, of which the thinking man is a part, is what it is—that is to say, a rational system, coherent and self-fulfilling. Accordingly, the criterion of successful thought is its systematic coherence; thinking is of relations, all of which are internal to the system. Advance in knowledge thus lies in the widening of systematic thought rather than in the discovery of isolated facts. At the same time, Bradley gave only qualified assent to the theory that knowing is creating; he thought it "a partial truth." Indeed, the statement that "man *makes* truth" is so erroneous, he thought, as to be misleading. What must be said, however, is that man's knowledge is participation in the Absolute's knowledge of itself. "The Reality," he explained, "was known always, and now [in me] its knowledge occurs. My contribution leaves it unincreased, and yet is indispensably requisite." That there is this human contribution to the totality of truth can, however, only be maintained if it be granted that there are degrees of truth, for no human knowledge can be complete and absolute in itself. Further, it is in terms of this incompleteness of knowledge (rather than of total ignorance) that human error is to be explained.

Realist Rationalism. Though Rationalism was supposed by many to have reached its definitive expression in absolute Idealism, there were certain Rationalist epistemologies that were not Idealist. Thus, contemporaneous with Bradley at Oxford, John Cook Wilson put forward a theory of knowledge that rested on a very close analysis of such concepts as knowledge, apprehension, judgment, statement, belief, thought, definition, inference, induction, and deduction. The members of the Cook Wilson school were critical both of Hegel and of Kant, finding their main inspiration in Plato and Aristotle, especially in the Platonic distinction between knowing and opining. Knowledge and belief differ in kind and not in degree, as the Idealists had taught. Both Cook Wilson and his follower H.A. Prichard insisted that knowledge is *sui generis*. A person knows that he knows, and he knows too that the experience of believing is a different experience. When he does know, he cannot then be in error. Thus, the epistemology of this school rests finally on the assertion of infallible knowing and is comparable in this respect to the epistemologies of Plato, of Aristotle, and indeed of Descartes, insofar as Descartes rested infallibility on the nature of the knowing itself rather than on the fact that God is not a deceiver. This Oxford school was in this respect essentially Rationalist but not Idealist.

Cook Wilson and the Oxford school

It can be said, too, of the most successful and most influential critic of British Idealism, G.E. Moore, that, though his "Refutation of Idealism" (1903) was the main factor in the decline of that school, the common-sense intuitionism of his writings ranks him still among the Rationalists. Indeed, a long list of post-Hegelian philosophers who rejected absolute Idealism and yet were Rationalists could be drawn up. Only two further instances are cited here. Prior to World War II, Edmund Husserl, the founder of Phenomenology, propounded a philosophy very different from both absolute Idealism and the Realisms of his day. The very strong Rationalistic basis of his theory is sometimes neglected, though he insisted upon it throughout. Knowledge in the strict sense is infallible certainty; it is insight into the self-evident, self-evidence being the mark of all sound knowledge. Logical law, which is so known, is in no sense an empirically based probability. Again, in the later writings of the U.S. philosopher Brand Blanshard (born 1892), while his earlier Idealism is abated, his Rationalist philosophy is still emphatically and powerfully maintained.

ATTEMPTS TO TRANSCEND REASON

In concluding this section, reference must be made to claims that man can transcend reason to achieve a higher knowledge. Plato's account of knowledge of Being and Kant's of the thing-in-itself are sometimes interpreted in this way, though safer examples are to be found in the epistemologies of Plotinus and Spinoza. In the 3rd century AD, Plotinus, the foremost philosopher of Neoplatonism, spoke of (1) an irrational soul, that which senses and imagines without ever using the faculty of reasoning.

Plotinus' views of the soul

But he also held that man, through intellectual self-discipline, may become (2) a rational soul and have true understanding of himself and of the world. Moreover, above the rational is (3) a soul in which *nous* provides immediate insight. This is human reason at its best. There is still a higher knowledge than *nous*, however—if indeed it can be called knowledge—viz., the soul's knowledge of God. How can man experience God? He does not see him, for God is no object standing over against the knower. Plotinus was led to use the metaphor of contact. Man touches God, and the touch becomes union. Toward the close of the sixth book of the *Enneads*, Plotinus described the ecstasy of this knowledge that transcends reason. It is not to be interpreted, however, as a rejection of reason; it is the outcome of disciplined thought and intellectual struggle leading to intuitive knowledge.

Plotinus' account of man's knowledge of God found a modern counterpart in Spinoza's *Ethicsa*. For Spinoza, however, God is not "beyond being" but is Being. To know anything as it is, not in isolation but as it relates to and depends upon other things, is already to begin to know the infinite. It is to understand and have rational knowledge of the Substance that is. Yet, insofar as this is knowledge of the general and universal features of Substance, it cannot be the highest knowledge, for the highest must be an intuitive and nonabstract knowledge of Substance or God; and this involves feeling as well as reason. It is an "intellectual love of God." *Ratio* thus passes into a higher stage of knowledge that Spinoza named *"scientia intuitiva."*

Irrationalism

In such epistemologies as these, Rationalism is accepted and an effort made to complement it. But there are other epistemologies that are critical of it: human reason, it is held, works through abstraction and generalization and misses what is individual and concrete; its scope is thus limited, and it is blind to important epistemological features. For man does know what is individual and concretely real. This is a point emphasized at the turn of the 20th century by the French intuitionist Henri Bergson, who was not antirational but rejected the view that true and reliable knowledge is to be found only by closing the mind to all possible procedures other than the rational. The truth is that discursive intellectual understanding needs to be supplemented by what Bergson called intuition. For example, clock time, which human intelligence has brought into being, needs to be understood in terms of the underlying intuition of real time. Again, rational argument pointing to determination in human life is to be tested by the intuition of free will. Furthermore, the prophetic intuition that God is love is of greater significance in religion than any logical demonstrations about God's nature. Human intelligence needs the supplementation and possibly the correction provided by intuition.

On the other hand, a note of impatience with Rationalism, if not of bitterness toward it, was present in an earlier writer, the Danish Christian Søren Kierkegaard, who flourished in the mid-19th century and who in his *Either/Or* (1843) and his *Philosophical Fragments* (1844) argued that the life of reason dehumanizes man; the intellectual fails to commit himself, so that, in a sense, he himself never exists; a man has to choose, to take sides, in order to be, and the intellectual tends to stay on the fence. In the second place, the intellectual assumes that nothing can be true that is contrary to reason and so rules out the "absolute paradox" of religion, particularly of the Christian religion. This same charge had been made earlier (in the 17th century) by Blaise Pascal, a scientist and religious philosopher, and indeed by all who accept the view that knowledge rests not on reason but on a fundamental act of faith.

Opposition to the domination of human life by intellectual disciplines has come, too, from those who see danger in the influence of modern science. The Neo-Kantians Wilhelm Windelband and Heinrich Rickert in Germany, for instance, opposed the doctrine that the naturalism and positivism of the sciences provided a true view of human life and sought to set out a new life philosophy.

As a final point it is notable that much so-called anti-Rationalist criticism is directed rather against the "understanding" than the "reason" in Kant's sense of these two terms. And, as has been seen, Kant himself, and other Rationalists, were only too well aware of the defects of the understanding.

IV. Empiricist orientation

ANCIENT GREEK AND MEDIEVAL EMPIRICISM

Classical philosophers. Empiricist elements are discernible in the Pre-Socratics and in Plato and Aristotle.

Pre-Socratics. The early philosophers of Greece were curious about the nature of the external world; in particular, they sought to solve the problem whether the apparently different objects in it are ultimately composed of one material—for instance, of water, air, or fire. Their concern was with what exists and not with the knowledge of it or with the mind that knew. They took it for granted that their method was to be observation plus speculation on what was observed. They are sometimes said to have been Empiricists. But, if they were, they were not consciously so, and it is questionable whether the description is appropriate, particularly to the Pythagoreans and the Eleatics. Nevertheless, such accounts as remain of, for instance, the philosophy of Leucippus, whom Aristotle said was the first to put forward the atomic theory, suggest that his opposition to Parmenides' monist theory of being rested on empirical grounds. The evidence of the senses does not support Parmenides' theory, and by this criticism Leucippus adjudged it false and put forward in its stead a pluralistic atomic theory, asserting the existence of a void between the particles of reality. Sometime later, Leucippus' atomic theory was defended by Democritus, perhaps the greatest of the physical philosophers. By then, however, Protagoras, Socrates, and others were grappling with epistemological problems.

Democritus himself is reported as affirming that sensation is due to the passage of atoms from outside objects through the sense organs into the soul, which, like all else, is material and consists of atoms. Though images are left in the soul, they are inevitably distorted in the process—and so is the "knowledge" thus gained. The ancient Skeptic Sextus Empiricus reported him as saying: "Sweet and bitter, hot and cold, and colour exist by convention only; in reality atoms and the void exist." This would seem to mean that any knowledge based on the senses is likely to be unreliable because of the condition of the human body and of those things that impinge and press upon it. Yet, Democritus, apparently, wished to affirm that the atoms and the void are fully known. (Aristotle's criticism of this claim was that no successful account of the knowledge of objects can rest on a purely Materialist philosophy.) Though an evaluation of the exact teaching of Democritus is impossible in view of the fewness of the extant fragments, it is interesting to find doubts being expressed from the first about the cogency of "information" gained through "distorting" sensory media.

Theaetetus of Plato. In dealing with sense perception, the most significant work in classical literature is undoubtedly Plato's *Theaetetus*, and it is a mark of his genius that almost at the beginning of the discussion of this problem he should have produced so penetrating a study. The *Theaetetus* is divided into three parts: the first (151d–187a), the part to be considered here, examines the theory that sense perception is knowing, the second that true judgment is knowing, and the third that true judgment, with the reason (*logos*) for affirming it, is knowing. Each of these three theories is rejected. The mathematician Theaetetus, a leading personage of the dialogue, defends the view that sense perception (*aisthēsis*) is knowledge. By sense perception he means the perception of objects, particularly their visual perception. Things are as they appear to the observer. This thesis is put forward as the position of Protagoras, the Sophist, and Theaetetus and Socrates sympathize with it. If a man now sees a patch of yellow in front of him, it is beyond all doubt and infallible that there is a patch of yellow before him. But the question remains whether this is infallible knowledge. If the object of true knowledge has any

Whether sense perception is knowledge

permanence and is one and the same for all, then seeing the yellow patch is not an instance of true knowledge. A wind feels warm to one person but cold to another, and it may be possible to argue that sense perception, while in a sense infallible, is yet always relative. Heracleitus had held that whatever is known is in a process of flux and is never permanent. Just now, and for a particular observer now, the wind is warm. The wind and he himself who now feels the wind to be warm are both changing ceaselessly, but for the moment it must be granted that for him the wind is warm. Such sense perception could then be said to be infallible knowledge of the truth.

It has to be asked whether this account of knowledge can stand examination, whether any and every man is to be the measure of truth, and indeed whether living things other than man might also serve. The theory would seem to hold that, since pigs and baboons presumably have sense perceptions, a pig can be the measure of truth. Again, a distinction is made between wise and foolish men on the ground that the former are more likely to know the truth and less likely to lapse into falsehood, but this distinction would be meaningless if what every man experiences, however foolish the man, is invariably the truth. This would appear to assume that the statements of the inexpert are as true as those of the expert, that, if doctor and patient disagree about the latter's illness, both know the truth equally well. Even though it be granted that in sense perception each and every individual is in a sense infallible, it is still reasonable to ask if it can be supposed that this is true outside sense perception. Surely the infallibility of sensation is no infallible knowledge of the truth. Truth is permanent: once true always true. But, if all things are changing, then what men wish to know would be constantly changing: nothing would stay from one moment to the next to be known. Furthermore, the observer who knew a moment ago will be changing. If this is indeed a world of constant flux and nothing more is to be said, it is a fair question whether there can be any place within it for permanent truth.

Sensation, truth, and change

Plato sums up. It is said that perception is knowledge in the strict sense, infallible knowledge. A man sees the colour that he sees, and no error can enter. Yet, it is a question whether this is what is meant by knowledge. Seeing a shade of colour is a very different thing from knowing something about a physical object. The former is the product of one sense only—vision. The latter involves more than all of the special senses can give. It involves a knowledge of the being of objects, that they exist; of the difference between them; of likeness or unlikeness. Such knowledge is given not through the operation of one special sense but through the power of the mind itself, which reveals a truth that holds for the future as well as for the present, since it is permanent. Strictly speaking, seeing a shade of colour, smelling a smell, or hearing a sound, in spite of their infallibility, are not genuine instances of knowing. In Plato's view, to find true knowledge a man must apprehend the Forms.

Importance of sense perception in Aristotle. Aristotle shared Plato's Rationalism even though he rejected his theory of Forms; on the other hand, he did not share Plato's general mistrust of sense perception. Some of Aristotle's most original contributions to learning were made in the field of biology, the outcome of his own close observation of nature and of the seriousness with which he took the information that the senses provide. Moreover, his father was a physician of Ionian stock, and the Ionians prided themselves on the accuracy of their observations. Sense perception, Aristotle thought, provides the starting point of learning. To survive, all living things must have first, powers of nutrition, but, in addition, they must have, those of sensation so as to be aware of objects and, in particular, to be able to distinguish between them. Living things need sensory powers in order to select what is nutritious, but these powers, too, provide experience, which—in the human case—can be stored up, compared with other experiences, reflected upon, and used to understand the world. "Without sense-perception there can be no learning and no understanding," for there would be nothing to reason about in respect to nature. Aristotle conceived of sensation as one faculty that specifies itself in the various ways of the five senses. In addition, however, the sensory faculty perceives "common" sensibles (common to two or more of the senses)—*e.g.*, being one or many, being of a certain shape, being in movement or at rest. These are known in sense perception but not through any one, and only one, specific sense. Further, a person may see a colour and at the same time incidentally perceive a concomitant, "as when a certain white object is perceived as the son of Diares." Today such a perception is frequently contrasted with the bare sensation of the white patch and said to involve recognition. Aristotle clearly counted both the perception and recognition of the physical object as a part of the knowledge that the senses provide.

Biological approach to sense perception

Nevertheless, he recognized, too, that sense perception is inadequate. He emphasized that sense perception is knowledge of the particular—the particular taste or colour and also the particular thing or person or the particular event. It is knowledge of the here and now, though it is true that a person's remembering the same particulars occurring over and over again may lead to an expectation and a generalization—for instance, that, given a certain occurrence, then a certain other occurrence will always follow. Aristotle was aware of the part induction can play in yielding general knowledge, but such knowledge is not completely certain; it is not necessary and universal knowledge. Human beings are capable of a necessary and universal knowledge that is true not for this moment but eternally. Though sense perception never provides such knowledge, reason does—not only in providing mankind with the ways of procedure, particularly the syllogism, which makes demonstrated, necessary, and universal conclusions possible, but also by providing the fundamental laws upon which the demonstrations rest. Aristotle agrees with Plato that the highest field of human knowledge is one into which sense perception does not enter; the most to be expected of sense perception is that it leads to generalizations that can be argued about rationally. But Aristotle would add that any conclusions produced by such arguments, insofar as they apply to the natural world, can and should always be tested by appealing to sensory experience. In *De generatione animalium* he remarked, for instance, that, though theories about the generation of bees are many,

> the facts have not yet been sufficiently grasped; if ever they are, then the credit must be given rather to observation than to theories, and to theories only if what they affirm agrees with the observed facts.

Aristotle was as clear as any Empiricist that the conclusions of reasoning, if they are about the world, are to be verified or falsified by the evidence of the senses.

Post-Aristotelian Empiricism. A word may be added about the two principal post-Aristotelian schools, those of the Stoics and Epicureans, particularly in view of 17th-century assumptions that these schools founded a thoroughgoing Empiricism. Not all of the extant fragments confirm this view, particularly in the case of the Epicureans. It is noteworthy that, in one account of Epicurus' views on the infallibility of sense perception, he is said to have held that what is seen in hallucinations and in dreams is true. This would seem to suggest that, by the infallibility of the senses, he meant no more than Socrates did in the *Theaetetus*—that is, that, if one sees a blue patch, one is undeniably seeing a blue patch—even if it be seen in a hallucination. But this is very different from claiming that sense perception gives infallible knowledge of the world. Cicero complained, possibly unfairly, that Epicurus failed to give due attention to deductive and intuitive rational knowledge and suggested that this made it easier for him to magnify sense perception.

The Stoics and Epicureans

Whatever be said of Epicurus, however, the ancient fragments make it clear that the Stoics were confirmed Empiricists. They held that the senses provide the soul with the first material for thought, the sensible "presentations"; up to this point, the soul is like a sheet of white paper with nothing written upon it. But this, in the Stoic view, is not the only source of presentations and is not reliable. In sensation the soul gives its assent. For the soul,

though analogous to white paper in taking the sensory impressions, is yet not actively passive. On the contrary, it receives the sensory presentations, deliberates upon them, presents conceptual material of its own, and after deliberation recognizes and firmly grasps the truth. Sensation of itself cannot provide such knowledge. It is notable that, in developing an epistemology, the Stoics and Epicureans were further hindered by the necessity of explaining human knowledge in terms of a narrow atomistic Materialism.

Medieval philosophers. Up to the time of Albertus Magnus, in the 13th century, the task of philosophy was primarily to justify the Christian faith by rational argument. This curtailed the growth of natural science. Albertus, a scientist as well as a theologian, was the first influential thinker to insist that, whereas the authority in disputes on the faith was the theologian, he was not necessarily the authority in more mundane inquiries. If the question was one of medicine, one should turn to Hippocrates and to Galen; if of physics, to Aristotle. Information in physics, Albertus held, comes first from the senses, and the development of the natural sciences comes best through disciplined observation.

This lesson was not forgotten by his most distinguished pupil, St. Thomas Aquinas. In the natural sciences, Aquinas acknowledged the authority of Aristotle, and his thought followed closely that of Aristotle both in the latter's account of sense perception and of its role in knowledge. But he did not follow Aristotle blindly. Thus, in his commentary on Aristotle's *De anima*, he sometimes openly disagreed—for instance, in the examination of Aristotle's doctrine of incidental sense objects. He did not disapprove, however, of the Empiricist strain in Aristotle's approach to nature or of the view that human knowledge in this field begins with and rests upon sense experience.

Yet, oddly enough, real advances in the natural sciences came not from the Dominican Aristotelians of the Albertus–Aquinas school but from their critics, the Oxford Franciscan Platonists, of whom the most important figure was Roger Bacon, a 13th-century friar, who owed much to his teacher, the influential Robert Grosseteste. The explanation of the success of the Franciscans is simple but also profound: they were mathematicians; they believed that the development of the natural sciences could take place only by applying mathematics to the findings of sensory observation. Experimental science was in part disciplined observation but also in part applied mathematics. Bacon's own experiments were of no great importance, but he had the genius to see the right way to proceed.

The movement in the 14th century to liberate the natural sciences from dominance by theology was furthered by skepticism about the adequacy of reason in metaphysics. The view that the universal (or general) concept known by the intellect was the true object of knowledge was now questioned; the true object is the real, and the real is the particular and individual. Concrete "thisness," as Duns Scotus termed it, is not to be ignored. William of Ockham regarded knowledge of the individual existent through sensory intuition as the foundation of knowledge. He did not deny the need for abstraction in order to help judgment, but it is erroneous, in Ockham's view, to suppose that abstracted objects or conceptions contemplated by the mind exist outside the mind; universals belong to the mind's abstraction and conception, not to the real world. It followed from this that a strictly Rationalist theology, concerned with the concept of a divine being, could not positively assert the existence of the divine being, because the concept is a creation of the human mind. Man has, says Ockham, no intuition of God (as one can now intuit this white patch before him). The being of God is revealed; it is not known either by intuition or again by abstract thought; it is a matter of faith. On the other hand, the being of a particular natural object is known intuitively in sense perception, and this fact completely justifies the Empiricist approach to the natural sciences, as contrasted with that of the Rationalist. Such was the solution of the problem of man's knowledge of God and of nature that finally emerged from the medieval schools, and it was a solution that eased the way of the Empiricist.

MODERN AND RECENT EMPIRICISM

Central to the modern discussion of Empiricism are the contributions of John Locke, George Berkeley, and David Hume. But three other slightly earlier figures cannot be ignored, viz., Francis Bacon, Pierre Gassendi, and René Descartes.

Francis Bacon's attempt to provide a method for science

Early forerunners. Francis Bacon, at the turn of the 17th century, expressed optimism about modern science in his *Advancement of Learning* (1605), in which he stressed the need for more experimental work and less speculation. In the *Novum Organum* (1620) he attempted to set forth the complete method for the new science. The proper approach to nature, Bacon held, is Empiricist rather than Rationalist. This means that the procedure is inductive, passing from the observation of particulars to generalizations, and not deductive. But Bacon is critical of those Empiricists who think their task completed when they have jumped from observations to any sort of theory. It must be "a gradual and unbroken ascent" ultimately to the most general axioms by "dwelling duly and orderly" on the observed phenomena. True advance demands patience and the greatest of care in view of "the dullness, incompetency and deceptions of the senses." Bacon had some important insights. In discussing negative and affirmative instances, for example, he held that "in the establishment of any true axiom, the negative instance is the more forcible of the two." Yet, his occasional insights do not add up to a complete theory. What can be claimed for him is that he taught Empiricists to question their assumptions and their procedures while continuing to reject the mere speculation, divorced of reference to experience, of the Rationalists.

The work of Pierre Gassendi

A contemporary of Descartes and a reviver of Epicureanism, Pierre Gassendi was influential in his day in France and was not without influence elsewhere, particularly in England. His scientific contribution was limited, largely because of his lack of mathematical understanding. He was, indeed, suspicious of demands for complete mathematical certainty in the natural sciences, for the observations of nature and of life provide a foundation for probability only. "I live," he said, "from day to day, accepting nothing which goes beyond the limits of probability." His fame rested rather on his influential advocacy of a new metaphysic based on the Epicurean physics of atoms moving within a void. He advocated, too, the moral teaching of Epicurus and his epistemology. Gassendi was charged in his day with being, in his theory of knowledge, a sensationalist (who reduced experience to sensations). It is true that he believed that there is "nothing in the intellect" not derived from the senses and that he asserted, with Epicurus, that the senses never err (the Theaetetan infallibility of bare sensation). But this does not make Gassendi a sensationalist, for he also admitted readily that the senses are frequently deceptive and are then to be corrected by reason, "which is superior to sense."

In his lifetime Gassendi, though much the lesser man, was regarded as Descartes's rival. And, indeed, the standpoints of the two philosophers were very different. Descartes sought certainty in metaphysics, in physics, in psychology, and in morals and believed that man has the powers to establish it if he chooses, whereas Gassendi thought that probability based finally on sense experiences is sufficient for human needs.

Descartes, in his *Principia Philosophiae* (1644), gave an account of the nature of sensation, showing how neural movements affect the brain and, through it, the mind. The information thus provided is of great utility in a practical sense, but none of it is wholly reliable. It tends to be deceptive and creates deep prejudices that are inimical to a true knowledge of the world. Descartes emphatically rejected the Empiricist view that there is nothing in the intellect that was not first in the senses; it contradicts his own Rationalist account of knowledge. The most that sensation can do is to provide material for

thought; it can suggest, for instance, the notion of extension, which the intellect can then handle geometrically. It suggests too the existence of an external world of physical objects. Sometimes an object taken to be external is not in fact external, as in the case of illusions; but Descartes argued that doubts about the existence of an external world on this ground can be dissipated by noting that the perception of it is provided by a veracious Creator who has also provided man with a means of testing the validity of the perception—by the criterion of clear and distinct ideas. Thus, sensation is useful and, in a limited way, illuminating; but for full illumination one must turn to reason and intellect—for instance, to the mathematical analysis of the external world. Given this illumination, observation and experiment can be exceedingly valuable, but without it they do not provide an adequate basis for human knowledge.

Classical British Empiricists. The classical sources for the study of Empiricism are the writings of three British philosophers—John Locke, George Berkeley, and David Hume.

Influence of Gassendi: the "clean slate"

John Locke's new way of ideas. John Locke was said by Leibniz to have been "established in the system of Gassendi" and was viewed on the Continent in the early 18th century as the champion of the Gassendists. He was not so regarded at home, however, for the influence of other (English) movements upon him was too obvious. Leibniz clearly had in mind the conflict between Gassendi and Descartes and held that, in this conflict, Locke was on Gassendi's side. This judgment, however, needs qualification, for, though Locke was suspicious of Cartesian speculation and theories, yet his was not a simple choice between Gassendi and Descartes. Locke had cooperated with leading English scientists, such as Robert Boyle, a pioneering physicist and chemist, and Thomas Sydenham, an eminent physician, and had studied the work of Isaac Newton. Locke had learned from these scientists that the new science was not merely the product of a better empirical observation of nature; it rested, too, on a successful application of mathematical techniques. This latter element had not been emphasized in the teaching of Gassendi.

If this qualification be made, it is then safe to acknowledge Locke's debt to Gassendi. In the *Essay* he set out to consider "the original, certainty, and extent of human knowledge, together with the grounds and degrees of belief, opinion, and assent." In Locke's treatment of the origin of knowledge Gassendi's influence is very obvious. At the beginning, Locke affirmed, the mind is as "white paper, void of all characters." It lacks innate ideas, though it does possess innate powers. It is "a clean slate." Ideas come to be written upon it through sensation of the external world and reflection (introspection) of the internal world; "these two are the fountains of knowledge, from whence all the ideas we have, or can naturally have, do spring." Locke showed in Book I of the *Essay* how, as against the Cartesians, there are not innate ideas in the mind, and his main purpose in Book II was to establish that all of man's ideas originate in sensation or reflection. This does not mean that all ideas, say of the external world, are themselves sensed, for, from what is sensed, other ideas may be derived. It means instead that, if a man lacked all five senses, if he had no sensory experiences, he would have no knowledge of the external world; for such knowledge, sense perception is essential. And this is Empiricism.

Locke's view of the representational function of perception

Locke acknowledged that there are difficulties in the theory. The senses, he explained, "convey into the mind several distinct perceptions of things; and the mind is thus stocked with ideas." How is this statement to be interpreted? Is "having an idea" itself a knowing? When Locke spoke in Book IV of knowing with certainty, knowing in the strict sense, this is not to be identified with the having of ideas in sense perception but with intuiting that ideas are, or are not, related in certain ways. Nevertheless, he did find it necessary to admit a "sensitive" knowledge that is more than merely probable. A person may imagine, for instance, a cat with webbed feet, but he would not then claim that it existed; if he saw such a creature, however, he would make this claim. A man cannot deny that the sun that he now sees exists. Sense perception is a knowledge of existents, and Locke never thought otherwise; yet, this is difficult to accord with what he said about the representative function of sensory ideas.

The senses, he explains, "convey into the mind several distinct perceptions of things." There are three elements that should be distinguished: (1) the mind, (2) the physical thing, (3) the idea or perception. The mind in Locke's day was generally held to be a substance that has real existence and that is distinct from a physical substance; on the whole, Locke accepted this view, though he was not prepared to rule out the possibility that God may have "super-added" to man's material body "the faculty of thinking." This would necessitate that there exist two substances in the world, material and immaterial, but not that man is a dualism of two substances. The second element is the physical thing known—for instance, the table. It is necessary to distinguish, however, between the mediately and the immediately known; what is known in sense perception (the table) is known mediately. For what one is aware of immediately is the third element, the idea or perception—*e.g.*, the brown colour of the table or its smoothness. A complex of simple ideas represents the table, and it is through being immediately aware of this complex that one becomes mediately aware of the table. The purpose of this doctrine is to explain how, in seeing a table, one may sometimes see its colour to be brown and sometimes some other colour. Things are not always what they appear to be, but men, nonetheless, know them through their appearances. Locke did think that in the case of certain qualities, which he called primary—*e.g.*, solidity, being extended, or being in motion or at rest—the observer could be sure that the ideas resemble the real qualities; but this is not so in the case of ideas of secondary qualities—*e.g.*, colours, tastes, and smells. Yet, he did not doubt that men do know physical things through sense perception.

Berkeleian Idealism. There are acute difficulties in Locke's account of sense perception as was made clear by contemporary critics, one of whom, Bishop George Berkeley, an immaterialist of the early 18th century, accepted the view that knowledge of the world originates in the having of ideas in sense experience. But he did not accept Locke's next step—*i.e.*, that these ideas represent qualities or things in the external world; on the contrary, he regarded representationalism as a millstone around Locke's neck. An idea may possibly represent another idea: *a* could be like *b* (in which *a* and *b* are ideas), but *a* can never be like a material thing and cannot represent it.

Immaterialism

Berkeley rejected the widespread assumption that a material world exists beyond the veil of ideas, but, even were it true, Berkeley would still hold that having ideas would throw no light on its nature. Further, Locke's distinction between ideas of primary and ideas of secondary qualities does not help him, for the very same arguments that show that ideas of secondary qualities cannot be representations of the qualities of material things show, too, that ideas of the primary qualities cannot be such representations. Yet, for the Empiricist the origin of knowledge lies in the gaining of ideas; and it would appear to follow that the knower cannot really know his world, because, apparently, the only means of doing so is through the discredited representations. Berkeley's answer is that ideas do not represent a material world and that this answer need not disturb the would-be knower dependent upon ideas; for there *is* no material world. One reason among several that Berkeley had for adopting an immaterialist philosophy—and an important one for him—is that it finally solves the difficulties of the representationalist theory.

If, however, a person does adopt immaterialism, there is a problem of what is to be said in a positive way about ideas, since the theory that they are representations is rejected. Berkeley's solution lies in the adoption of Idealism, the theory that what exists and what is real is solely mind having ideas. Ideas exist as being perceived by a mind, their *esse* ("to be") is *percipi* ("to be perceived"). Berkeley

Idealism: "To be is to be perceived"

himself acknowledged that this system of Idealism might be criticized on the ground of its subjectivity. If a person claims that what is is what he perceives, then what is would be subject to his whim at any moment, and there would be no objectivity and no order in the world. But this, Berkeley replied, is to fail to grasp the nature of the process of having ideas. When a man opens his eyes in broad daylight, he is constrained to see what he does see. What he sees is not determined by him. In imaging or daydreaming, he is free to conjure up what idea he chooses, but, when he sees something, then what he sees is in no way dependent on his will. Nonetheless, the idea seen is dependent on mind, though clearly not on his mind; thus, when he sees it, it must be given to him by some universal mind able to produce all these ideas. They are certainly not of the viewer's willing; they are lively and distinct and come in such an order and with such a coherence that one soon gets used to their regularity. The rules followed are, in fact, the laws of nature, which are thus revealed to man not by reasoning nor by intuitively grasping an objective necessity but by experience. Knowing the world of nature is a process of becoming familiar with the established ways "wherein the mind we depend on excites in us the ideas of sense." In this way, one knows objective law, and Berkeleian Idealism is not to be equated with subjectivism.

Impressions in Hume. Whereas reflections on the difficulties in Locke's Empiricism helped to lead Berkeley to Idealism, they led the Scottish philosopher David Hume, of the ensuing generation, to Skepticism—though in both cases other factors were also at work. Berkeley had argued that, beginning with Lockean Empiricism, a coherent system of philosophy could be maintained but only if it were an Idealist system. Hume, however, could not accept the view that the physical object could be identified with an idea. He was also more conscious than Berkeley of the deep skepticism present in Locke's own philosophy. The 18th century was Rationalist; but paradoxically, it acknowledged Locke, the Empiricist, as the source of its Rationalism. Locke's appeal to the life of reason, as contrasted with that of emotion and passion, had been very effective. Furthermore, there were, in fact, Rationalist elements at the heart of his epistemology, since he thought man capable of absolutely certain knowledge by rational intuition and by it alone. Nonetheless, man's link with the existing physical world, he held, is through ideas, and the correctness of their representation is suspect, at least in the case of ideas of secondary qualities. As a consequence, Locke explicitly affirmed that man's knowledge of the physical world does not extend to that certainty that would permit it to be called scientific. Observing the phenomena gives useful hints and clues; but "certainty and demonstration," he held, "are things we must not, in these matters, pretend to." This is the skepticism present in Locke, a skepticism regarding the possibility of exact and certain knowledge of the natural world; and it is the skepticism that Hume developed.

Skepticism

Hume himself acknowledged that certainty is possible in a very limited field, viz., when the ideas are clear and necessary and the relations between them can be grasped by reason. Such is the case, for instance, in arithmetic and algebra. Others had argued that the causal principle that whatever begins to exist must have a cause of existence is also known rationally and that knowledge of it provides a base for the objective, necessary science of nature. Hume, in a brilliant argument, claimed to refute this view on Empiricist grounds.

To understand the refutation, however, a student first needs to understand Hume's earlier account of ideas. Knowledge begins with "impressions"—this is Hume's term—and impressions come with "a force and liveliness" that make them unique. "Ideas" are in turn "faint images of these" impressions. This blue colour is a sensed impression; the perceiver can later recall the colour, and then it is an idea. An impression and the corresponding idea can be simple or complex, and, in the case of a simple impression, say the taste of a pineapple, the corresponding idea exactly resembles the impression. If someone thus speaks of a simple idea, it is perfectly justifiable to ask him what simple impression it exactly resembles. If he cannot answer, his idea lacks clarity. In the same way, it is reasonable to ask to what simple impression the idea of a causal relation corresponds. In view of Hume's phenomenalism, one answer is impossible for him: he cannot speak of an impression of a physical thing exerting power and so causing change. Certain impressions may suggest the presence of a continuing existent, but such an existent is never experienced. This is not to say that it does not exist but that man cannot know it, for he is confined to the phenomenal. For him it is the perceptions that are real, and, since "nothing is ever present to the mind but perceptions," a knowledge of what exists beyond these is impossible. It is true that men talk of a physical object; but what they have in mind then is "a bundle of perceptions," though this bundle may recur so constantly in their experience that they tend to think of it as a continuing identity. This is a part of their experience; yet, it does not establish the existence of a continuing physical object in the world. Thus, it makes no sense to say that a person has an impression of one real thing acting on or affecting another.

Impressions and ideas

The philosopher must speak in terms of simple impressions. But, even when he does so in the present context, he is still baffled. For apparently no one has an impression of causing. The observer does not perceive p causing q to move away. Strictly speaking, the most that a person sees is p moving in q's direction; he next sees the contacting, and he then sees q move away. So far, there is no knowledge of a causal relation. But, when he sees this sequence occur frequently, by an association of ideas, which Hume finds to be an important principle of the imagination, he does tend to assume that p causes q to move away. The actual simple impressions present are (1) contiguity (p and q are contiguous) and (2) priority in time (p's movement being prior to q's). Though these impressions alone would not give the feeling of a causal relationship, these plus the regularity of the experience and the constancy of the conjunction could do so. The expectation of an event, given the prior conditions, is strong and grows in strength with the increase in experience, until, finally, the observer feels that one *must* cause the other. There is a necessity in the causal relation; yet, the necessity is not now objective. It is a subjective necessity only, a necessity for the imagination controlled by the association of ideas. The observer must believe that p causes q to move.

Causation and belief

The argument throws light, Hume thought, on the nature of belief. Belief is having a present *impression*, for instance of p moving, and associating with it an *idea* of q's moving, so that, before one sees q moving, one believes it will move. In this sense, one may be said to know the causal relation and only in this sense. Knowledge of cause and effect, therefore, is not a rational intuition nor a rational demonstration, but a belief determined by imagination and its laws. The much-prized sciences are thus not rational; they in no way rest on rational insights into the objective nature of the world. They rest, instead, on an experience of impressions in a certain constant order and on the beliefs that the imagination is induced to bring into being as a consequence.

Clearly, one cannot accept the Rationalist account of the knowledge of the causal relation and still remain an Empiricist. The Empiricist may concede the possibility of rational intuition, say, in mathematics but not in the natural sciences nor in psychology. The physical existent is what it is, but men do not know what sort of being it has. The theory of representative perception is wholly unnecessary and solves no problem, for what is phenomenal can never represent what is not phenomenal. The principle of causal relationship belongs to the phenomenal world and makes possible a science, but it is only a science of which the necessity is subjective. Finally, knowledge of substantival continuants—*i.e.*, of identities (whether physical objects or persons)—rests solely on the experience of a series of impressions gained on many occasions in a regular order and on the "bestowing of an identity" on the perceptions, thus "producing the fiction

of a continued existence." Whatever the nature of its true existence, a "physical existence" is experienced by man as a recurring series of impressions. Personal identity, too, is explained in the same way. The abiding personality, like the continued existence of external things, is an "illusion of the imagination"; yet, an illusion without which men could not live their lives.

Summary of classical Empiricism

Assessments and recent expressions of Empiricism. The half century from 1690 to 1740 thus witnessed a bold attack on the problems of epistemology. Starting from the dictum that "there is nothing in the intellect which was not previously in the senses" and taking this dictum seriously, conclusions were drawn that were both brilliant and disturbing. Locke interpreted it to mean that a complete reduction of all ideas into simple ideas was feasible and that this reduction would help to show that the most complex ideas had reference, however remote, to what really existed; for they were reducible to simple ideas, and these, he assumed, resembled the real things. Contradictions in sensory experience, however, caused him to doubt whether the ideas of secondary qualities did resemble the qualities of real things, but at least the ideas of primary qualities resembled them. Further, he failed to reduce all complex ideas—for instance, the idea of substance—into simple ideas. Berkeley contended that Locke in his theory of knowledge had attempted the impossible. Complete skepticism was inevitable if knowledge was held to depend on the establishment of correspondences between ideas in the mind and independent material things outside the mind, and the distinction between ideas of primary and ideas of secondary qualities could not save Locke. What was necessary, in Berkeley's opinion, was a revolutionary change of outlook; it was necessary to deny independent matter and assert that what existed consisted solely of mind and ideas. Hume agreed with Berkeley's diagnosis of Locke's failure but rejected the remedy that he proposed. Idealism was unacceptable. Human experience was of the phenomenal only, and no one could say what lay behind the veil of phenomena. Instead of Idealism, Hume offered a phenomenalist account of human knowledge that enabled one to speak significantly of continuing identities of things and persons and of changes in and between them, but it was a knowledge that arose from familiarity with the phenomena and from inductions based on this familiarity. As for sure knowledge of what exists other than phenomena, Hume admitted that he had none.

18th-century assessments. That these consequences of Empiricism were disturbing is clear from many of the writings of the time, two of which may be briefly mentioned. Richard Price, a British moralist, was quite emphatic on one point: that no account of sense perception can be given in terms of sensation alone; for such notions as substance, space, and duration, which are not given in sensation, show the presence of understanding, "the faculty that discerns truth." The understanding is itself "a spring of new ideas," and these ideas are made use of in sense perception and in the comparison and relating of ideas.

Assessments by Price, Reid, and Kant

Another—and fuller—examination of the Empiricists is that of Thomas Reid, founder of the Scottish commonsense school of philosophy. Reid criticized the failure of the Empiricists to distinguish between sensation and perceiving. Sensing is having an impression; but perception is

> first, some conception or notion of the object perceived; secondly, a strong and irresistable conviction and belief of its present existence; and thirdly, that this conviction and belief are immediate and not the effect of reasoning.

More than all he criticized Berkeley and Hume for allowing their speculations to lead them to positions (immaterialism and total Skepticism, respectively) that are obnoxious to common sense.

On the Continent, while recognition of Locke was widespread, few recognized the significance of Berkeley and Hume. Kant, however, did appreciate the importance of Hume's teaching in epistemology and found it necessary, as a consequence, to make radical changes in his Rationalism. In seeking to understand the world, the knower must first wait on experience for he is dependent upon sense-data. He does not construct the concepts that he thinks from nothing. And there is no possibility of his "calculating the universe" in the extreme Rationalist sense. On the other hand, no attempt to reduce intellectual concepts into sensibles can succeed. For, whereas human knowledge is admittedly a matter of ordering the manifold of the senses, it can only be ordered when principles are given that have been gained not in experience but rationally and a priori. Pure Rationalism of the Leibnizian kind and pure Empiricism of the Humean kind both fail. Kant attempted to frame a theory of knowledge that has both Rationalist and Empiricist elements, but he found himself encumbered by the difficulties of both schools. His partial acceptance of Empiricism, for instance, involved him in a representationalist interpretation of ideas, which contributed to his difficulties with an unknowable "thing-in-itself" and an inexperienceable "real self."

Culmination of sensationalist thought in Condillac and the Materialists

The emphasis on sensation in Locke, Berkeley, and Hume influenced various aspects of 18th-century life. It had a profound effect upon the ethical teaching of the age, leading through the moral-sense theory ultimately to Utilitarianism. On the more theoretical side, the extreme teaching of the French sensationalist Étienne Bonnot de Condillac, influential in France and Italy, is significant. He argued that the full life of the human mind emerges out of the power of sensing. Condillac is noted for having imagined a statue that is lifeless until one first superadds to it a sense of smell. If the statue is genuinely able to smell, then it will soon come to desire and to will; if, further, the sensation of touch is superadded, then will follow a consciousness of a real external world, and so on. Given sensation, all the rest of human experience follows. In England the Materialist Thomas Hobbes, a political philosopher, had argued a century earlier not merely that conscious life is a development of sensations but also that the explanation of sensation is a physical one, the affecting of physical organs by physical motions. David Hartley, an 18th-century physician and philosopher, explained the associationist theory, which Hume made central to his epistemology, in terms of brain vibrations, whereas his disciple, the clergyman Joseph Priestley, discoverer of oxygen, made clear the Materialist implications of this theory. These were emphasized, too, by several other French Materialists of the 18th century, one of whom used the analogy of digestion and spoke of the brain as secreting thought. Many of these writers were medical men; their plea, in effect, was for a purely empirical science of physiology, on the one hand, and of psychology, on the other. They found support in Condillac's sensationalism.

Empiricism and Positivism. Along with sensationalism, another offshoot of the Empiricist school was the Positivism of the 19th century. Condorcet had already expressed the spirit of the new movement, and his work influenced Auguste Comte, the founder of the Positivist school. Comte lectured privately in Paris in the 1830s and published his *Cours de philosophie positive* (1830–42), which set down the Positivist principles. He claimed that the development of science, on the one hand, and the Industrial Revolution, on the other, had given birth to a new culture that derived its principles neither from theological dogmatism nor from abstract metaphysics but from science. The methods that had worked so well in natural science needed to be applied to a new science, sociology. In Britain, Comte had an admirer in John Stuart Mill, whose Empiricism and Positivism are apparent in all of his writings, as they are, too, in those of his father, James Mill, and of Jeremy Bentham, known—as were the Mills—for his Utilitarian ethics. John Stuart Mill objected, however, to what he took to be an illiberal and doctrinaire feature of Comte's teaching, viz., that, once sociological laws are established by scientific methods, they are then—like laws of physics—not to be questioned. Mill objected on moral grounds and also on epistemological grounds, because the sciences, as he thought, provide men with beliefs only and never with absolute certainties. Mill accepted the fundamental epis-

Comte, J.S. Mill, and "empiriocriticism"

temological positions of Hume—*i.e.*, the rule of induction, the regular sequence theory of causality, the greater ontological significance of the particular, and phenomenalism. (Mill held that a physical object is a permanent possibility of sensation.)

Mill's doubts about Comte's teaching were shared by Positivists later in the century. Toward its end, in the German-speaking world, a school arose that was Positivist in outlook but critical of the earlier naïveties of the movement. These men questioned the earlier uncritical assumptions of Comte as to the reliability and universal applicability of scientific method. The school included in its ranks Karl Eugen Dühring, author of a Materialistic "philosophy of reality," and Richard Avenarius, the founder of "empiriocriticism"; but the most prominent figure was a physicist, Ernst Mach, at the University of Vienna. For Mach, who later taught philosophy, science is first and foremost a matter of economy, "the completest possible presentation of facts with the least possible expenditure of thought." The facts are known in experience; science is not fact-finding but fact relating; it does not claim to present the absolute truth about the world. Its so-called laws are not absolute but are at best "limitations of possibilities." Thus, the view of the early Positivists that strict adherence to scientific method guarantees truth was not shared by the later Positivists; yet, they were as emphatic as their predecessors had been that metaphysics is to be avoided. The outcome of their thinking was a fairly general Skepticism.

The Vienna Positivists

In the meantime, the pioneers of mathematical logic, Gottlob Frege, Bertrand Russell, and Alfred North Whitehead, had established that an element of necessity is present in mathematics and logic, and this helped a group of new Positivists in Vienna—among them Moritz Schlick, Rudolf Carnap, Friedrich Waismann, and Otto Neurath—to put forward a Positivism that rests on a sure logical basis. It was argued that some statements can be shown to be true or false by appeal to experience and others by appeal to logic and that both classes of statement are significant. But allegedly factual statements about objects beyond experience—*e.g.*, metaphysical and theological statements—cannot be held to have meaning. This new philosophy was advanced forcibly in *Language, Truth and Logic* (1936), by a British Positivist, A.J. Ayer. Many criticisms have been made of Logical Positivism, the most damaging being that of A.C. Ewing, a Cambridge philosopher, that it would appear to make any philosophic statement whatsoever (including the statement of Logical Positivism itself) meaningless. Moreover, granted that the necessity of logic and mathematics is that of tautology, which, however, involves exact synonymy, W.V. Quine, a U.S. logician, questioned the assumption of analysts and tautologists that there is such a thing as complete synonymy. Thirdly, contemporary Positivists disagree about what experiencing is, for some think of it as the observation of colours, tastes, and sounds, others as that of physical objects, and this ambiguity still remains to muddy the waters of Empiricist arguments.

Empiricism and Pragmatism. Finally, in discussing Empiricism and Positivism, reference must be made to Pragmatism, a movement that flourished, mainly in the United States, at the turn of the 20th century. Its leading figures were Charles Sanders Peirce, an engineer and logician, and William James, a psychologist and philosopher. James's charm, liveliness of style, and generosity of spirit made him the better known of the two. In his suspicion of speculative metaphysics, he was a Positivist; thinking is always a means to an end, he held, and the end is practice. He argued that "truth in our ideas means their power to 'work,' " and that "any idea upon which we can ride, so to speak . . . is true *instrumentally*." Replying to outraged critics in *The Meaning of Truth* (1909), he argued that this view does not imply subjectivity; that, on the contrary, his Pragmatism is throughout objective and empirical, for it rests on the awareness of "brute fact"—*e.g.*, "that this paper is white." But brute-fact knowledge is limited and is to be eked out by way of conceptual knowledge; the latter, however, is largely dispositional in character, involving "a procession of mental associates and motor consequences."

The epistemological features of Pragmatism were given a fuller and more adequate treatment in the highly original writings of Charles Sanders Peirce, a gifted logician and painstaking analyst whose interest in practice was less dominating than that of James. He was less concerned, too, with the reality of individuals and concrete particulars and not at all attracted by James's nominalist philosophy of "radical Empiricism." The link between the two lies in the adoption by both of the pragmatic method that Peirce and James, along with other friends, thought out together in Harvard in the early 1870s.

For Peirce, man's knowledge begins with what he observes through his senses and what he is told by others. He may hold tenaciously to what he has so gained. But doubts inevitably occur, and he finds his senses occasionally deceptive. The authority of others, particularly of those whom he deeply respects, may assuage these doubts, at least temporarily. Other philosophers have held that, if the doubts return, he must then appeal to his own powers of intellectual intuition. But Peirce denies that human knowledge is ever intuitive in this sense; all of the evidence shows that it proceeds in another way—not through intuition but through thinking.

Pragmatism and science

Many would agree that the thinking that has given the most satisfaction to mankind (in the sense of enabling man to ward off doubts and so to believe) is scientific thinking. Consequently, the best approach to the understanding of human knowledge is a study of the method and purpose of science. Scientific thinking, however, is conceptual thinking. Hence, some account must be given of the concept. This account provides the core of the Pragmatisms of both Peirce and James. Peirce defined Pragmatism as follows:

> Consider what effects that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object.

A person may conceive of a diamond—*i.e.*, he may think of what it does and conceivably can do. The concept, then, is the conception of these actual and conceivable effects. Thus, part of the concept of diamond is of an object that can scratch a glass surface. Scientific thinking, by the method of hypothesis (or "abduction"), originates new concepts and also verifies them in experience and thus enriches the conceptual content of scientific thought and establishes its truth or falsity. The truth is what survives the verification processes of science, not of one scientist but of all. "The opinion, which is fated to be ultimately agreed to by all who investigate, is what we mean by truth." The investigation by the community of scientific inquirers, "if continued long enough," is destined to attain truth. Such is the fundamental nature of science and so of human knowledge at its best. But this account inevitably provokes the question: "How long is long enough?" Peirce seems aware of the possible criticism and confesses that his philosophy is "fallibilist," since "perfect certitude and exactitude" can never be attained.

A century after its birth, no school of Pragmatism survives; but its influence, blended with that of Positivism, remains. John Dewey's instrumentalism owed much to James and Peirce, and their influence is felt, too, in other circles—*e.g.*, in the operationalist Perry Bridgman, and in operationalism generally, and in the conceptualistic Pragmatism of C.I. Lewis. It is notable, too, that there are elements of Pragmatism in W.V. Quine's *Word and Object* (1960) and in his papers.

V. Conceptual thinking

CONCEPTS

In the two schools of epistemology considered up to this point, the role of reason, on the one hand, and that of sense perception, on the other, have been stressed. It would be acknowledged by both schools, however, that there are other factors in the situation that need attention. Both, for instance, would acknowledge that thinking is one such factor. This is not reason; neither is it sense

experience, but it is, nonetheless, part of the whole cognitive experience. Admittedly, this statement would be disallowed by Hegelian Idealists, because their Rationalism admits no final distinction between thinking and knowing: thinking is the self-development of an ideal content, and this is knowing. The more usual view, however, is that thinking may lead to knowledge, may help to extend knowledge—though it need not do so. The question that arises is what "leading to knowledge" and "extending knowledge" signify in this context.

Both reason as intellectual apprehension and sense perception are intuitive, but thinking proceeds mediately through concepts. Thinking involves having concepts. This has been interpreted to mean that concepts are the objects of thinking. Locke took this view at the outset of modern epistemological speculation. For him, the term idea covered concepts, but it also covered the impressions of sensation—*e.g.*, the colour now sensed and the impression of feeling, for instance, of anger. He used the term, he said, "to stand for whatsoever is the object of the understanding when a man thinks." Thinking is directed to an object, and any such object is an idea. The view that the idea is the object of thinking, when widely applied, caused many difficulties in subsequent discussion. To guard against these difficulties, it is necessary, in the first place, not to confuse an impression with a concept and, in the second place, not to assume without question that a concept is an object of thought.

Whether concepts are impressions or objects of thought

A concept is not an impression. A person may now see a red patch, but this red patch is not a concept. Furthermore, this particular red patch, when remembered, is still not a concept. Usually one thinks of a concept as pertaining to a stage of mental activity higher than that of merely sensing or even remembering. It is something intellectual as involving language and as contrasted with the barely imaginative. Yet, it does not seem to be the case that having a concept invariably presupposes the ability to speak. It is true that the phrase having a concept may be defined as an experience possible only for a person who can speak, but this definition is arbitrary. A deaf mute who has never learned to speak may have seen many storks. But seeing this particular stork or remembering a particular stork is not having a concept. Yet, the deaf mute may think of storks, though none is now visible and though he is not now recalling in memory any particular stork. It is reasonable, then, to suppose that the deaf mute could have the concept though he has no language. It is not unusual to use the word concept in this way, and anyone who does so is implicitly rejecting the view that all concepts are necessarily verbal.

But most concepts are verbal. This statement requires interpretation as to whether it signifies that the concept is the object of conceptual thinking and that the word is the concept's name. Locke seems to say something of the kind. The concept (for Locke, "the general idea") that is the object of thinking has a name, viz., the general word. Here, two questions arise. The first is how the name functions. Presumably it functions somewhat as a proper name, designating the object; but this is surely wrong. The second is whether it is correct to speak of the concept at all as an object. It is admittedly not an object in the sense in which this table is an object, for it is in the mind; it is an internal accusative, what the thought is directed to within the mind. One can immediately object, however, that no one, in his thinking, is ever conscious of the existence of such an inner object or accusative. To this claim it may be replied that it must, nonetheless, be there. It is argued, for instance, that, when a person thinks of a transcendent object, he must be thinking of a concept, since the object transcends all experience. Or, again, when he translates the English word liberty into the German *Freiheit*, there exists the concept as well as the two words. In this view, there is a world of material existences and also a world of psychological occurrences; but, in addition, there is a third world, that of concepts existing as objects of the mind.

If it be felt that this multiplication of worlds is to be condemned, an alternative theory must be sought. One such alternative is to regard the concept not as a mental entity but rather as a capacity. This view stems from David Hume's account of abstract ideas, in which he suggested that the abstract idea (as contrasted with impressions and simple and compound ideas) should be thought of not as an entity in the mind but as a disposition of the mind, or, in his own words, as a propensity. This suggestion provides the basis for the alternative theory. The philosopher should speak not of concepts but of having concepts, and his best way of determining whether a person has a concept is not by asking him whether he finds such a concept in his mind but, for example, by observing his behaviour. When he observes heavy clouds, he looks for his umbrella; and his response is the same if, when he is on his way out of the house, somebody tells him, "There are heavy clouds about." Familiarity with the English language, on the one hand, and with the climate, on the other, has disposed him to respond to the statement in that way. In either instance, he has a concept of heavy clouds presaging rain.

Whether concepts are capacities or dispositions

Having a concept is having a disposition or a capacity. This seems to be a more hopeful theory than that concepts are mental entities, but it, too, is not without difficulties. For, in the first place, the investigator can never directly study a disposition but only the consequences of being so disposed, and this makes for vagueness. Secondly, the mind is said to frame concepts; but, if these concepts are capacities, it must be questioned whether it makes sense to speak of the mind framing its own capacities. Finally, it is reasonable to ask whether it really makes sense to say that the concept is just a capacity, whether it would not be better to think of it as a capacity or capacities along with a certain content or core—which is not to be thought of as an internal accusative. Thus, a man might think of the concept in terms of Frege's account of the reference of the logical predicate term. The predicate term, he held, is a general word (or words) that is to be understood in terms of a function of which the values are truth values. Something is said about an object *a*, and the schematic logical form of the statement is "*f a*," in which *a* stands for the logical subject and *f* for the concept. The question, then, is whether there is in this function "*f* . . ." a clue to the nature of the concept in verbal thinking. If, in the open sentence ". . . is a tyranny," a person can use this locution significantly, if he possesses the necessary empirical and linguistic dispositions, then he has the concept of tyranny. The open sentence is not itself a disposition, but a significant use of it presupposes the presence of dispositions in the mind that are essential if this locution is to be used significantly. There is a core, which in this case takes the form ". . . is a tyranny," that has dispositional accretions around it that enable the locution to be used significantly. This approach would explain the nature of concepts in verbal thinking. A like theory could be worked out for concepts in nonverbal thinking. In respect to nonverbal thinking, having a concept would be explained in terms of accretions of dispositions and capacities around images; in respect to verbal thinking, it would be explained in terms of accretions around general words or open sentences.

LANGUAGE

The epistemologist is much concerned with language and, particularly, with the use of general words in conceptual thinking. But, beyond general words, all aspects of language are of interest to him. Much of what is known by each person is known by listening to others, by speaking with them, and by reading what they have written. Language is, first, a means of communicating between men. For the most part, the communication proceeds through statements that purport to state what is the case. But language may be used also to ask questions, to command, to cajole, and to express one's feelings. In addition, one uses it in thinking, and this latter use is the philosopher's main interest. Yet, this approach may give his study of language a slant that may hide a good deal from him.

Logical versus broader uses of language

For instance, exact thinking is logical thinking, and the philosopher may tend to think of language generally as a completely logical instrument obeying logical laws. He will think of statements as strictly logical statements and

of words as fixed logical terms. Thus, he will analyse "Hitler is dead" as follows: "Hitler" is a logical proper name designating one person and one person only, and "being dead" is a precisely defined predicate term the meaning of which is fixed so that "Hitler is and is not dead" is ruled out logically. Here, language is being used in accordance with strict logic, but this is not its only use nor even its usual use. If the philosopher confines his study to this use of language, his account of it will be both narrow and misleading.

It is sometimes said that Ludwig Wittgenstein, a seminal philosopher of language between the wars, gave too narrow an account of the use of language in his *Tractatus Logico-Philosophicus* (1922), first published as *Logische-philosophische Abhandlung* (1921), and that this was one of the "grave mistakes" to which he confessed in the preface to his *Philosophical Investigations* (1953). It should be noted, however, that the *Tractatus* is primarily a study of logic and philosophy. It expresses his interest in what can and what cannot be stated in logically principled factual statements; and this is what can be thought logically, for what cannot be stated cannot be thought. It expresses an interest, too, in the logical principles involved in the making of these statements and holds that these principles are tautologous. Such was Wittgenstein's interest in the *Tractatus*.

But in *Philosophical Investigations* he looked at language in all of its uses and found it to be much more flexible and less restricted than he had held it to be in the *Tractatus*. Early in the *Investigations* he talked of language as a sort of game in which the rules vary according to custom. He who would understand how the game is played can do so only by participating in it. So, too, the meaning of the words used in a language is not immutably fixed; one can only know it by observing how the words are, in fact, used at any time. Obedience to linguistic rules is a matter of "custom, use or institution" rather than of following deep, unchangeable, logical principles. To understand a language means to be master of a technique. Thus, the attitude toward the study of language in the *Investigations* is very different from that in the *Tractatus*. It is Empiricist rather than Rationalist; but it is an Empiricism with a difference, for it is not scientific Empiricism. It does not seek to pass forward from the observation of particular cases to systematic generalizations about language; it is content to observe piecemeal and to bring out the puzzles that arise on reflection upon each new observation. Wittgenstein reflected, for instance, on uses of the utterance "I am in pain," and out of the reflection came the question whether and in what sense a language can be private.

The contributions of linguists

Neither the method of the *Tractatus* nor that of the *Investigations*, however, is the only method of language study. The study by linguists of the general structure of language itself gives rise to puzzles: for instance, whether, as Noam Chomsky, a generative grammarian, thinks, the facts of grammar are such that a distinction must be made between deep structures and surface structures of grammar and whether the phonemic and the semantic components of language are dominated by these deep syntactical components. Furthermore, the exact study of the natural languages by linguists and of speech by psychologists can provide results that are of value to the epistemologist, and, in fact, many scholars consider it regrettable that he does not make greater use of them. In particular, they could be useful in his study of thinking and his attempt to distinguish it from and relate it to knowing.

As has been seen, not all thinking or conception is verbal, though most of it is; and the question arises how language, as an instrument used in thinking, affects the thinking. Natural languages, such as English and French, have certain features in common (though not a common grammar)—for instance, they are conventional in their use of phonemic symbols; they employ analysis and synthesis; and they can pluralize and speak in general terms. The first use of language was for communication; in this use and in the development of language as a means of communication, thinking was already involved; and no hard-and-fast line can be drawn between thinking and using a language. When men became more and more reflective, the use of language as the instrument of thought affected that thought; the techniques of convention, analysis, and generalization were at hand for the thinker. Symbols could be used in new ways to further thought; deductive argument became possible; and thinking could be carried on in "eternal" sentences. But the epistemologist's question—*i.e.*, whether there is any guarantee that such conceptual thinking can ever be used to extend knowledge—remained to be answered. This is a question that the study of language cannot of itself solve, for it involves metaphysical as well as epistemological issues. In particular, it raises the issue of universals, which must now be briefly examined.

UNIVERSALS

Universals, which are expressed in common nouns or terms, may be described—perhaps inadequately—as either recurrences or principles of grouping or classifying. They have been a bone of contention among philosophers throughout the centuries. At various times, by various schools, the universal has been presented as the ultimately real existence, the Form, as a thought or a concept, as a word or series of words, as an identical recurring quality, as a group of qualities that resemble one another, as a class of things, as a genus, as a species, as a principle of classification, and, finally, as what can be predicated of an individual thing, a species, or a genus. In view of this complexity of definition, disagreement is not surprising. Much of the complexity is due to the fact that the theory of universals is advanced to solve a complex problem, viz., (1) to explain and justify grouping, classifying, and the use of the general word; (2) to find a standard by which to establish what is and is not valuable; and (3) to provide a metaphysic in which what is real can be clearly distinguished from what appears.

History of the problem of universals

Socrates and Plato between them set forward the theory of Forms or Ideas (see above *Rationalism prior to Kant*: *Classical Rationalism*: *Theory of Forms in Plato*) that claimed to solve the whole problem in its full complexity: human beings, closely tied to sense perception, think of visible and tangible things as real, whereas, in fact, they are appearances only; the beautiful things known in nature are manifestations of beauty itself; and, when a person knows beauty, he knows what is real, and he has a standard by which to determine what is and what is not beautiful; finally, knowledge of the Form of beauty enables him to group beautiful things together, provides him with a principle of classification, and enables him to use a general word significantly. Aristotle was critical of the view that the Form had this separate existence and found the universal in the common quality.

The argument was passed on to medieval philosophers through Porphyry, one of the founders of Neoplatonism, in the 3rd century and through the late Roman statesman and philosopher Anicius Boethius at the turn of the 5th century; it was hotly debated in the 11th and 12th centuries; and in the 13th century the different theories were neatly formulated by Albertus Magnus, the leading scholar of the century: universals are held by some (*e.g.*, by Plato) to exist before the things that exemplify them, to be *ante rem;* by others (*e.g.*, by Aristotle) to exist in the things, to be *in re;* by still others to exist in thought or conception only as abstractions or even as names only, to be *post rem*. In the 13th century it was the second theory that was most widely accepted, but there was more support for the third in the 14th century. *Ante rem* and *in re* theories were called Realist and *post rem* Nominalist, though in fact the medieval thinkers of this school were, in modern parlance, more often Conceptualists (the universal is a concept) than Nominalists (the universal is a word or set of words only).

In the modern period the Empiricists thought of universals as concepts abstracted from experience and tended to shift the discussion from the nature of the universal to the manner of its abstraction. The Rationalists maintained that some universals are known a priori. In the writings of contemporary philosophers, Realism of an

Aristotelian type and Conceptualism have flourished; but Positivists have advocated Nominalism, while Idealists and members of the mathematical logicist school have advocated a Platonic Realism.

Requisites for a solution

In epistemology the problem of universals is clearly linked with that of the ontological relevance of conceptual thinking. Its solution would seem to require the following: (1) a full understanding of the manner in which general words are used in thinking; (2) a study of qualities to help determine the precise nature of the link between the quality and the thing to which it is said to belong; (3) an examination of the theory of common qualities—asking whether the theory rests on the supposition of the existence of identical qualities or of resembling qualities or of both; (4) a determination of the principal dispositional features that are present, though not overtly expressed, in conception; and (5) a definitive answer to the question of what it is "to have a concept." It would seem that enough is known of these matters to enable one to say that there are universals. There are recurring resemblances between qualities, and there are also identities of qualities. In the second place, there are universals in the sense of principles of grouping or classifying. For the epistemologist the question then arises whether such a theory of universals can provide the necessary foundation for a theory establishing the ontological relevance of human thinking.

VI. Concluding reflections

The scope of the foregoing discussions has been limited by various considerations. In the first place, the article has been concerned solely with the thought of the West. (For the epistemologies of India and of China, see INDIAN PHILOSOPHY and CHINESE PHILOSOPHY.) In the second place, some matters that are in certain respects epistemological are not usually studied under this head. It is not usual, for instance, to discuss the nature of moral consciousness or, again, of aesthetic appreciation in works on epistemology, though such discussion might be illuminating. (See ETHICS and AESTHETICS.) Again, problems that are discussed in epistemological studies, such as knowledge of the self, of other selves, and of God, are not given special treatment here since they are covered in other articles (see MIND, PHILOSOPHY OF and RELIGION, PHILOSOPHY OF).

CONFLICT OF EMPIRICISM WITH RATIONALISM

The present inquiry into epistemology has concentrated on the main flow of Western speculation from the days of the Greeks to the present. As has been seen, two schools of thought are prominent, namely, the Empiricist and the Rationalist, but it is clear that neither of these schools has been able to offer a wholly satisfactory account of knowledge.

The Empiricist thesis is that human knowledge is either directly acquired in sense perception or derived from what is thus acquired. It follows from this that nothing is or can be thought that was not "previously in the senses." For such a theory a satisfactory account of the function of sense perception obviously is a crucial necessity.

Sense experience and the external world

Whether sense perception reveals the external world as it is and shows up its characteristics becomes an important question, though reflection on the contradictions in sensory experience has made it impossible to hold on for long to an affirmative answer to it. A person may observe that the moon he is looking at is covered wholly by a coin held at arm's length; yet, the moon for an observer standing on its surface is a vast object of which he sees a very small part. One concludes that what he now sees does not, in fact, show the moon's real size. Yet, he does see something. If he says that this visible something is not the real moon, he has then assumed a dualism. What he immediately sees is an object given to the senses, a sense-datum or sensum, but it is not the physical object, the real moon. The problem, then, is that of relating the two. It has to be asked how the Empiricist can say that sense perception reveals the real external world if what is immediately seen is a sensum. The dualism of real object and sensum would seem to throw doubt on the very possibility of knowledge of the external world in sense perception. It is, perhaps, possible to argue that the sensum is a representation of the real object or that it represents it in certain selected aspects, so that the real object is in part known. But this is to raise the question of whether and how the claim that the sensum represents, say, such and such a quality or qualities of the real object can be tested. If the difficulties of the representationalist theory have driven philosophers to reject dualism, some might suggest that, in order to re-establish a monist theory of knowledge, one should adopt then an Idealism denying the independent reality of the physical object or affirm that human knowledge is phenomenalist. In the latter view, man is concerned with phenomena only and the question of whether physical objects exist or do not exist beyond the veil of phenomena does not arise. Yet, both the Idealist and phenomenalist accounts, too, are full of difficulties, and it is doubtful whether the original thesis, that all human knowledge of the outer world is directly acquired in sense perception or derived from what is so acquired, can be sustained by the adoption of either account.

Hearsay and abstracted scientific knowledge

There have been philosophers who have held that sense perception is a snare and that it provides no reliable information; yet, this seems extreme. What is objectionable in the Empiricist position is rather the claim that sense perception is the sole provider of information. On the face of it, it seems incorrect to say that all that a person knows of the world around him is known through sense perception. For instance, it may be true that much of what one knows he learns from what others tell him (in speech or writing)—*i.e.*, from "hearsay knowledge." And the surest source of information may well be the knowledge provided by an expert in any field. To this the Empiricist might reply that in both cases the information in question, though not given to the layman in sense experience or derived from what is so given, is nonetheless given in or derived from the sense perception of his informant or of someone else. Thus, a man is told that there are black swans in Australia. He has never seen a black swan and has never been to Australia; yet, if he is satisfied about the reliability of his informant and especially if the latter says, "I saw them there," he accepts this information. The layman's gaining of the knowledge, however, is dependent not only on the informant's sense experience but also on the layman's understanding of what he says; thus, one condition of hearsay knowledge is the successful use of a language. But this, the Empiricist would say, has to do with the medium in which the information is conveyed to another person; the information itself is gained in sense experience.

A more complex case is presented when an inductively established scientific generalization is reported. Though it will be said that this generalization is still empirically based, "based" is an ambiguous word in this context. The information has been gained by the scientist using the methods that he and his fellow workers accept in this field. It involves sense experience or sensory observation together with scientific reasoning. The knowledge is admittedly empirically based, but this leaves open the question whether it is correct to speak in this case of sense experience as the sole provider. The point can be taken a step further. One of the most important features of scientific method is hypothesis making. The scientist, having observed widely and being thoroughly acquainted with the body of knowledge about the matter in hand, has an insight and proffers it as a hypothesis. Such an insight could possibly be described as a perception but certainly not as sense perception. A natural science is an empirical inquiry; it uses information gained in sensory experience; further, it verifies hypotheses by appeal to the senses. Yet, the Empiricist's claim that such knowledge is either gained in sense perception or derived from sense perception must be dismissed as misleading.

Rationalist infallibility

It is generally agreed that the empirical sciences do not give certainty but only probability. To assert this is to assert that there are degrees of knowledge; a person may know a statement P with complete certainty (know "in the strict sense"), or he may opine or believe P without

being certain, holding it to be just probable or probable with a high or low degree of probability. In the traditional Rationalist view, as has been seen, it is man's reason alone that enables him to know at the highest level—to say "This is true absolutely." Absolute truth is not given in sensory experience nor can it be derived from it. But man is endowed with a pure intellectual power that enables him to know truth positively—truth that is not qualified in any way, which is unchanging and necessary. The information presented by sense experience is uncertain, holding relatively with a certain context, so that to build upon it is to build on shifting sand. In contrast, the information that reason provides is unchanging: "once true always true." To know by reason is to know infallibly; and it is on this rock of infallible rational intuition that human knowledge rests.

This doctrine contrasts strongly with Empiricism. It is undoubtedly comforting, but reflection on its claim still generates doubts. The main doubt is the all-embracing one of whether human beings ever have a knowledge that is strictly infallible. Plato thought they do. But only a few, the "true philosophers," have it, and even in their case infallible knowledge is of Forms or Ideas only. Others supposed it to be more widely spread—for instance, every rational creature could see certain mathematical and logical statements to be true beyond all doubt. Thus, the statement, "The interior angles of a triangle are equal to two right angles," is infallibly true and so is the statement, "This figure is a triangle or not a triangle." These are necessarily true statements that cannot possibly be false, because they are tautologous. No one can doubt that *a* is *a*. But, if the Rationalist thesis is taken to mean that reason provides and is the sole provider of infallible knowledge and if this infallible knowledge is of the tautologous, it might appear that reason's contribution is slight. Further, it may be asked whether even the barely tautologous statement is known infallibly.

BLEND OF EMPIRICISM WITH RATIONALISM

The blended theory and its defects

It could be argued that, in order to make the contrast between Empiricist and Rationalist clear, the immediately preceding paragraphs have tended to exaggerate the differences between them. For, when the history of the conflict is recalled, one finds few Empiricists, if any, who do not have some place for reason in their theory of knowledge and few Rationalists who are not prepared to acknowledge the contribution of sense experience. The difference is one of emphasis: the part played by experience is more important in the Empiricist's mind, and that by reason is more important in the Rationalist's mind. But the general view is that human knowledge is dependent on both experience and reason; experience provides the "manifold," or material, of knowledge; reason provides the principles for ordering this material. Such is the general line followed by epistemologists.

If human knowledge is regarded as a blend of experience and reason, it is necessary to ask whether the theory is then acceptable. It is reasonable to ask whether the blending will result in the modification of material given in experience and reason. Sensory experience, it is supposed, is an awareness of the concrete particular; and there could be a danger that the ordering zeal in the blended knowledge will diminish the concreteness of the immediately experienced and so mislead. On the other hand, reason in its purity provides necessary, infallible knowledge. The question, then, is whether the material that it is now called upon to order will be such that no necessary knowledge is henceforward possible.

Problems of this sort arise. They may be dismissed by some philosophers as the consequences of making two abstractions—sense experience and reason—into one thing and of viewing sense experience in this context as something working against generalization and reason as struggling to secure infallibility. It must be admitted that the tendency is strong to think and speak of sense experience and reason as two faculties operating more or less independently in the attempt to increase knowledge.

As against the above criticisms of the blended theory of knowledge, however, two points can be made. In the first place, the assumption that sense experience gives the particular and never the general can itself be questioned. The evidence available suggests that very young children do not distinguish between particular and general. What one begins with, it would seem, is as general as it is particular. It is only later, when the child has reflected, that he comes to conceive a particular thing, isolated, substantial, and enduring. It may be wrong, therefore, to suppose that early experience is of the concrete and particular and that later conceiving generalizes the experience and to a degree falsifies it. In the second place, it can also be questioned whether the function of reason is to provide man with infallible principles by which to order the manifold of experience. It may be that these principles are provided in some other way: they may be stipulations of the thinker, or they may be hypotheses. Mathematics and logic, it is said, do provide complete certainty within a framework of tautologies. But one may then ask whether the function of reason is to grasp the truth of tautologies. It is necessary to question the assumptions of the critics of the blended theory both in respect to experience and to reason.

Rejoinders to the criticisms

There is, further, a double assumption in arguing that man is capable of infallible knowledge and therefore that reason must be one of his faculties, since reason alone knows the infallible. Even if man enjoyed infallible knowledge, it would still be necessary to prove that the possession of the faculty of reason is a requisite for such knowledge; one cannot assume this. Nor can a person assume that man does have infallible knowledge. It might be answered: "But surely he is certain of many things!" This is so, but to be certain of something is not the same as having infallible knowledge of it: it is not the same for the simple reason that a man can be certain, absolutely certain, and yet be in error.

Remaining problems

It is clear that many problems are left for the epistemologist to solve. No one would say that the subject of sense perception is satisfactorily disposed of, in spite of the considerable attention paid to it. Interest in the analysis of conception is comparatively new, but it has become apparent that difficulties abound in this field. Along with the study of concepts goes that of universals, and linked with both is the study of thinking, which in turn involves an examination of language and of its relation to thought. Finally, there looms the problem of reason and of its function in human knowledge and in human life. The study of these and other matters will certainly keep epistemologists busy for the foreseeable future.

BIBLIOGRAPHY. The subject of epistemology is wide and a vast number of books have been written in this field. What follows is a brief selection, mostly of mid-20th-century books, which the student might find useful as an introduction.

General introduction: A.J. AYER, *The Problem of Knowledge* (1956); R.M. CHISHOLM, *Theory of Knowledge* (1966); and BERTRAND RUSSELL, *Problems of Philosophy* (1912). Also A.D. WOOZLEY, *Theory of Knowledge* (1949) and Panapot Butchvarov, *The Concept of Knowledge* (1970).

Sense-perception: Highly recommended is H.H. PRICE, *Perception*, 2nd ed. rev. (1950). The following are further useful studies: A.J. AYER, *The Foundations of Empirical Knowledge* (1940); R.M. CHISHOLM, *Perceiving* (1957); R.J. HIRST, *Problems of Perception* (1959); J.L. AUSTIN, *Sense and Sensibilia*, reconstructed from the manuscript notes by G.J. WARNOCK (1962); D.W. HAMLYN, *Sensation and Perception* (1961); A.R. WHITE, *G.E. Moore: A Critical Exposition* (1958); A.O. LOVEJOY, The *Revolt Against Dualism* (1930); H.A. PRICHARD, *Knowledge and Perception* (1950); M.H. MANDELBAUM, *Philosophy, Science, and Sense Perception* (1964); and WILFRED SELLARS, *Science, Perception, and Reality* (1963), reprinted 1968.

Belief and certainty: One may begin by referring the reader to C.I. LEWIS, *An Analysis of Knowledge and Valuation* (1947); BERTRAND RUSSELL, *Human Knowledge: Its Scope and Limits* (1948); and G.E. MOORE, *Some Main Problems of Philosophy* (1953). For later views, see NORMAN MALCOLM, *Knowledge and Certainty* (1963); JAAKKO HINTIKKA, *Knowledge and Belief* (1962); BRAND BLANSHARD, *Reason and Analysis* (1962) and *Reason and Goodness* (1961); RICHARD I. AARON, *Knowing and the Function of Reason* (1971); and A.C. DANTO, *Analytical Philosophy of Knowledge* (1968). Very relevant, too, are LUDWIG WITTGENSTEIN's last investigations

On Certainty (1969). The collection of papers entitled *Knowledge and Belief*, ed. by A. PHILLIPS GRIFFITHS (1967), is to be highly recommended.

Thought and language: On the logical (or philosophical logic) side, the reader will find much help in W. and M. KNEALE, *The Development of Logic* (1962). Valuable, too, is W.V.O. QUINE, *Word and Object* (1960), and also his *From a Logical Point of View*, 2nd ed. rev. (1961) and *The Ways of Paradox, and Other Essays* (1966). The most influential modern work in this field is LUDWIG WITTGENSTEIN, *Philosophical Investigations*, 2nd ed. (1958, reprinted 1967), with which one may read F. WAISMANN, *The Principles of Linguistic Philosophy* (1965). See also: G. RYLE, *The Concept of Mind* (1949) and *Dilemmas* (1954); H.H. PRICE, *Thinking and Experience* (1953); BRAND BLANSHARD, *The Nature of Thought*, 2 vol. (1939), a work reflecting the influence of Hegelian Idealism; J. HOLLOWAY, *Language and Intelligence* (1951); RICHARD I. AARON, *The Theory of Universals*, 2nd ed. (1967); P.T. GEACH, *Reference and Generality* (1962); and P.F. STRAWSON, *Individuals* (1959).

Schools of epistemology: Elsewhere in this encyclopaedia are biographies of the leading philosophers together with accounts of their principal works. See also the bibliographies of articles on the various epistemological schools and movements mentioned in the text. The reader may also consult further the *Encyclopedia of Philosophy*, ed. by PAUL EDWARDS, 8 vol. (1967); and D.J. O'CONNOR (ed.), *A Critical History of Western Philosophy* (1964). Below are some useful commentaries.

The Greeks: On early Greek philosophy the best general guide for the newcomer is still JOHN BURNET, *Greek Philosophy: Thales to Plato* (1914). On Plato, see the comprehensive *Plato: The Man and His Work*, by A.E. TAYLOR (1926). The standard commentary for Aristotle is W.D. ROSS, *Aristotle* (1923). A number of excellent commentaries are available on individual works by Plato and Aristotle: F.M. CORNFORD, *Plato's Theory of Knowledge* (1935), is noteworthy. See also GREGORY VLASTOS, "The Third Man Argument in the *Parmenides*," *Phil. Rev.*, 63:319–349 (1954). G.E.L. OWEN and H. CHERNISS are leading contemporary scholars in this field; see the latter's *Aristotle's Criticism of Plato and the Academy* (1944) and *Aristotle's Criticism of Presocratic Philosophy* (1935, reprinted 1964). An introductory account of Aristotle's philosophy is D.J. ALLAN, *Philosophy of Aristotle* (1952). The main difficulty with minor Greek philosophic writers is to find out exactly what they say, since it is necessary to depend on fragments. G.S. KIRK and J.F. RAVEN, *Presocratic Philosophers* (1957); H.F. VON ARNIM, *Stoicorum Veterum Fragmenta* (1905–24); and H. USENER, *Epicurea* (1887, reprinted 1966), provide good collections.

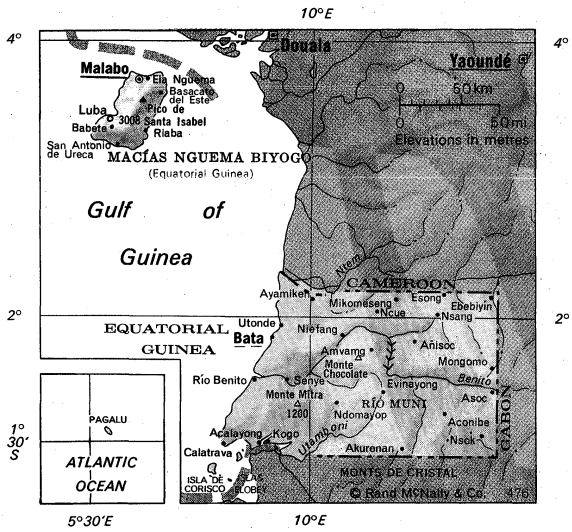
Medieval: Two useful general commentaries are ETIENNE GILSON, *A History of Christian Philosophy in the Middle Ages* (1955); and F.C. COPLESTON, *A History of Philosophy*, vol. 2–3 (1950–53). A good introduction to Aquinas is F.C. COPLESTON, *Aquinas* (1955); a fuller work is ETIENNE GILSON, *Le Thomisme*, rev. ed. (1922; Eng. trans., *The Philosophy of St. Thomas Aquinas*, 2nd rev. ed., 1929).

Modern: Not unexpectedly, the best commentaries on Descartes are in French: see HENRI GOUHIER, *La Pensée métaphysique de Descartes* (1962); and OCTAVE HAMELIN, *Le Système de Descartes* (1921); but there are good introductory studies in English, for instance, S.V. KEELING, *Descartes*, 2nd ed. (1968). On Spinoza, two contrasting works for those beginning to read this philosopher, who is not too easy to understand, are L. ROTH, *Spinoza* (1924); and STUART HAMPSHIRE, *Spinoza* (1951). RUTH SAW has written on *Leibniz* (1954); but BERTRAND RUSSELL, *A Critical Exposition of the Philosophy of Leibniz*, 2nd ed. (1937), may not prove too difficult. General studies of John Locke are D.J. O'CONNOR, *John Locke* (1952); and RICHARD I. AARON, *John Locke*, 3rd ed. (1971); of Berkeley, G.J. WARNOCK, *Berkeley* (1953); of Hume, D.G.C. MACNABB, *David Hume* (1951). The beginner will find Kant difficult and introducing him is not an easy task, but compare S. KORNER's general introduction on *Kant* (1955). On Idealism generally, an excellent study is A.C. EWING, *Idealism* (1934). (R.Aa.)

Equatorial Guinea

The Republic of Equatorial Guinea (República de Guinea Ecuatorial), on the west coast of Africa, is the only independent Spanish-speaking state on the African continent. Formerly Spanish Guinea, a colony of Spain, it consists of two geographically separated provinces—Río Muni, a coastal enclave, and Macías Nguema Biyogo (formerly Fernando Po), an offshore island—whose unity is only political and cultural. This fragmented republic has a total area of 10,831 square miles (28,051 square kilometres) and a population that in 1975 was estimated to be about 310,000.

Río Muni, or continental Guinea, with an area of 10,038 square miles (25,998 square kilometres), is a roughly rectangular territory bounded by Cameroon to the north



EQUATORIAL GUINEA

and Gabon to the east and south. The estuary islands of Corisco (about six square miles), Elobey Grande (less than a mile square), and Elobey Chico (less than a tenth of a mile square) are administratively included in the province of Río Muni, which thus has a total area of 10,045 square miles. The second province, Macías Nguema Biyogo consists of the island of Macías Nguema Biyogo proper—off the coast of Cameroon, with an area of 779 square miles (2,018 square kilometres)—and Pagalu (formerly Annobón), a volcanic island of less than seven square miles that lies south of the Equator almost 400 miles (640 kilometres) to the southwest of Macías Nguema Biyogo.

By any standard Equatorial Guinea forms a small political unit, and it is beset by regional differences, geographic isolation, a fragile economy, a lack of trained personnel, and the incubus of a heavy psychological legacy from the colonial era, which ended when the country obtained its independence on October 12, 1968. The capital of the republic is Malabo (formerly Santa Isabel), on Macías Nguema Biyogo. Bata, on the mainland, is only the administrative capital of Río Muni but aspires to become the national capital. (For coverage of historical aspects, see WEST AFRICA, HISTORY OF.)

Independence

Relief, drainage, and soils. *Río Muni.* The enclave of Río Muni is too small to have a geographic unity of its own. A coastal plain about 12 miles (20 kilometres) wide abuts on the coastal hills, which lead to inland plateaus (called *mesetas* in Spanish) that rise toward the frontier with Gabon. There are several ranges of hills. The central range divides the Río Benito Basin to the north from the southern basin of the Río Utamboni. The Niefang-Mikomeseng range north of the Río Benito is somewhat lower. All these ranges form segments of the Monts de Cristal in Gabon. The province is divided by the Río Benito (known as the Woleu River in Gabon), which runs generally from east to west and is nonnavigable except for the first 12 miles inland. To the north, the Río Campo (called the Ntem in French-speaking Africa) marks part of the frontier with Cameroon. In the south, Río Muni, from which the province takes its name, is not a river but the estuary of various Gabonese and southern Río Muni rivers, among which the Utamboni is the largest. To the east, the border with Gabon follows the meandering course of the Río Kié (Kyé) rather than the legal frontier, which runs along a line 11°20′ east of the Greenwich meridian. Except for some limited use of waterpower on lumbering sites, the rivers of Río Muni are not exploited. The coast

consists of a long stretch of beach with low cliffs toward Kogo to the south. The province has no natural harbour, and Bata, Río Benito, and Kogo are no more than rudimentary ports of call for the ships that infrequently visit the coast.

The coastal plain is overlaid by sedimentary deposits. The hinterland is composed primarily of ancient metamorphic rocks that have undergone a lengthy process of leaching and erosion, so that the resulting soils are mediocre. Exploration for petroleum in offshore waters was discontinued in 1971. Gold, titanium, coal, and iron ores are said to exist inland, but no economic exploitation has begun.

Macías Nguema Biyogo. The main island, Macías Nguema Biyogo itself, is roughly 45 miles long and 22 miles wide (72 by 35 kilometres). Its volcanic cones, crater lakes, and rich lava soils form a contrast with the landscape of Río Muni. In the north, the Pico de Santa Isabel soars to a height of 9,868 feet (3,008 metres); this extinct volcano is now the site of a television transmitter. In the centre of the island, Moca Peak and the Moca Heights present an alpine type of landscape. The southern part of the island, remote and scarcely developed, consists of the Gran Caldera range, which is rugged and indented by torrents. There are two crater lakes, Loreto and Claret.

Agricultural productivity

Despite its tortuous relief, Macías Nguema Biyogo is highly productive agriculturally. Torrents are exploited for hydroelectric power; the Río Musola provides electricity for the township of Malabo. The coast is largely inhospitable, consisting for the most part of a cliff about 60 feet high, broken occasionally by small inlets and beaches. The southern coast is very steep and dangerous to shipping; San Antonio de Ureca, located there, is the most isolated settlement on the island. Malabo has a relatively good harbour, being built on the partially sunken rim of a volcano.

Pagalu is an isolated fragment of the republic, about 93 miles (150 kilometres) southwest of the island of São Tomé in São Tomé e Príncipe and about 400 miles southwest of Macías Nguema Biyogo. Like the latter, it is a volcanic island but is less high; it consists of a conglomeration of cones of which Monte de Santa Mina (about 2,460 feet) is the highest. Not quite four miles long by two miles wide, it is a rugged little island; the only settlement of note is centred on the Catholic mission. The inhabitants are mostly fishermen who speak a Portuguese patois.

Climate. The climate of Río Muni is equatorial, with local variations due to differences in altitude and proximity to the sea. The wet seasons are from February to June and from September to December. Rainfall is higher on the coast than inland. In Bata the rainiest months are September, October, and November, with rainfall averaging more than 94 inches (2,388 millimetres) a year. At Calatrava, farther south on the coast, it sometimes reaches 180 inches (4,572 millimetres). Inland, however, rainfall diminishes; Mikomeseng, for example, receives only about 58 inches (1,473 millimetres). The average yearly temperature is about 79° F (26° C), the maxima being somewhat lower than in Macías Nguema Biyogo. The relative humidity, however, is higher than in Macías Nguema Biyogo, with an average of 86 percent, occasionally rising to 94 percent.

Macías Nguema Biyogo has an extremely debilitating climate. The so-called dry season lasts from November to March, and the rest of the year is rainy. The average annual temperature is 77° F (25° C) with a maximum of 93° F (34° C) and a minimum of 63° F (17° C). Most of the time the sky is cloudy and overcast. Extreme rainfall occurs in the south, with rain brought by monsoon winds amounting to about 450 inches (11,400 millimetres) a year around San Antonio de Ureca.

Vegetation and animal life. Río Muni is covered by a dense jungle that is exploited by a lumbering industry. Over 140 species of wood are found, the most important commercially of which are okume (*Aucoumea klaineana*), African walnut, and various mahoganies. Intensive exploitation close to the coast led the Spanish timber companies to venture deeper into the interior. Reforestation is minimal, and a secondary forest growth has replaced the original rain forest. African farmers clear land by burning off the vegetation cover, after which they grow manioc, beans, yams, roots, and even cash crops such as coffee and cocoa, to the extent that the heavily leached soils allow. Mangroves fringe long stretches of the coast as well as riverbanks.

Macías Nguema Biyogo has a greater variety of tropical vegetation, including mangroves, and—at higher altitudes—vegetable gardens and pastureland.

Varied animal life

Río Muni has a rich animal life that includes gorillas, various rare species of monkeys, leopards, buffalo, antelopes, elephants, hippopotamuses, crocodiles, and various species of snakes, including pythons. Insects, including the tsetse fly and the malaria-bearing *Anopheles* mosquito, as well as hosts of ants, beetles, spiders, and termites, abound. Macías Nguema Biyogo has no big game but has various monkeys, dwarf antelopes, and rodents, as well as the ever-present mosquito and other insects.

The landscape under human settlement. *Río Muni.* Río Muni is sparsely settled by coffee and cocoa farmers who practice traditional methods of agriculture. During the colonial period, Roman Catholic missions did much to encourage the population to construct "corridor" villages by the sides of roads; in most villages the church and the school figure prominently. Spanish plans for building model villages lapsed after independence. Río Muni was never a settler colony, and the few European plantations—mostly Spanish or German—that survived the colonial era have been abandoned. After independence an exodus of Spanish technicians occurred, so that Río Muni has largely reverted to a subsistence economy. While subsistence agriculture sustains the population, it remains to be seen whether the modern sector of the economy, represented by lumbering, coffee growing, and oil palm production, can recover from the crisis of confidence that occurred after independence.

Macías Nguema Biyogo, by contrast, is a plantation island; it retained for several years a larger number of plantation owners and managers and consequently withstood longer than Río Muni the effect of independence upon its economy. Before independence there were about 1,900 plantations (known as *fincas*), which ranged in size from one hectare to more than 2,000 hectares (4,900 acres). The Bubi people, the indigenous population, live for the most part in mission villages in the northern part of the island on the lower slopes of the Pico de Santa Isabel, as well as in their traditional homeland, the Moca Heights. Unused to plantation labour, they gain their living mostly as small farmers and minor civil servants. The workers on the *fincas* were for many years migrant labourers from Nigeria, who served under contract. During the 1960s the Nigerian workers often brought their families, settling in numbers believed to have reached 50,000 to 80,000 by the end of the decade. Political and economic conditions after independence gradually reduced these numbers, despite an agreement with Nigeria in 1972 for the recruitment of new labourers. Reports of virtual slave-labour conditions on plantations and of repressive killings by authorities in 1975 turned this gradual exodus into a flood, with the anticipated effect of further impoverishing Equatorial Guinea's economy.

Nigerian migrant labour

Many of the formerly European-owned plantations were extensive; in 1962 more than 300 plantations occupied over 60,000 hectares (148,000 acres), leaving about 7,000 hectares in the possession of 1,600 African farmers grouped in cooperatives. Some of the larger plantations, employing hundreds of Nigerian labourers and occupying the most fertile land, were local economic powers whose political influence, if subdued, was nevertheless felt by the local Bubi population.

Malabo, both the provincial and the national capital, is a small city standing behind its crater harbour. Created by the British in the 19th century, it was remodelled and developed by the Spanish. A rambling tropical city, it has a distinctly Spanish atmosphere—especially in the European district near the cathedral, the mission, and government house. Farther inland, the African districts have been inhabited mostly by Nigerian and other workers who chose not to return to the mainland. While the

Malabo township and its Ela Nguema (formerly San Fernando) satellite had a total population of about 20,000 in 1960, it is safe to assume that the population has risen to between 25,000 and 30,000. Another town of some importance is Luba, on the southwest coast, linked with the capital by a good tarred road that runs through a series of Bubi settlements. Basilé, on the slopes of Pico de Santa Isabel, provides a cool refuge for heat-weary residents of the capital.

Río Muni was settled much later by the Spanish, so that the township of Bata lacks the amenities of Malabo. Its population was less than 4,000 in 1960 but now numbers perhaps 15,000, as Fang migrants from the interior have built new suburbs around the sprawling port city.

People and population. Population figures are outdated, since the only reliable estimate dates back to 1960, when there were almost 250,000 people, of whom 7,000 were Europeans. It may be assumed that the 3,000 Europeans living in Río Muni have dwindled to a few dozens at most, and that the 4,000 Europeans on Macías Nguema Biyogo have declined to a few hundreds. In 1960 Río Muni had about 180,000 inhabitants and Macías Nguema Biyogo about 60,000 (including 1,400 on Pagalu). The majority of the population is African, but its composition is complex for a political unit so small in size.

Río Muni. The Fang people, who fought their way to the sea in the 19th and early 20th centuries by subjugating the weaker tribes in their path, form about 80 or 90 percent of the population of Río Muni. North of the Río Benito are the Ntumu Fang, and to the south of it are the Okak Fang. Holding political power on the mainland, the Fang tend to migrate to Macías Nguema Biyogo, where their leaders hold most of the levers of political control. Coastal tribes—such as the Kombe, Mabea, Lengi, Benga, and others—have been in contact with European traders much longer, and a limited amount of miscegenation has taken place, especially on the island of Corisco. Spanish ethnographers refer to these coastal tribes as *playeros* (literally, "those who live on the beach"). Both the Fang majority and the *playero* minority are Bantu. Some Pygmies are reported to roam in the vicinity of the Río Campo. In addition to these peoples, there are substantial colonies of Africans from other countries in Río Muni, including Hausa traders in Bata and Ibo, Ibibio, and Efik contract workers from Nigeria who are employed in the lumber camps. There is also a small number of Africans from Cameroon and Gabon, as well as a few Indian shopkeepers in Bata.

The Bubi

Macías Nguema Biyogo. The original inhabitants are the Bubi, descendants of Bantu migrants from the mainland. Contacts with Europeans decimated them, and only a few thousand remained early in the 20th century. They now number about 15,000 and constitute the most pro-Spanish element of the African population, viewing the end of Spanish rule as a signal for the invasion of their island by the Fang. Certainly, numbers of Río Munians, most of them Fang, have flocked to the island since the mid-1960s, seeking to join the civil or military forces or to receive political patronage. In addition to these two groups, there are a few thousand Fernandinos, descendants of former slaves liberated by the British during the 19th century who mingled with other emancipated Africans from Sierra Leone and Cuba as well as with immigrants from West African countries. Formerly constituting an influential bourgeoisie, they lost much of their status when the Spanish acquired the island. The inhabitants of Pagalu are descended from slaves imported by the Portuguese when the island was a dependency of São Tomé; some of them now live on Macías Nguema Biyogo.

By about 1970, all these different strata together constituted a minority on the island, the majority being formed by the mass of Nigerian contract labourers, who lived in compact colonies either in Malabo or on the plantations. The repatriation by Nigeria of an estimated 45,000 workers beginning in mid-1975, following reports of repressive conditions in Equatorial Guinea, could be expected to lead to extensive realignment of the demographic, social, and labour structures of the island and, indeed, of the country. Additional communities on the island are formed by *crioulos* (of mixed Portuguese and African origin) from the formerly Portuguese islands of São Tomé and Príncipe; there are also Cameroonians.

The island's population thus forms a true ethnic kaleidoscope, ruled by a Spanish-speaking group who control the Guardia Nacional, composed of Fang, who are garrisoned on the island.

Language. While each ethnic group speaks its own language, two linguistic influences are at work. The first is Spanish, which is the official language of the republic, is taught in schools and used by the press, and is the only means of communication common to both Río Muni and Macías Nguema Biyogo. The second influence is pidgin English, which forms the lingua franca on Macías Nguema Biyogo and enables the Nigerians to make themselves understood. A Portuguese patois is also spoken both in Macías Nguema Biyogo and Pagalu.

Religion. While about 80 percent of Equatorial Guineans are nominally Roman Catholic, the Bubi and Río Munians have all more or less openly retained traditional forms of worship, which are emphasized in times of crisis. The Mbwiti cult in Río Muni, banned by the Spanish authorities, still has adherents. There are American Presbyterian and English Methodist missions on the mainland. The only Muslims are the Hausa.

The national economy. Equatorial Guinea's economy has depended primarily upon three commodities—cocoa, coffee, and timber. Before independence, the Spanish subsidized cocoa and coffee exports to Spain. Cocoa, which is of high quality, is the mainstay of the economy of Macías Nguema Biyogo, which possesses the right soil and climate for its intensive cultivation. About 90 percent of Equatorial Guinea's cocoa is produced on the island, only about 10,000 hectares (25,000 acres) of land being devoted to cocoa cultivation in Río Muni. Equatorial Guinea's coffee is of poor quality; most of it comes from Río Muni, with Macías Nguema Biyogo's share of production being marginal. The timber companies are entirely European; there were at one time 30 timber concessions in operation in Río Muni. The industry is heavily mechanized and requires a large investment. Friction between the Fang-dominated government and the timber companies, however, led in 1969 to the disruption of the economy of Río Muni.

Cocoa, coffee, and timber exports

Bananas are grown on Macías Nguema Biyogo, where they are exported from the port of Luba. The output of palm oil from Río Muni plantations has been adversely affected by the unsettled state of the country. A Spanish fishing fleet was based on Macías Nguema Biyogo in the late 1960s. Cattle raising in Moka Heights is of minor importance. There is little industry in either of the country's two provinces. Some cocoa and coffee processing takes place locally on plantations and in African cooperatives. There is no mineral production. External trade usually shows a favourable balance, but this is largely because internal consumption of goods and commodities is so low. Since independence the national budget has been balanced chiefly by large Spanish subsidies.

Transport. The road network in Río Muni was adequate for the light traffic it was required to carry before independence, but since local autonomy was granted in 1964, neglect has led to deterioration of the earth roads. Bata is linked with Río Benito by a tarred road. The trans-Río Muni road from Bata to Ebebiyín, near the Cameroon frontier, is partially tarred. There are no railways. On Macías Nguema Biyogo the road system is of a higher standard, with a semicircular tarred road linking Batete to the eastern Bubi villages. The main harbour is Malabo, which is connected to Spanish and other European ports by regular steamer services. There is a regular steamer service between Bata and Malabo, but the service to Pagalu is erratic. Communications are maintained with Calabar in Nigeria and with Douala in Cameroon to transport migrant workers to and fro. There is an international airport at Malabo with links to Madrid, Douala, and Bata as well as weekly flights to Lagos (Nigeria) and Moscow and Odessa in the Soviet Union.

Administration. The system of government is presidential; and while all governing bodies are elected by

universal adult suffrage, the rule of the first president, Francisco Macías Nguema, is highly personal. The Council of Ministers is appointed by the President and is responsible to him. At least one-third of the 10 ministers on the council must be from Macías Nguema Biyogo. The National Assembly of 35 members is the legislative body. Río Muni has 19 deputies, Macías Nguema Biyogo 12, Pagalu two, and the Río Munian islets two. A Council of the Republic, consisting of six members elected by the respective provincial councils, has the function of mediating any dispute between the Council of Ministers and the assembly and between the central government and the two provincial councils, which are also elected by universal suffrage. This complex structure has been devised to protect the rights of the Macías Nguema Biyogo minority. The supreme tribunal in Malabo is the highest judicial authority.

In March 1970, all the political parties that existed before independence were merged into a Partido Único Nacional (Single National Party), which later changed its name to Partido Único Nacional de Trabajadores (Single National Party of Workers), under the leadership of the President, who assumed most of the powers in the country and subsequently was made president for life. Since he is a Fang, and most of the other political figures are also Fang—in addition to the members of the Guardia Nacional—the Fang now constitute the dominant element in Equatorial Guinea.

Cultural life. Despite a veneer of Spanish culture and of Catholic religion that is thicker in Macías Nguema Biyogo than in Río Muni, Equatorial Guineans live largely according to ancient customs, which have undergone a revival since independence. Among the Fang of Río Muni, witchcraft, traditional music (in which the Fang harp, the xylophone, the great drums, and the wooden trumpet are used), gorilla and elephant hunting, and storytelling all do much to relieve the monotony of life in the forest. Among the Bubi farmers of Macías Nguema Biyogo, some ancient customs are still followed.

Role of ancient customs

Education at primary levels was encouraged by the Spanish. In 1970–71 there were more than 6,000 students in secondary schools and technical institutions. At the time of independence, fewer than 30 Equatorial Guineans had completed higher education in Spain.

There is one daily newspaper in Malabo with a circulation of 1,000, and one weekly in Bata with a circulation of 500; both are in Spanish. There is also an official gazette, a monthly, and a Catholic bulletin. There are two radio stations, one in Bata and the other in Malabo. Both are operated by the government, as is the television station that operates from Malabo with programs in Spanish and in African vernacular languages. Cultural life centres on the Catholic mission in Malabo.

Prospects for the future. Equatorial Guinea received independence after a four-year transitional period that could have been used to prepare the small country for its future role but was not. The stability of its fragile economy rests upon the confidence of foreign investors in the Fang leadership; meanwhile, whatever goodwill existed between the local Spanish settler community and the Fang leaders was largely dissipated in the conflicts that occurred in 1969 and after the disturbances of 1970 in Malabo. The situation is complicated by the fact that, while the economic equation includes the necessary presence of tens of thousands of Nigerian migrant workers on Macías Nguema Biyogo, these were being repatriated in 1975–76 because of brutal treatment.

For a variety of reasons—including the reluctance of some Spanish settlers to cooperate with the new leadership, tribalism, and the influence of vested interests—the newly independent country has made a false start. It now struggles with a shortage of trained personnel and capital and is also not yet psychologically prepared for its role as an independent state. Because of its isolation from other West African states, which have been subjected to French and British influences, it is obliged to rely upon Spanish, Soviet, and Chinese help in overcoming the formidable obstacles presented by geographical fragmentation, ethnic diversity, and economic uncertainty. In addition, the strategic position of Macías Nguema Biyogo cannot be discounted in evaluating Equatorial Guinea's future prospects. It has on several occasions (once through its use as a base for a humanitarian airlift during the Biafran conflict) played a role in West African history and may do so again.

BIBLIOGRAPHY

Geography: MANUEL DE TERAN, *Síntesis geográfica de Fernando Poo* (1962), a useful geographical analysis; LUIS BAGUENA CORELLA, *Guinea: Manuales del África Española* (1950), the only detailed compendium of Guinean problems under the colonial regime; RENE PELISSIER, *Les Territoires espagnols d'Afrique* (1963; trans. and updated in 1964, *Los territorios españoles de África*), a handy reference on Spanish Africa; "Spanish Guinea: An Introduction," *Race*, 6:117–128 (1964), the only general review in English; SERVICIO INFORMATIVO ESPANOL, *España y Guinea Ecuatorial* (1968), an official record of developments until independence.

Economy: PRESIDENCIA DEL GOBIERNO, *Plan de desarrollo económico de la Guinea Ecuatorial: años de 1964 a 1967* (1963), the only reliable study of the subject.

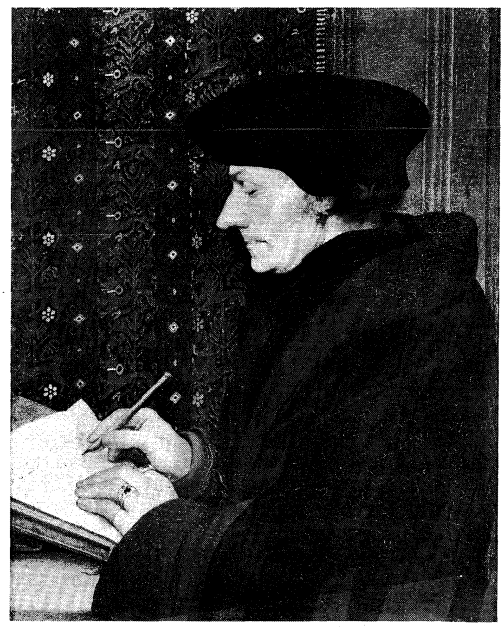
History and politics: ABELARDO UNZUETA, *Geografía histórica de la Isla de Fernando Poo* (1947), *Guinea Continental Española* (1944), *Islas del Golfo de Guinea* (1945), three standard, although poorly organized, works on history; RENE PELISSIER, *Études Hispano-Guinéennes* (1969), the only detailed analysis of the pre-independence period.

(R.P.)

Erasmus, Desiderius

Desiderius Erasmus, Dutch Humanist, scholar of classics and patristics, and the first editor of the Greek version of the New Testament, was the greatest scholar in what became known as the northern Renaissance.

Giraudon



Erasmus, oil painting by Hans Holbein the Younger, 1523. In the Louvre, Paris.

Early life and career. Erasmus was born at Rotterdam or, possibly, Gouda, in the Low Countries, October 26/27, in about 1466. He was the second child of Margaret, a physician's daughter, and Roger Gerard, a priest. Though a romantic story woven by Erasmus around these facts is suspect, the legal impediments of illegitimacy did haunt his later career. He went first to school at Gouda, where Peter Winckel (later one of his three guardians) was his rather obtuse usher (assistant teacher), and spent the years 1475–83, till his father's death, at the school of St. Lebwin's, Deventer (with a short spell as choirboy at Utrecht). There, though he could have had little contact with them, such teachers as John Sintheim (Sinthius) and Alexander Hegius, members of the Brethren of the Common Life, were opening the school to Humanist ideas. In this part of Holland the

influence of the so-called *devotio moderna* ("modern devotion"—a late medieval lay religious movement emphasizing education) was strong, and Erasmus owed much to its practical Christocentric piety, though he was to be repelled by the obscurantism and austerities of its latter end. When their guardians sent Erasmus and his elder brother, Peter, to school at 's Hertogenbosch, Erasmus found the prospect of a cloistered life and its discipline irksome. The brothers, however, were in no position to withstand their guardians' plans, and Peter entered a monastery at Sion, near Delft, while Erasmus became an Augustinian canon at Steyn and he was ordained a priest on April 25, 1492.

Though he now had an opportunity to study the classics and the poets, he felt the constraint of monastic life, which bred in him a passion for personal freedom and an irritation with restraint that became a trait of his complex character. Grasping eagerly at an opportunity to leave, probably in 1494, he became Latin secretary to Henry of Bergen, bishop of Cambrai, only to find that he did not take kindly to the sort of diplomatic career that suited many scholars of the age. He was glad when a friend secured permission for him to study theology at the University of Paris, where he entered the Collège de Montaigu at the time when the rector, the formidable Jan Standonck, was attempting to graft onto Paris the piety of the *devotio moderna* in its most austere form. It was too much for Erasmus, whose health gave way. Nor did his heart lie in the Scholastic theology or in preaching, in which he dabbled at this time. Like other Humanists, he was affronted by the dogmatic theologians, especially of the religious orders, with their party cries and intolerance and their violence and wooden hostility toward new methods. In Paris, however, he made acquaintance with the Humanists, and a commendatory epistle in Robert Gaguin's *De origine et gestis Francorum compendium* (1495) was his first writing to attract public attention. He returned to Paris from a convalescence in Holland and took pupils, first Christian and Henry Northoff from Lübeck and then the Englishmen Thomas Grey and Robert Fisher, for whose instruction he wrote the manuals *De conscribendis epistolis*, *De copia verborum*, *De ratione studii*, and *Colloquia familiaria* (*The Colloquies of Erasmus*), which underwent many revisions and became one of his most famous works.

Early works

First visit to England. What Erasmus really wanted was to be allowed to study in reasonable comfort, and he saw his various employments as of value only to that end. He had gradually to build up a precarious income (which always lagged behind his reputation) from pensions, gifts, and dedications, which in those years involved fulsome touting. His tutorial adventures introduced him to William Blount, Lord Mountjoy, who proved a steadier patron than most and whose invitation to England Erasmus accepted in 1499. Despite his often acid criticisms of England and the English, their barbarism, their weather, their beer, the country fascinated him and drew him back again and again. One reason, no doubt, was the patronage of Archbishop William Warham and John Fisher and even more the congenial friendship of such scholars as William Grocyn, Thomas Linacre, William Latimer, and, above all, John Colet and Thomas More. Of these last he left verbal portraits that have the vivid detail of a Holbein painting. Colet made an early vivid impression on him; his more intimate friendship with More ripened through the years. Though Colet was not the first to turn his thoughts from "good letters" to "sacred letters," he did challenge Erasmus to join him at Oxford in the battle against the Obscurantists and in the exposition of a more directly biblical theology. Erasmus rightly saw that he had not the technical equipment for such intervention, but he was now evidently more deeply concerned for "sacred letters." "I wished," he wrote, "that good letters should find that Christian character which they have lacked in Italy, and which, as you know, ended in glorifying pagan morality." Erasmus' new concern did not diminish his interest in the classical authors, for he stood for a combination of "good" with "sacred letters." He had long been acquainted with such poetic models as Virgil, Horace, Ovid, Juvenal, Statius, Martial, Claudian, Persius, Tibullus, and Propertius and in prose with Cicero, Quintilian, Sallust, and Terence. Throughout his life he edited works by Latin authors (*e.g.*, Pliny, Seneca) and produced translations of Greek classical writings (*e.g.*, Lucian).

The wandering scholar. In 1500 he left England, horrified at having to relinquish his earnings to the customs at Dover, and, in June, worked off some of his disgust by producing a collection of proverbs, the germ of his great *Adagia*. He spent the next months between France and Holland and in this unexciting period began arduously to gain that mastery of Greek that was to become the foundation of his later pre-eminence. His thoughts began to turn toward a great edition of Jerome, the archetype of his own ideal combination of good and sacred letters. At Saint-Omer he learned to know and admire the saintly Franciscan Jean Vitrier, who may have deepened in him a genuine strain of Pauline piety not to be overlooked. One result was the little *Enchiridion militis Christiani* (1503), which reflects some of the most attractive elements in his religion and at the close of which he wrote,

> Make the prophets, Christ and the Apostles your friends. Above all, choose Paul . . . it has long been my cherished wish to cleanse the Lord's temple of barbarous ignorance and to adorn it with treasures from afar, such as may kindle in generous hearts a warm love for the scriptures.

He continued to produce edifying discourses at intervals, though none of them attained the quality of the *Enchiridion*. There followed his *De interdicto esu carnium* (1522), *Exomologesis* (1524), *Modus orandi Deum* (1533), and *Christiani matrimonii institutio* (1526), dedicated to Catherine of Aragon. He commented on certain psalms in 1528 and at the end of his life discussed preaching in a comment on the book of Ecclesiastes.

In the summer of 1504 Erasmus found in the library of the monastery at Park, near Louvain, a manuscript of Lorenzo Valla's *Annotationes* on the New Testament, which criticized the Vulgate and advocated a critical method giving priority to exact, philological exegesis. The discovery was momentous for Erasmus, who published his own edition of this manuscript in 1505. He returned to England (1505–06), where he made his own copy of the Greek New Testament from such manuscripts as he could borrow. When a half-promised benefice failed to materialize, he seized an opportunity to visit Italy as tutor of the sons of Giovanni Battista Boerio, the physician to Henry VII. He took his D.D. at Turin (September 4, 1506) and in the next months saw at first hand the splendour and secularization of the church in Italy and those bloody consequences of papal diplomacy that much offended his temper and his conscience. He moved to Bologna and to Venice, where he collaborated with the great publisher Aldus Manutius. There, in the house of Andrea Asolani, he shared lodgings with the phenomenal young scholar Giralomo Aleandro. At Venice in September 1508 Erasmus published a great edition of his *Adagia*, now swollen to over 3,000 proverbs collected from the classical authors, a work that established his reputation as the foremost scholar in northern Europe. He became tutor to Alexander Stewart, the young archbishop of St. Andrews, and went with him to Padua and Siena. He visited Rome, but though he found friends at the papal court England drew him back, and he returned there to stay with Sir Thomas More. While there he wrote (1509) the *Encomium moriae* (*The Praise of Folly*, 1549) in the Lucianic style that was congenial to both him and More. The *Encomium moriae* drew not only upon ancient and medieval inspiration but also on Erasmus' own wit; it justifies J. Huizinga's comment that "only when humour illuminated that mind did it become truly profound," for the sharp satire became one of the most popular and enduring of his writings. He moved to the University of Cambridge, where he lectured on Greek and on the letters of Jerome. Archbishop Warham gave him the benefice of Aldington in Kent, which he commuted into a pension. But Erasmus could not settle anywhere for very long and was off again in 1514 to Basel, to a new fruitful partnership with the publishing house of

Erasmus in Italy and England

Johann Froben. About this time there appeared a scurrilous satire against the late pope—the *Julius exclusus*. The probabilities were slightly in favour of Erasmian authorship and were reinforced by the curiously evasive character of Erasmus' denials.

A shock came in June 1514, when an old friend Servatius Roger, now prior of Steyn, wrote re-calling him to the monastic life. His reply was a passionate plea for freedom, a defense of his own evident vocation for letters. It now became urgent for him to find authoritative support for his way of life, and he drew up an appeal to the papal chancery in which he put his case under the name of an imaginary Florentius. (This appeal, addressed to "Lambertus Grunnius," together with the letter to Roger and his own *Compendium vitae*, compiled in 1524, are primary sources of the life of Erasmus.) In January 1517 he received two dispensations from Pope Leo X, one allowing him to live in the world, and the other dispensing him from the dress of his order. On his return to England he received these dispensations formally on April 9, 1517, at the hands of his friend Ammonius Andreas, the papal legate.

Edition of the New Testament. In 1516 he had published his most immediately influential work, the edition of the New Testament, which gave priority to the Greek text. He compared such manuscripts as were available to him and various readings of the Fathers, while other scholars assisted with the philological apparatus, especially with Hebrew. It was perhaps less lack of energy that prevented him from mastering Hebrew than it was that he was as instinctively repelled by it as he was drawn by Greek. His general distaste for the Old Testament led him nearer than he knew to a view in which the moralists of the classical world supplied the place of the Old Testament. Although he did not have access to the earliest manuscripts, he provided a much more accurate text than that of the Vulgate. Alongside the Greek text he placed his own elegant Latin version, which greatly attracted scholars; and he appended critical notes, some of which were far from academic, many being as biting as the glosses of later Reformers. His prefatory essay was a noble plea for the study of the Scriptures, including the lines:

Views on knowledge of the Scriptures

> I could wish that every woman might read the Gospel and the Epistles of St. Paul. Would that these were translated into each and every language so that they might be read and understood not only by Scots and Irishmen, but also by Turks and Saracens . . . Would that the farmer might sing snatches of Scripture at his plough, that the weaver might hum phrases of Scripture to the tune of his shuttle, that the traveller might lighten with stories from Scripture the weariness of his journey.

The work was dedicated to Pope Leo X, who accepted the dedication. Nonetheless, it was strongly attacked. There were indeed obvious inaccuracies in the work, many of which were corrected in later editions. To forward-looking scholars the new version offered a more attractive, truer text, whereas its prefaces amounted to a manifesto of the new learning, expounding new critical method, and at the same time transmitting important gleanings from Erasmus' reading of the Fathers. Important too were his paraphrases, first of the Letters, then of the Gospels, the English translation of which had an important and official vogue in the reign of Edward VI. Erasmus has some claim to be regarded as the first New Testament scholar and in this lies his importance as theologian, scholar, and divine.

Years at Louvain. Erasmus spent the years 1517–21 at Louvain, where he took great interest in the new college for the study of the sacred languages founded by Jerome Busleiden. During this period his correspondence greatly increased and he became one of the most prolific letter writers among the Humanists for whom semiprivate correspondence of this kind afforded a prime means of learned conversation. A high proportion of the letters of Erasmus are distinguished from the bulk of Humanist correspondence only by their more polished Latin; but every now and again, writing to some intimate something new appears, pictures of contemporary life wittily observed, that Bruegel-like quality, which in his *Encomium moriae* and in his *Colloquia familiaria* anticipated the writing of a new age.

Years at Basel. In November 1521 he settled at Basel, perhaps the most satisfying to him of his many abiding places. From there he published a great series of works and became a central figure among those scholars for whom the return to the Fathers of the church was an integral part of the renewal of sacred letters. His edition of Jerome in nine volumes (1516) was a great achievement, but others followed: Cyprian (1520), Pseudo-Arnobius (1522), Hilary (1523), Irenaeus (Latin, 1526), Ambrose (1527), Augustine (1528), Chrysostom (Latin, 1530), Basil (1532), Origen (Latin, 1536). In some of the later editions Erasmus took little direct personal share. In Basel, too, he revised and extended the range of the *Colloquia familiaria* into which he contrived to pack an amazing amount of comment on contemporary manners, some of it "replete with malice spiteful."

Erasmus and Luther

The advent of the Reformation ended the Erasmian dream of a golden age of letters. His own criticism of the church had been formidable enough for Aleander to denounce him as the real author of schism and worse than the German reformer Martin Luther. To Erasmus, however, Luther must have seemed a typical dogmatic theologian, violent, aggressive. Luther in his first literary contacts (they never met) was polite, though he soon diagnosed a deep divergence of which the Erasmian preference for Jerome against Augustine was symptomatic.

As the church struggle widened, both sides appealed to Erasmus, who was at one time hailed as the author both of Luther's *De captivitate Babylonica ecclesiae praeludium* (1520) and of Henry VIII's reply. Yet it counted for Luther that in critical months during 1519–21 Erasmus did not come out against him and indeed offered his services and mediation; while in an interview with Frederick the Wise of Saxony, Erasmus himself may have stiffened the resolve of the prince that Luther should not be refused a hearing at the Diet of Worms. In the next years Erasmus was more and more alienated from the Reformers, though some of them, including Huldrych Zwingli, John Oecolampadius, Martin Bucer, and Philipp Melanchthon, had been among his fervent admirers. A long dispute with the French classical and biblical scholar Jacques Lefèvre d'Étaples had shown Erasmus to be hypersensitive to criticism, and in 1523 a bitter polemic with the German Humanist and writer (also a vigorous supporter of Luther's reforms) Ulrich von Hutten resulted in *Spongia Erasmi . . . adversus aspergines Hutteni* ("Erasmus' Sponge Against Hutten's Mire"), the least creditable of all his writings. In 1524 Erasmus yielded at last to the prompting of his patrons and in great haste attacked Luther in *De libero arbitrio*. The work incensed his opponents without much satisfying his friends, and, though an able enough essay in Erasmian moderation, it exposed too his theological limitations. Luther's reply, the *De servo arbitrio* (1525), violent and perverse, is in a more profound dimension. Erasmus worked off his resentment in the tiresome *Hyperaspistes* (1526), but he much improved a second edition, which contains some of his finest mature utterances.

Last years. Though the triumph of the Reformation in Basel under Oecolampadius in 1529 was less violent than elsewhere, Erasmus with other Humanists left and went to Freiburg in Breisgau; there he stayed until 1535 when, aging and unwell, he returned to Basel. Although an imperial councillor, who from the time of writing his *Institutio principis Christiani* (1516) had sought to advise the secular powers, he did not attend the Diet of Augsburg. But his interest in the peace of the empire and of the church brought from him in 1533 *De amabili ecclesiae concordia*, and among those who were trying to heal the breach there were moderate men who were his disciples. A new pope, Paul III, nominated him as dean of Deventer, and at last there was talk of a cardinal's hat. But it was all too late in the day, for Erasmus had outlived his preoccupation with dignities and turned instead to paraphrase a psalm for a humbler but loyal friend, Christopher Eschenfelder, the customs officer at Bop-

pard; to set his affairs in order (he had sold his library in 1534 to the Polish nobleman Johannes à Lasco); and to prepare for an end that could not be far off. He died in Basel on July 12, 1536. That he died without the sacraments shocked Luther, but Bonifaz Amerbach, his heir, who was with him to the end, wrote of his death, "As was his life, so was the death of this most upright of men. Most holy was his living, most holy his dying." Poignantly the last words of the great cosmopolitan, *Lieve God* ("Dear God"), were in Dutch.

MAJOR WORKS

THEOLOGICAL WORKS AND COMMENTS ON TOPICAL PROBLEMS: *Enchiridion militis Christiani* (1503; *The Manuell of the Christen Knyght*, trans. by W. Tyndale, 1533); *Ratio seu compendium verae theologiae* (1518); *De libero arbitrio* (1524; *Discourse on Free Will*); *Hyperaspistes diatribae adversus servum arbitrium M. Lutheri* (1526); *Christiani matrimonii institutio* (1526); *Dialogus, cui titulus Ciceronianus, sive, de optimo dicendi genere* (1528); *De sacrienda ecclesiae concordia* (1533); *Modus orandi deum* (1533).

EDUCATIONAL WORKS: *De ratione studii et instituendi pueros* (1512); *Institutio principis Christiani* (1516; *The Education of a Christian Prince*); *Ecclesiastae sive de ratione concionandi libri quatuor* (1535; *Ecclesiastes; or, the Preacher: An Essay on the Duties of a Public Religious Instructor*, 1797).

BIBLICAL WORKS AND EDITIONS OF THE FATHERS: *Novum instrumentum* (1516); *Omnium operum divi Eusebii Hieronymi* (1516); *D. Caecilii Cypriani opera* (1519); *Sancti Aurelii Augustini opera* (1529); *Sancti Basilii opera omnia* (1532).

OTHER WORKS: *Adagia* (1500; extended edition, 1508); *Encomium moriae* (1509; *The Praise of Folie*, trans. by Sir Thomas Chaloner, 1549; republished 1965); *Colloquia familiaria* (1518; *The Colloquies of Erasmus*); *Compendium vitae Erasmi* (1607); *Opus epistolarum D. Erasmi Roterodami*, ed. by P.S. Allen, H.M. Allen, and H.W. Garrod (1906–58).

BIBLIOGRAPHY. J.A. FROUDE, *Life and Letters of Erasmus* (1895, reprinted 1971), perceptive and still valuable; PRESERVED SMITH, *Erasmus: A Study of His Life, Ideals and Place in History* (1923, reprinted 1962), a useful biography with still more useful and plentiful quotations; JOHAN HUIZINGA, *Erasmus of Rotterdam* (1924, reprinted 1952), a brilliant analysis; P.S. ALLEN, *The Age of Erasmus* (1914, reprinted 1963); ROLAND H. BAINTON, *Erasmus of Christendom* (1969), embodies recent research in a scintillating narrative; AUGUSTIN RENAUDET, *Études érasmiennes* (1939), a basic source for French-reading students; MARCEL BATAILLON, *Érasme et l'Espagne* (1937); ERNST W. KOHLS, *Die Theologie des Erasmus* (1966), the only recent survey of Erasmus's theology; MARGARET M. PHILLIPS, *Erasmus and the Northern Renaissance* (1949), a weak account of theology, but splendidly sympathetic and a good historical study; also by Phillips, *The "Adages" of Erasmus* (1964). Important English editions of Erasmus's work include CRAIG R. THOMPSON, *The Colloquies of Erasmus* (1965), a fine collection of extracts and commentary; E. GORDON RUPP and PHILIP S. WATSON, *Luther and Erasmus* (1969), on the controversy between Erasmus and Luther in 1525 with commentary and essays; and paperback translations of *Praise of Folly*, *Selected Letters*, and *Selected Writings*.

(E.G.R.)

Ericales

The Ericales is an order of about 2,700 species of plants, largely shrubs or small trees, arranged in ten families and about 135 genera.

The number of species used as ornamentals is inordinately large when compared with other plant groups of similar size. Many of the most beautiful and prized horticultural plants throughout the world are found in the Ericales, for example, azaleas and rhododendrons. Blueberries and cranberries are economically significant; there are other species of less commercial value.

GENERAL FEATURES

Structural diversity and distribution. Although predominantly woody shrubs and small trees, the order includes herbs, undershrubs, vines, epiphytes (plants that grow upon other plants), herbaceous saprophytes (nonphotosynthetic plants that obtain nourishment from dead organic matter), and root parasites. A few tree species attain heights of 75 to 100 feet (20 to 30 metres).

Worldwide distribution

Ericads are very widely distributed in the tropics and throughout temperate regions of the Northern and Southern Hemispheres and are a significant component of the sub-Arctic circumpolar flora. A few extend into the Arctic and the Antarctic regions. A number of genera are indigenous to southern Africa, and the large family Epacridaceae is in a sense the Australian segment of the order.

It has been customary to think of the species of the Ericales as inhabitants of temperate and cooler regions, with tropical representatives being confined to the cool, high mountainous regions. A majority of species do occupy these areas, but the recent inclusion of the families Actinidiaceae and Saurauiaceae in the order has extended the group's range into typically tropical situations of Asia and America. As a general rule, ericalean plants prefer acid soils.

Economic importance. Ericads have served as tea substitutes; early explorers of Canada and the eastern and northern regions of the United States, for example, brewed tea from the leaves of Labrador tea, or bog tea (*Ledum palustre*); Lapland rosebay (*Rhododendron lapponicum*); and checkerberry, or teaberry (*Gaultheria procumbens*). The Ojibwa Indians similarly used bog rosemary (*Andromeda glaucophylla*) and leatherleaf (*Chamaedaphne calyculata*).

For centuries the best material for making briar pipes has been wood from burls of *Erica arborea* or related Mediterranean species.

Edible berries

Ericalean plants have been and still are most important to man as foodstuffs and as cultivated ornamentals. Checkerberry (*Gaultheria procumbens*), certain of the bearberries (*Arctostaphylos* species), and creeping snowberry (*Chiogenes hispidula*) produce edible fruits. In Asia the fruit of the tara vine (*Actinidia arguta*) is much sought after. But it is the huckleberries, the blueberries, and the cranberries that are most prized as fresh fruits or as jellies or preserves.

Cranberries (*Vaccinium macrocarpon*) are cultivated in Massachusetts, especially on Cape Cod; in New Jersey; in Wisconsin; and in the Pacific Northwest. They are also harvested from wild stands in these areas. The tart fresh berries are processed with sugar and sold as cranberry sauce or jelly. Cranberries, first cultivated in New England in the early 19th century, are still one of the leading exports of Massachusetts. Cranberry sauce and jelly have come to be thought of as uniquely American, but the Scandinavians hold in high esteem their own native cranberry (*V. vitis-idaea*), the berry of which is spicier than the American fruit.

Huckleberries (*Gaylussacia baccata*) are harvested only from wild stands in the eastern United States and Canada. Although the flesh of the fruit is generally considered to be sweet and delicious, the "berry" is actually a drupe, or stony-seeded fruit, with ten large, hard nutlets, making it less desirable and marketable than blueberries.

Blueberries of many types are harvested from wild stands in New England and the eastern states and from northern California into western Canada. Several species of *Vaccinium* are involved. They range from lowbush to highbush species and from diploid to hexaploid forms (*i.e.*, genetic variants with two to six sets of chromosomes, respectively). Blueberries, with their small and insignificant seeds, are widely considered a delicacy; in constant demand, their commercial production is economically significant, especially in New England.

Ornamental uses

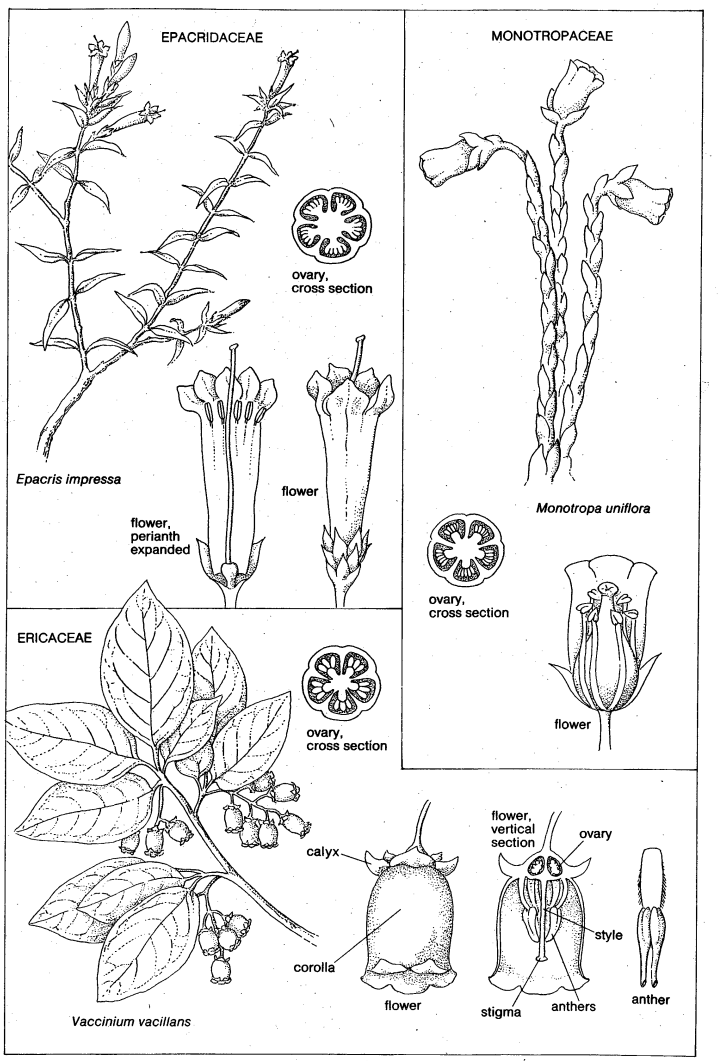
The other major use of ericads is as ornamentals. In Asia some of the vinelike forms (*Actinidia arguta* and *Clematoclethra* species) are cultivated for their deep green foliage. Several species of sweet pepper bush (*Clethra*) are used horticulturally to take advantage of the highly fragrant summer flowers. The leatherwood (*Cyrilla racemiflora*) and buckwheat tree (*Cliftonia monophylla*) are used similarly and also are desirable for their showy autumn coloration of foliage. The sourwood (*Oxydendrum arboreum*), one of the few trees in the order, is also cultivated because of the brilliant colour of its foliage in autumn. In Europe and New England, the heather (*Calluna vulgaris*) and some of the low bushy

heaths (*Erica* species) are used as garden border plants. Many ericads are planted in gardens as novelties or to provide variation in foliage colour and texture.

The most important ornamentals in the entire order, however, are found in the genus *Rhododendron*, the azaleas and the rhododendrons, about 850 species of which are or have the potential to be ornamentals. Some of the most spectacular flower clusters known are produced by these shrubs, and they are widely used not only for the beauty of their blooms but also for the value of the many evergreen species as effective background shrubs. Natural species are cultivated, and many hybrids have originated in cultivation; it is thus difficult to develop a classification that includes all species, varieties, and forms. It is safe to say, however, that there are at least 1,000 varieties of rhododendrons and azaleas in cultivation.

Noxious ericads. The order is not noted as a group of noxious plants, although a few species can be so described. Some contain a toxic substance identified as andromedotoxin. This compound is known to occur in certain species of *Rhododendron*, *Leucothoë*, *Menziesia*, *Ledum*, and *Kalmia*, and it is probably more widespread in the order than is now known. Humans can be poisoned by chewing on leaves and twigs or by sucking nectar from the flowers of these plants. Poisoned honey has been reported in areas in which bees visit large *Rhododendron* or laurel (*Kalmia*) stands, but the honey is so bitter that very little of it can be eaten. The poison causes watering of the mouth, eyes, and nose; vomiting; slow pulse; and depressed blood pressure. Massive ingestion of andromedotoxin results in convulsions and slow and progressive paralysis until death; fatalities, however, are rare.

Several species of laurel (*Kalmia*) long have been known to be poisonous to domestic animals. *Kalmia angustifolia* is known colloquially as sheepkill, lambkill, or calfkill, and one common name of *K. latifolia* is poison laurel. Sheep and cattle may become ill and die from eating foliage of these plants, but deer do not seem to be affected. The course of the disease parallels that in humans who have taken massive ingestion of andromedotoxin—convulsions, progressive paralysis, and death. Laurel poisoning is not a serious problem in sheep or cattle, however, probably because the tough, leathery leaves are not desirable except in the absence of more palatable foliage. Cases of poisoning are more frequent in sheep than in cattle.



Representative plants from three of the principal families of the order Ericales.

Drawing by M. Pahl based on (Epacridaceae, Ericaceae all but ovary) G.M.H. Lawrence, *Taxonomy of Vascular Plants* (1951); The Macmillan Company, (Ericaceae ovary) J. Hutchinson, *The Families of Flowering Plants;* The Clarendon Press, Oxford

NATURAL HISTORY

Pollination. Compared with some of the very highly specialized plant groups, such as the orchids, few unusual or unique features exist in the life cycle of the ericads. There has been only the slightest tendency toward development of irregular flowers, and highly involved and complicated methods of pollination are rare.

Probably most ericads may be pollinated in more than one way, but studies on pollination in the Ericales have been sketchy. In the heather (*Calluna vulgaris*), the nectary in the flower is a continuous, fleshy ring around the base of the ovary (female structure), and the lower parts of the petals and stamen filaments (basal parts of the male structures) are also succulent. An insect forcing its proboscis into the flower shoves the anthers (pollen sacs) apart. Anther chambers open laterally, and hornlike appendages attached to the anthers help deposit pollen on the insect's proboscis by acting as levers. Pollen is carried to another flower by the insect and deposited upon the stigma (pollen receptive surface of the female structure) as the insect probes for nectar in the flower.

Insect pollination

Nectar production decreases near the end of the pollination period. The filaments then elongate and extend the open anthers beyond the confines of the flower; thereafter wind pollination occurs.

Another kind of pollination also takes place in the same heather flower. Certain small insects (*Taeniothrips ericae* [Thysanoptera]) appear to be present at all times in *Calluna* flowers. These tiny insects are very lively, move about actively, and reach parts of the flower inaccessible to larger insects. Pollen discharged during their activity adheres to their bodies, which become sticky with nectar. The males of *T. ericae* are much rarer than the females and cannot fly. Females, therefore, take the aggressive role and fly from flower to flower searching for males. They depart from projecting stigmas, their bodies heavily laden with pollen, and presumably land on stigmas, thereby probably being able to effect both self- and cross-pollination.

The laurels (*Kalmia* species) have two pouches in each of the five partly fused petals. Anthers are retained under tension in these pockets until the flower opens; then the filaments snap the anthers up, and pollen is forcibly ejected. Laurel flowers are visited by bees that almost certainly effect pollination. Presumably elevation of the anthers above the saucer-shaped flowers of *Kalmia* also permits wind pollination.

Mycorhizal relationships. Mycorhizae are structures formed by plant roots and the fungi that live in close association with them. Mycorhizal associations are common and widespread in the Ericales. It has been suggested that probably all ericads are in some mycorhizal relationship. Although considerable research has been done on the nature of the relationship, much of the work has been challenged and is of doubtful validity. In most instances, the species with an obligate relationship and those without cannot be identified with certainty; the relationship is obligate, however, in the few members of the family Pyrolaceae that occasionally lack chlorophyll and in the family Monotropaceae, all of whose members lack chlorophyll. In addition to these families, mycorhizae have been reported in the Clethraceae, Epacridaceae, and

Empetraceae. The phenomenon, which is widespread in the Ericaceae, probably occurs in all families of the order.

The structure of the ericalean mycorhiza has been described as being intermediate between the conditions called ectotrophic and endotrophic, and the term ectendotrophic has been used to describe the mycorhiza of *Monotropa*. Host roots are enveloped by masses of fungal hyphae (filaments of fungus cells) as in ectotrophs, but hyphae penetrate the cells of the host root tissues, as in endotrophs.

The fungi involved in the relationship are members of the fungal classes Deuteromycetes and Basidiomycetes. A number are of the genus *Rhizoctonia*. Little is known as yet of the physiological relationships between the fungus and the plant. The view that species of *Phoma*, a deuteromycetes associated with some ericads, serve to fix free atmospheric nitrogen appears to be unfounded. There are advocates of a theory that fungi may serve in some capacity to synthesize various carbon-containing materials from the soil, which are then passed along to the host plant. This explanation is especially attractive in explaining saprophytic situations of nongreen ericads. It must be assumed that the fungus utilizes organic compounds from the soil and transfers its products in forms useful to the plant.

EVOLUTION AND CLASSIFICATION

Fossil record and phylogeny. Fossil records of ericads are not numerous, especially considering the number of species in the order. All have been found in deposits of the Tertiary Period (2,500,000 to 65,000,000 years ago) and mostly from the Oligocene Epoch (26,000,000 to 38,000,000 years ago) and the Miocene Epoch (about 7,000,000 to 26,000,000 years ago).

Fossil leaves of *Saurauia* have been found in Alaska, and *Actinidia* leaves have been recovered from California. Miocene layers in Oregon have yielded leaves of *Arbutus trainii*, which closely resemble those of the living madrona (*A. menziesii*) of the Pacific Coast. Fossil records of *Arctostaphylos*, *Vaccinium*, and *Andromeda* also have been reported.

Fossil pollen grains

Many fossil pollen grains have been taken from Miocene deposits at Rypin, Poland. In some cases, some of them have been assigned to the Ericales because of structural resemblances; in other cases, they have been so assigned because of structure and the fact that they occur in tetrads (groups of four), a condition common in living ericads. *Cyrilla* appears to be represented by three species. Several pollen tetrads have been assigned to the Ericaceae, but only those of the genus *Rhododendron* are clearly identified.

The fossil record is of little help in tracing phylogenetic relationships within the order or between the Ericales and other orders. Any such relationships must be determined largely from comparative morphological studies of living forms. Close affinities between the order Ericales and the order Diapensiales are recognized, and a common ancestry for the two is postulated in the order Theales. A close relationship of these groups has been recognized for some time.

Classification. *Distinguishing taxonomic features.* Although there are exceptions to most features, ericads generally are shrubs with alternate, simple, leathery leaves. Their flowers have functional structures of both sexes and are regular (radially symmetrical) with a four- or five-parted corolla (collection of petals) of partly fused petals. Stamens number twice the number of corolla lobes and usually are attached directly to the receptacle (expanded stalk tip that carries the flower parts). Anthers open by terminal pores, and pollen is discharged in tetrads. The pistil (female structure) is compound with one style (upper, narrow, elongated part of pistil). The ovules are numerous and are attached along the central axis of the ovary. The only characteristics shared by all ericalean species are that the ovule has but one integument, or covering cell layer, and the nucellus (central part of ovule) is a single cell layer thick.

Variations in the above characteristics provide the bases for establishing families and for determining relationships between them. Major variables of taxonomic significance are: number of floral parts, degree of fusion of floral parts, degree of fusion of floral parts with the ovary, type of anther dehiscence (opening), structure of pollen, structure of ovary, and type of fruit.

Annotated classification. The classification system used below is a recent one that differs from more traditional schemes mainly in the inclusion of the family Grubbiaceae.

ORDER ERICALES

Herbs, shrubs, trees, herbaceous saprophytes, or root parasites. Leaves alternate (also opposite or whorled in Ericaceae; opposite in Grubbiaceae), simple, commonly leathery, without stipules (basal appendages). Flowers mostly bisexual, regular or slightly irregular. Calyx of 4 to 10 (rarely fewer) separate or partly united sepals. Corolla of 4 to 10 (rarely fewer) separate or partly united petals or, in a few instances, lacking. Stamens usually as many or twice as many as petals or sepals, very rarely more numerous, and sometimes partly reduced to staminodia (sterile stamens). Ovary of several, usually 2 to 7, fused carpels (ovule-bearing segments) and 2 to 7 (rarely more) chambers, infrequently with incomplete partitions; rarely with 1 chamber. Ovary superior, partly inferior, or inferior (positioned above, partly enclosed, or enclosed by the other flower parts). Styles united except in rare cases. Ovules usually, but not always, numerous. Ovule attachment usually axile (central in Grubbiaceae). Fruit a berry, drupaceous, or a capsule. Seed with copious fleshy endosperm (nutritive tissue). Ten families, about 135 genera, and some 2,700 species.

Family Saurauiaceae (saurauia family)

Trees or shrubs clothed with bristly hairs. Leaves alternate, simple, frequently serrate (toothed). Flowers mostly bisexual, regular, hypogynous. Sepals 5, separate, greatly overlapping. Petals 5, separate or fused into a short tube at the base. Stamens numerous, attached to petals. Ovary superior with 3 to 5 locules (chambers). Styles 3 to 5, separate or partly to completely united. Ovules numerous. Fruit a 3- to 5-locular berry, rarely somewhat dry and dehiscent. Seeds small. One genus (*Saurauia*) with about 250 species in tropical and subtropical America and Asia.

Family Actinidiaceae (actinidea family)

Trailing or climbing shrubs. Leaves alternate, simple, more or less rounded, sometimes serrate. Flowers bisexual or unisexual; regular. Sepals 5, separate and overlapping. Petals 5, separate. Stamens 10 or more. Ovary 3-locular or more, the carpellary walls sometimes hardly reaching the central axis. Ovules 2 or more in each locule. Styles 5 and completely united or many and spreading. Fruit a dry capsule or a berry. Seed small. Three genera with about 25 species of tropical (a few in temperate) East Asia.

Family Clethraceae (pepperbush family)

Tall shrubs or low trees. Leaves alternate, simple, often serrate. Flowers very fragrant or aromatic, bisexual, regular. Sepals 5, partly fused, persistent about the fruit. Petals 5, separate, forming a saucer-shaped corolla. Stamens 10 (sometimes 12), separate. Pistil with 3 carpels. Ovary superior with 3 locules, mostly 3-lobed. Style 1. Ovules numerous. Fruit a 3-valved capsule. Seed sometimes winged. One genus (*Clethra*) with about 30 species native to eastern North America, Brazil, Mexico, and East Asia.

Family Ericaceae (heath family)

Shrubs; subshrubs; occasionally woody, perennial herbs; or small trees. Leaves opposite, alternate, or whorled; simple; often leathery and persistent. Flowers bisexual, regular or slightly irregular. Sepals 4 to 5, separate or partly fused, usually persistent, sometimes fused with the ovary. Petals 4 to 5, rarely separate, usually partly fused. Stamens as many or more commonly twice as many as the petals, inserted on the outer edge of a disc that surrounds the ovary at its base or higher up. Pistil 4 to 10 carpellate. Ovary superior, partly inferior, or inferior, 4- to 10- (typically 5-) loculate. Ovules usually numerous. Stigma and style 1. Fruit a capsule, berry, or drupe. Seed usually small. Seventy to 80 genera with 1,500 to 1,900 species widely distributed on acid soils throughout temperate Northern and Southern Hemisphere regions, in sub-Arctic regions, and in high mountains in the tropics.

Family Epacridaceae (epacris family)

Shrubs or small trees. Leaves alternate, often crowded, simple, small leathery, usually stiff. Flowers bisexual (rarely unisexual), regular. Sepals 4 to 5, partly fused, persistent. Petals 4 to 5, partly fused, or rarely petals separate. Stamens usually 5, but sometimes 4, alternating with the petals, and sometimes with alternating staminodia. Pistil 4- to 5-carpellate. Ovary superior with 1 to 10 locules often surrounded at the

base by a disc. Ovules 1 to many in each locule. Style 1, stigma capitate. Fruit a 5-valved capsule or a drupe. Twenty-four genera with about 340 species, mostly in Australia, New Caledonia, New Zealand and the Antarctic islands, a few species in the islands of the Pacific, in India, and the Malay Archipelago, and a single species in the mountains of South America.

Family Pyrolaceae (pyrola family)
Evergreen, perennial herbs with creeping rootstocks. Stem above ground short, sometimes slightly woody at the base. Leaves alternate, almost opposite, or in false whorls, simple, leathery, and mostly toothed. Flowers bisexual, regular or slightly irregular. Sepals 5, separate or slightly united. Petals 5, waxy, separate, hypogynous. Stamens 10, separate. Pistil compound, 4- to 5-carpellate. Ovary superior, 4- to 5-loculate. Ovules numerous. Style 1. Fruit a capsule. Seed numerous. Ten or 11 genera with perhaps as many as 45 species of the Northern Hemisphere.

Family Monotropaceae (Indian pipe family)
Saprophytic or root parasitic herbs, almost or, more commonly, completely lacking chlorophyll, usually white, brown, purple, red, yellow, or pink, mostly succulent. Stems single or clustered from matted roots. Leaves short or rarely elongated scales. Flowers borne singly, erect, or nodding, bisexual (very rarely unisexual), regular or slightly irregular. Sepals 2 to 6, separate. Petals 3 to 6, separate or partly united (petals absent in *Allotropa*). Stamens 6 to 12. Pistil compound of 4 to 6 carpels. Style 1. Ovary superior, mostly 4- to 6-loculate, or sometimes 1-celled. Ovules numerous. Fruit a 4- to 6-celled capsule or, rarely, 1-celled, somewhat fleshy. Seed numerous, minute. About 9 genera with about 16 species chiefly, if not all, North American.

Family Cyrillaceae (cyrilla family)
Deciduous or evergreen shrubs or small trees. Leaves alternate, simple, frequently leathery. Flowers bisexual, regular. Sepals 5, separate and overlapping or fused at the base. Stamens 10, in 2 whorls of 5 each, the inner whorl sometimes reduced to staminodia or even missing. Pistil compound, 2- to 4-carpellate. Ovary superior with 2 to 4 locules. Ovules 1 or 2 in each locule. Fruit a dehiscent capsule or a spongy berry. Three genera with about 12 species. Monotypic species (genera with only one species) of *Cyrilla* and *Cliftonia* are indigenous to the southeastern United States. *Costaea* has several species extending from Cuba into South America.

Family Empetraceae (crowberry family)
Evergreen, small, cushion-shaped, heathlike shrubs. Leaves alternate, linear, the lower surface deeply grooved. Flowers usually bisexual, irregular. Sepals 2 to 6, often petaloid, bracteate (resembling bracts, small green structures), sometimes absent, separate, often overlapping, in 2 whorls. Petals absent (inner whorl of sepals interpreted by some to be petals). Stamens 2 to 4, separate, alternating with the sepals. Pistil 1, 2- to 9-carpellate. Ovary superior, 2- to 9-loculate. Ovules 1 in each locule. Style 1, variously branched into 2 to 9 lobes. Fruit a fleshy or dry berry with 2 or more 1-seeded pyrenes (nutlet-like structures). Three genera with 8 species native to the mountains of North and South America extending into the Antarctic, the Arctic, and northern circumpolar regions.

Family Grubbiaceae (grubbia family)
Heathlike shrubs. Leaves opposite, simple, linear or lance shaped, leathery. Flowers bisexual, regular. Perianth (sepals and petals) 1- to 4-lobed of undifferentiated segments. Stamens 8, in 2 whorls of 4 each. Pistil 1, 2-carpellate. Ovary inferior, 1-loculate at maturity, with 2 pendulous ovules. Style 1, stigma slightly bifid (two-parted). Fruit a drupe with a crown composed of the disc and style. One genus (*Grubbia*) with 4 species endemic to southern Africa.

Critical appraisal. Close relationships among the families Clethraceae, Pyrolaceae, Monotropaceae, Ericaceae, and Epacridaceae have been recognized for many years, although taxonomists have not agreed in their treatment of the families. The family Monotropaceae frequently is combined with the Pyrolaceae, and both families are sometimes included in the Ericaceae. The family Vacciniaceae which is composed of those species of the Ericaceae that have partly to wholly inferior ovaries is sometimes recognized.

In any of these manipulations, however, the five "core families" noted above comprise some 2,200 species, or more than 80 percent of the approximately 2,700 species of the order.

Inclusion of the family Empetraceae in the Ericales has become almost universally accepted in recent years, although one important taxonomist retains it in the order Celastrales.

Fairly recently the family Cyrillaceae has been placed in the Ericales, an arrangement that is gaining acceptance. Previously, it was assigned to the Celastrales, and, in some systems, it remains there.

Five principal families stable in all taxonomic systems

The families Actinidiaceae and the Saurauiaceae are obviously closely related to each other, and it is common to find the Saurauiaceae within the Actinidiaceae. These families have affinities with both the orders Dilleniales and Theales, and modern taxonomic treatments are about equally divided in placing them in one or the other order. Referred to the Ericales, as is done here, they are the most primitive families in the order and may be regarded as near ancestral to the families more traditionally comprising the Ericales.

Inclusion of the family Grubbiaceae in the Ericales has been to date quite widely rejected. The family consistently has been placed in the order Santalales for a great many years. The main stumbling block to general acceptance of the Grubbiaceae in the order Ericales is the structure of the pistil. The ovary is inferior and, as it begins its development, is more or less two-locular (chambered). As the ovary continues to mature it becomes unilocular with a standing, free-central placenta that bears two pendulous ovules. This ovular structure is so typical of the order Santalales and so completely atypical of the order Ericales that few taxonomists consider the ericalean characteristics of the Grubbiaceae to outweigh the santalalean structure of the ovary.

BIBLIOGRAPHY. G.C. AINSWORTH and A.S. SUSSMAN (eds.), *The Fungi: An Advanced Treatise*, 3 vol. (1965–68), a discussion of mycorhizal relationships; ERIK BJORKMAN, "*Monotropa Hypopitys* L.—an Epiparasite on Tree Roots," *Physiologia Pl.*, 13:308–327 (1960), an original description of mycorhizal role in root parasitism of *Monotropa* on roots of other plants; C.G. BOWERS, *Rhododendrons and Azaleas: Their Origins, Cultivation and Development*, 2nd ed. (1960), a comprehensive treatment of native and cultivated rhododendrons and azaleas of the world; A. CRONQUIST, *The Evolution and Classification of Flowering Plants* (1968), a modern description of taxonomy and phylogenetic relationships of flowering plant families, which are arranged generally in conformance with Takhtajan's system; J. HUTCHINSON, *The Families of Flowering Plants*, 2nd ed., 2 vol. (1959), a more traditional taxonomic treatment of flowering plant groups than that followed in this article; C.J. HYLANDER, *The MacMillan Wild Flower Book* (1954), excellent illustrations and descriptions of some of the more beautiful ericads; *North American Flora*, Series 1, vol. 29, pt. 1 (1914), extensive descriptions of ericad families of North America; ALFRED REHDER, *Manual of Cultivated Trees and Shrubs, Hardy in North America*, 2nd ed. rev. (1940), classic taxonomic treatment of woody ericads under cultivation; R.W. SCHERY, *Plants for Man* (1952), an interesting discussion of economic aspects of ericalean and other plants; A. TAKHTAJAN, *Flowering Plants: Origin and Dispersal* (1969; Eng. trans. from the 2nd Russian ed., 1961), an exposition of the classification system followed throughout in this article.

(H.T.C.)

Erskine, Thomas Erskine, 1st Baron

Renowned in his own time for an unrivalled forensic skill, Thomas Erskine (1750–1823) made important contributions to the defense of personal liberties during his brilliant career as a British lawyer.

Early life and career. He was born Jan. 10, 1750, in Edinburgh, the youngest son of Henry David Erskine, 10th earl of Buchan. Upon completion of his preliminary education, though he wanted to enter a learned profession because of the straitened financial circumstances of his family, he sought a career in the Royal Navy instead. He became a midshipman in 1764 but left the service in 1768 and purchased a commission in a regiment of the 1st Royals. He added to his education by extensive reading, particularly in English literature. His unsigned pamphlet, "Observations on the Prevailing Abuses in the British Army" (1772), gained a wide audience. Finding opportunities for advancement in the British army no more favourable than in the navy and encouraged by the friendly interest of Lord Mansfield, Erskine decided to

enter the law. He was admitted to Lincoln's Inn in 1775, and in 1778 he received an honorary M.A. degree from Trinity College, Cambridge, after which he was called to the bar.

By courtesy of the National Portrait Gallery, London

Erskine, oil painting by Sir William Charles Ross (1794–1860). In the National Portrait Gallery, London.

Professional life. Within a few months, his future was assured by his defense of Capt. Thomas Baillie, lieutenant governor of Greenwich Hospital, who had published charges of corruption in the administration of the hospital. Those accused instituted a proceeding to show cause why Baillie should not be prosecuted for criminal libel. Erskine was retained by Baillie as his junior counsel and in his first appearance at the bar vindicated his client with remarkable eloquence and courage. He very quickly rose to the leading position in the English bar. In the following year, he successfully assisted the defense in the court-martial of Adm. Augustus Keppel. His successful defense of Lord George Gordon on the charge of high treason for instigating the anti-Catholic riots of 1780 substantially destroyed the English legal doctrine of constructive treason—*i.e.*, treason imputed to a person from his conduct or course of actions, though none of his separate actions amounts to treason. Erskine appeared in most of the major cases that arose out of the disruption of commercial relations with France, which had entered the United States War of Independence against Britain in 1778.

Defense of Lord George Gordon

In 1784 he unsuccessfully represented a clergyman defending a charge of criminal libel, but his contention that it is for the jury, not the judge, to determine whether a publication is libellous was vindicated by the passage of the Libel Act of 1792. In 1789 he won an acquittal for a bookseller who was charged with criminal libel for selling a pamphlet criticizing the trial of Warren Hastings, a former governor general of India who was impeached for alleged misconduct. Erskine's speech on that occasion is one of the monuments in the literature of English freedom. His unsuccessful defense of Thomas Paine, whom William Pitt, the prime minister, had caused to be indicted for treason for publishing *The Rights of Man*, cost him his position as attorney general to the Prince of Wales.

His defense of various politicians and reformers on charges of treason and related offenses placed a powerful check on the repressive measures taken by the ministry of William Pitt in response to the insecurity and hysteria engendered in England by the French Revolution and its aftermath. In 1800, by establishing the defendant's insanity, he successfully defended James Hadfield, who had attempted to assassinate George III. Erskine's argument at the trial is an important contribution to the law of criminal responsibility.

Erskine, who was an intimate of the Whig leaders Charles James Fox and Richard Brinsley Sheridan, the playwright, sat in the House of Commons from 1783 to 1784 and from 1790 until he became a peer in 1806. His undistinguished parliamentary career was almost wholly devoid of the forensic triumphs that marked his legal practice. In 1806 he became lord chancellor during the so-called Ministry of All the Talents. Erskine was not well qualified for the position either by training or experience. At the same time, his tenure of 14 months in office was not sufficiently long to permit any outstanding contribution.

Lord chancellor

His latter years were marked by private sorrows and misfortunes, which caused him almost completely to withdraw from public affairs. Toward the close of his life, however, he again achieved widespread prominence by his role in defense of Queen Caroline, whom her husband, George IV, had brought to trial before the House of Lords for adultery in order to deprive her of her rights and title. Erskine died at Almondell, Linlithgowshire (now West Lothian), Scotland, on Nov. 17, 1823.

Erskine's published works include his pamphlet "The Unconditioned Freeness of the Gospel in Three Essays" (1778); a student declamation published in 1794 in a pamphlet entitled "Sketch of Erskine, with Anecdotes: A View of the Causes and Consequences of the Present War with France" (1797); *Armata, a Fragment* (1817), written in the style of *Gulliver's Travels*; a burlesque of Gray's "Bard" entitled "The Barber" in *Poetical Works* (1823).

Published works

Numerous collections of Erskine's courtroom speeches have been published, the most useful of which include a collection edited by James Ridgway in 1847 and another edited by James L. High in 1874.

Assessment. Erskine excelled principally as a jury lawyer. His courtroom speeches are characterized by vigour, cogency, and lucidity and often by great literary merit. They consistently reveal a courageous and passionate dedication to the causes he represented. It was the judgment of Lord Campbell, a later lord chancellor and author of *Lives of the Lord Chancellors,* that, as an advocate, Erskine was "without an equal in ancient or in modern times."

BIBLIOGRAPHY. L.P. STRYKER, *For the Defense, Thomas Erskine* (1947), is a modern biography written in a popular style by a well-known American courtroom advocate. LORD CAMPBELL, *Lives of the Lord Chancellors . . .*, vol. 6 (1845), presents an interesting and anecdotal account of Erskine's life, marred by some inaccuracies. On the historical background out of which Erskine's famous cases arose, see PHILIP ANTHONY BROWN, *The French Revolution in English History* (1918); and CARL B. CONE, *The English Jacobins: Reformers in Late Eighteenth Century England* (1968).

(F.A.A.)

Eschatology

Eschatology (the doctrine about last things) is originally a Western term, referring to Jewish and Christian beliefs about the end of history (or of the world in its present state), the resurrection of the dead, the Last Judgment, and related matters. The term has been extended by historians of religions to cover similar themes and concepts in the religions of primitive peoples, ancient Mediterranean and Near Eastern cultures, and Eastern civilizations.

Eschatological ideas and beliefs played a central role in the development of Judaism; the Kingship of God, "the end of days," "the world to come," the Messiah and the messianic era, the Day of Judgment, and the images of a perfected future were basic concerns in biblical (Old Testament) and rabbinic Judaism.

NATURE AND SIGNIFICANCE

In New Testament Christianity, history is viewed throughout in eschatological terms: the future of God has already begun with the appearance of Christ; the end of history is near; the end time is therefore filled with danger and salvation, faith and unfaith, Christ and Antichrist, and will be consummated through the resurrection of the dead, the judgment of the world, and its salvation through a new creation. Christianity, with this biblical heritage, has influenced many religions, revolutions, and

civilizations through its orientation toward hope in the future. Biblical eschatological archetypes can be found in the various secular liberation movements leading up to the present day.

In the general history of religions, the term eschatology refers to *conceptions of the beyond* that express the destiny of man after his death (immortality of the soul, rebirth, resurrection, migration of the soul). These eschatological concepts stand in a mutual relationship with the experiences of men in the present world, the turning points of life, and the understanding of death. They often pose a contrast to the present experience of suffering within nature and society, within this whole "perverted world." Eschatological themes thrive particularly in crisis situations, whether they serve as consolations for those who hope for a better world or as the motives for revolutionary transformation of the world.

A distinction has often been made between individual (personal) eschatology and collective (social) eschatology. The eschatological expectations of either type, however, are as extensive as the respective believer's interest and involvement in life and his suffering from its miseries. These expectations can embrace individual souls; they can just as easily embrace a people or group, humanity, or the whole cosmos.

ESCHATOLOGICAL THEMES AND FORMS

Origins and last things. Because biblical eschatology is grounded in what are interpreted as uniquely occurring historical events (such as the Exodus of the Hebrews from Egypt in the 13th century BC), certain difficulties occur when the biblical concept of eschatology is translated into the eschatological framework of other religions. In other religions (especially Eastern and primitive religions), cosmic representations of the eternal struggle between cosmos and chaos prevail. Therefore, a distinction must clearly be made between *mythical* eschatologies and *historical* eschatologies. The term mythical refers to the concept of the philosophically conceived truth of the human condition in relation to the realm of the sacred and the profane as defined in nontemporal terms and stories. The term historical, on the other hand, refers to the concept of the philosophically conceived truth of the human condition in relation to the realm of the sacred and the profane as defined in temporal terms and stories.

Mythical eschatologies

Mythical eschatologies emphasize the reproduction of the origin of the world at the end of the world. The end time repeats the primordial time; out of chaos (disorder) there arises anew the cosmos (order). At the origin there stands the cosmogony (creation of order) and the laws that govern the world and the pure order of things. Time is experienced as decay, degeneration, and guilt. Salvation, therefore, is found in a return of the origin or a return to the origin. All historical events are interpreted as representations of an eternal struggle in which the world order is defended against chaos. This struggle occurs as myth in the world of the gods and as history in the world of men. History thus becomes a cultic drama in which priests and kings play out the ritual roles foreordained to them.

The religious symbolism of mythical eschatology can be defined in terms of the "myth of the eternal return." This concept contains not only a cyclical view of history but also a cultic view of the annihilation of the horrors of history itself. In the ever-recurring cultic festivals, the lost time of history is regenerated and eternity is represented. Through the ritualistic repetition of all events in the creation of the cosmos, the impression of transience is proved to be a mere semblance. Everything basically remains in its place. In the framework of the myth of the eternal return, hope is inherent in memory. This basic structure is not limited to the great cultic religions; many messianic and revolutionary movements (such as the nativistic religious movements in Africa and Oceania, and sectarian movements in pre-1917 Russia) also exemplify it. Within a history that is generally regarded as evil, the saving future is depicted as a return to the primordial origin. In the terms of mythical eschatology, the meaning of history is found in a celebration of the eternity of the cosmos and the repeatability of the origin of the world.

Historical eschatology

Historical eschatology, on the contrary, is not grounded in a mythical primal happening but in historically datable events that are perceived as root experiences (past events that are narrative and paradigmatic in the present) and are regarded as fundamental for the progress of history. Biblical and biblically influenced eschatologies are grounded in such a view of historical experiences and are directed toward the historical future inaugurated by them. In this view, such experiences are never universal but always particular; they are not grounded in natural happenings but in historical election. Such events and experiences are not repeatable in cultic forms and rituals but are remembered through the telling of history and the relating of tradition. They do not abolish history; they rather inaugurate a new process (or new age) of history. They are events in which a *novum* (new or extraordinary thing) is perceived, that have a greater future than a beginning. Hope is thus grounded in historical remembrance but goes beyond what is remembered historically.

Because history is viewed as unrepeatable, the future of history is final. History is understood in this context not as the arena of the horrors of chaos but as the field of danger and salvation. The meaning of history is thus not found in its cultic abolishment (as in the case of mythical eschatologies) through the presence of eternity but in its future goal and its fulfillment. The divine or sacred is not experienced in the eternally recurring orders of nature and of the cosmos through ritualistic reenactments. Rather, God's freedom, faithfulness, and promises of the future are known and comprehended in the contingent and irreversible events of history. If here the future is greater than the beginning, hope at the end is more extensive than at the beginning.

Historical eschatologies are found in the faith of Israel and Judaism, which is grounded in the Exodus event (the liberation of the Hebrews from Egypt in the 13th century BC) and which in the course of its historical experiences is more and more directed towards the expected revelation of the glory of God in all lands. Historical eschatologies are also found in the Christian faith, which is grounded in the history of Jesus and in the root experience of his resurrection from the dead. The hope of Christian faith is aimed at the Kingdom of Christ and the Kingdom of God, through which history is ended as well as fulfilled. In both cases the unique occurrence of a historical foundation event serves as a basis for the final goal of the long-awaited and hoped for future. A historically experienced *novum* opens up hope in a new creation that will be more than the reproduction of the primordial condition.

Apocalypticism and messianism. In the sphere of historical eschatology, distinctions should be made between the hopes of messianism, millenarianism, and apocalypticism. Messianic hopes are directed toward a king of the end time who will lead the people of God, now suffering and oppressed, into a better historical future. Millenarian hope is directed toward the 1,000-year Kingdom of Christ and of his own people, in which the ones who are suffering now will rule over their enemies. In political and nativistic messianism, visions of the vengeance and of the equalizing justice on the side of the oppressed are aimed at political and religious leaders. Always at work in these instances are inner and often local historical expectations of a certain fulfillment of history before the end of history. Apocalypticism should thus be distinguished from this point of view. Apocalypticism upholds the view that God will intervene in history on the side of a faithful minority and that the intervention will be accompanied by sudden, cataclysmic events. According to this view, "this world" cannot bear the "justice of God." Therefore, against what is perceived as the perverted world, the followers of apocalyptic views hope for the creation of a new world on the basis of God's righteousness. If this hope is universal, it is, nevertheless, not generally represented by a people

but rather by individual holy men or, perhaps, by an ascetic community.

The negation of the negative and the analogy of the future

Eschatological terminology. Eschatological language ordinarily uses two elements of style in conjunction with each other: the negation of the negative and the analogy of the future. Objective statements about the future that is not yet present are not possible in history. Such statements are possible only in the form of the negation of the negative in this life and this world. An example of this style may be found in Revelation 21:4: "And death shall be no more, neither shall there be mourning nor crying nor pain any more." Thus, the positive aspects of the eschatological future are circumscribed by the negative aspects of the present. If this future is to be meaningfully related to this life, however, this life, despite all its negativity, must also be presently capable of pointing toward or of foreshadowing the future life. Eschatological imagery and language, therefore, constantly use comparisons or analogies from everyday life (such as the several "the Kingdom of God is like . . ." analogies in the New Testament) and employ the hopes and anticipations of people and events from history in analogical ways.

A problem of eschatological language exists because of the necessary connection between negation and analogy. If negation is predominant, the tendency is towards apocalyptical and metaphysical dualism and towards mysticism. If analogy or foreshadowing of the future dominates eschatological language, the tendency is toward a one-sided belief in progress. In both cases the *novum* of eschatology becomes inexpressible. A hermeneutic (methodological interpretive principle) of eschatological traditions must verify the negation of the negative in face of the presently experienced negative, and simultaneously seek the traces of the coming positive in the ambiguous history of liberations. There are always negative and positive signs of the future in history. "Where danger is, the saving also flourishes" (Hölderlin), but "where the saving draws near, danger also grows" (E. Bloch). Eschatology understands history as a growing crisis: the good provokes the evil, and the growing danger makes the action of redemption necessary. Authentic eschatology is neither world denying nor faith in progress but rather it can be seen as anticipation of freedom in the midst of slavery and of salvation in the midst of lostness and alienation.

VARIETIES OF ESCHATOLOGY IN THE RELIGIONS OF THE WORLD

The end of the order of the world

Primitive religions and nativistic movements. Eschatological motives must be understood in their religious and cultural contexts. According to the eschatological views of the people of the Andaman Islands, at the command of the god Puluga an earthquake will destroy the earth and the bridge of heaven; the souls and spirits of the dead will arise and be reunited. Then men will lead happy lives in power, without sickness, death, and marriage. Even the animals will appear again in their present form. The impatient spirits of the underworld are already now shaking the roots of the palm tree, which supports the earth, in order to bring about the end of this present world and the resurrection more quickly. The Semang pygmies in Malacca hold a somewhat different eschatological view:

> In the beginning of the ages there was still absolutely nothing. Then Ja Pudeu [the highest being] blew with her mouth causing storms to rage over the earth. This was the means by which stars, water, trees, and everything came into being. At the end of the ages when all men shall have died, she will destroy everything with the same storms; a great flood will complete the destruction and the bones of men will swim together. Finally, the bones will rise.

The Australian aborigines claim that the end of the world will come when the moral world order legislated by the gods is no longer upright.

The Gabonese Pygmies in West Equatorial Africa believe that once Kmvum (the original man) lived with them faithfully. Then their guilt brought on the day of separation. He will come again, however, and bring back with him joy, abundance, and happiness. Among the Altaic Tatars of Central Asia there is an eschatological belief that Tengere Kaira Khān (the "graceful emperor of heaven"), who once lived on earth with men, will return at the end of the world in order to judge men according to their works. At his departure from earth he sent a mediator who remains faithful to him. At the end of the world the mediator will be victorious over evil. The Salish Indian tribes of the Pacific Northwest of North America believe that the creator god, before he vanished from the earth, promised the "elder" or "chief" his return at the end. When earth (a female figure) has become old, the coyote will return as the first sign of the world's end. This will be followed by the "chief" himself returning to earth. "After this there will no longer be a land of the spirits. All men will live together, the earth will receive her natural form and will live as a mother among her children. Then all things will be made right and happiness will reign."

Such mythical themes arranged around the origin–fall–return motif are experiencing a revival as primitive peoples suffer cultural shock in their encounters with Western civilization and Christianity. These themes are particularly reflected in various nativistic movements. In part, they take over biblical-messianic hopes and turn against the church. Cultural catastrophes are viewed as signs of the nearing end. The end is expected to bring about the return of the old gods and the eternal orders of creation. The presently "perverted world" is turned upside down. "The blacks become white and the whites become black. The whites will work for us just as we must now work for them." The Mamaia prophets on Tahiti in 1832 prayed for the second coming of Jesus so that he would drive all whites and missionaries out of the land. In the case of the Dayaks on Borneo there is a similar expectation. The Dayaks believe that the white officials and missionaries, when their time has run its course, will become slaves of the free Dayaks. Nativistic and messianic movements are a regular occurrence in acculturations. The self-consciousness of the oppressed is restored in them, and they can be easily politicized. In the 20th century, therefore, large numbers of nativistic–messianic cults have arisen and are continuing to do so in Africa and Asia.

Eschatological views within nativistic movements

Religions of the East. Buddhism has four "noble truths"—the fact of human suffering, the understanding of the origin of suffering, the removal of the causes of suffering, and the path to the transcendence of suffering —and is partly therefore a religion of redemption. According to the Buddhist world view there is a macrocosm composed of innumerable worlds. In recompense for good and evil deeds, creatures are reborn in an unceasing process in the region that they deserve. Beyond all worlds is found Nirvāṇa ("bliss" or enlightenment), "the indescribable goal," whose attainment means redemption from the cycles of existence. Each one of the innumerable worlds passes through periods of destruction and recreation. There is no soul migration because there is no soul substance. Rather, each new existence is defined by *karman*, the deeds of the earlier existence. Only insofar as this is true can one speak of continuity in the reincarnations.

Buddhist and Hindu "eschatological" views

If anything at all in Buddhism can be considered as eschatological it is the yearning for an ultimate redemption from the cycle of rebirth and the final abolishment of the suffering bound up with it. In the religious community founded by the Buddha this yearning finds its peculiar way, the "eightfold path," on whose highest step occurs illumination or enlightenment. This illumination or "life beyond grief and woe" (Nirvāṇa) is a condition of eternal peace that cannot be conceived or described in ordinary terms. Rather, it comprehends even the worlds and their cycles themselves: "I do not know the end of suffering if we have not reached the end of the world" (Gautama, the Buddha).

According to Hīnayāna (the Lesser Vehicle, or Southern) Buddhism, the individual believer must strive after redemption for himself through the exertion of his powers of being. At the centre of Mahāyāna (the Greater

Vehicle, or Northern) Buddhism, on the contrary, stands the concept of the redemption of all living creatures which will occur through the sympathetic assistance of a redeemer figure.

In Hinduism the world of Brahman (the Universal Soul) is created by demiurges (creator beings) in egg-form and contains numerous zones and levels around the golden world mountain Meru. It passes through a series of temporal cycles, and every temporal cycle ends with the destruction of the world followed by a new creation under a new demiurge. Man is now living in the dawn of the last, worst Kali *yuga* (age) of such a cycle. Through migration of the soul, man is drawn into the circle of the animals and plants. Rank and kind of birth are determined by the individual *karman* (acts and their consequences). The *karman*, which accompanies the soul and even has an effect on the destruction of the world, determines the following existence. After death, the soul returns to earth or it goes the "gods' way" (*devayāna*) of redemption. Redemption occurs when *karman* can no longer be produced, or it can happen through a divine act of grace that blots out the existing *karman*. Redemption is popularly viewed as entrance into the highest heaven of the god worshipped, where the redeemed one awaits a spiritual reflection of earthly joy. In modern Hinduism the soul that is identical with God is redeemed through a recognition of the organic wholeness that has vanished from consciousness because of the soul's imprisonment in matter. Self-recognition (*Ātman*) then leads to identity with the absolute being (Brahman). Redemption lies in the accomplishment, or rather recognition, of the *Ātman*-Brahman identity, for it already frees man from the chains of *karman* and *saṃsāra* (cycle of rebirths).

Buddhism and Hinduism do not have a historical eschatology. They rather emphasize an ultimate redemption from the cycles that have no beginning. Their redeemer figures are impersonal transparencies of the ultimate-universal.

Religions of ancient civilizations. In the religions of the West, there is a tendency to understand time as irreversible and therefore the end as occurring once and for all, as ultimate. The final judgment will be followed by the creation of a new and sacred world that is eternal. Origins of this kind of eschatological thinking are found in ancient Egypt in texts such as the "Shipwrecked Sailor" and the "Conversation between Atum and Osiris." The idea of an eschatological individual judgment of the dead is developed here in the strongest sense. Ancient Greek and Roman eschatological views depict a shadow life for the individual departed soul in Hades. Other than this focus on the individual, however, the cyclical concepts of periodic world destruction and world renewal are also found in these religions.

Zoroastrianism and Islām

Religions of the West. In Parsiism (Indian Zoroastrianism), the end-time struggle between the forces of the good against the forces of evil is the ruling element in its eschatology, and Parsis look toward a "wonderful creation" (*frashokereti*) at the end. Under biblical influence, the concept of the resurrection of men is given rich embellishment in Islām. The Antichrist and ʿĪsā (Jesus) play important roles. ʿĪsā will usher in Islām as the world religion. Before him, however, will appear a mahdī (divinely guided one) who is to establish the Islāmic ideal state and fulfill the divine law. After the appearance of ʿĪsā, the last judgment will begin: the good will enter paradise; the evil will fall into hell. Heaven and hell possess various goals and steps of recompense for good and evil. The time before the end is viewed pessimistically: God himself will abandon the godless world. Ka'bah (the great pilgrimage sanctuary of the Muslim world) will vanish, the copies of the Qur'ān will become empty paper, and its words will disappear from memory. Then the end will draw near.

Judaism. The real inception of historical eschatologies stems from the Old Testament. Israel's faith in God there is rooted in the historical experience of the Exodus, and because of this experience of liberation it contains a hope in the guidance and the promises of God in history. The basic structure of this faith is found in the law of promise and fulfillment. The eschatology of the Old Testament is grounded in the identity of faith in God and hope in the future (Gen. 12:1ff.). It has its beginning in the promise of a "good and broad land, a land flowing with milk and honey" (Ex. 3:8). In the Pentateuch (first five books of the Old Testament) the promise is broadened to the increase of people and possessions, the blessing of God and the victorious presence of God (Gen. 49:8–12; Num. 23; Deut. 33:13–17; Num. 23:21). The history of the occupation of the land of Canaan (Palestine) and the victory of the Hebrews is to be understood as "realistic hope." Through the experiences of Israel's own disobedience to the laws and the will of God and defeats at the hands of its enemies, the concept of the "day of the Lord," which is to bring salvation and victory, came into existence. The happiness of the establishment of the Kingdom by David in the 10th century BC led to a hope in the future Messiah (Anointed One) of God from the house of David (II Sam. 7).

Concept of the "day of the Lord"

In the midst of the political catastrophes of the 8th century BC, the great prophets took up the concept of the "day of the Lord" and proclaimed it as a day of judgment (Amos 5:18) over the disobedient people and also over all other peoples. It was through such a process that the day of the Lord concept became the bearer of eschatological hopes. Isaiah also viewed salvation in an eschatological light as happening only after the universal judgment (Isa. 4:3; 6:13; 11:11; 37:31) and combined it with the presence of a messianic mediator of salvation (7–12).

In face of the political destruction of Israel (8th century BC) and Judah (6th century BC) the prophetic hope kept Israel alive religiously. This prophetic hope was aimed primarily at a comprehensive and total new creation, a new heart, a new covenant (Jer. 31; Ezek. 36; Isa. 41; Isa. 51). Through the vicarious atonement of the servant of God (either the people of Israel or a messianic figure) this hope was to include not only Israel but also the Gentile world (Isa. 42:6; 49:6). The future of Israel is thus bound up with the future of all peoples and of the whole earth.

In the incipient apocalyptic views of the prophet Daniel (2 and 7), hope is transcendent. His apocalyptic eschatological hope expects the "Kingdom of the Son of man" following the consummation of evil in the fourth and final kingdom of the world. Since that time, hope in a Messiah and hope in a Son of man have been bound to a kind of eschatology that unites the fulfillment of the history of Israel with the end of world history.

Individual eschatology emerges only on the periphery of the Old Testament. Amazingly, there were in Israel no known death cults and no vivid conceptions of life after death. The late expectation of the resurrection from the dead to judgment (Dan. 12:2) is not a yearning for salvation but hope in the victorious righteousness of God. Rabbinical messianism continued this same line of thought.

Christianity. If for Judaism the peculiar eschatological event lies in the future, this future of God, according to the New Testament, has already begun with Christ ("Today this scripture has been fulfilled in your hearing." [Luke 4:21] "The Kingdom of God is at hand." [Mark 1:15]). New Testament eschatology differentiates between the eschatological present and the eschatological future. Yet they are both one unique event: the presence of Christ is the anticipation of the future of God, and the future of God is the fulfillment of the presence of Christ. God has reconciled the world in Christ but not yet redeemed it. Faith saves, but the believers are saved in hope (Rom. 8:24). At the centre stands the hope in the "Kingdom of God," in which God comes to his lordship and his right in the world.

Eschatological present and eschatological future

Individual eschatology and collective eschatology are constituent parts of the theocentric (God-centred) hope in the coming God. Jesus proclaimed the Kingdom of God and brought it through his words and deeds as grace and joy to the poor and the lost. He united the Kingdom of God with his person in such a way that he is under-

stood as the incarnated Gospel of the Kingdom. His word calls men to eschatological decisions and makes the end judgment a present concern. His Crucifixion in loneliness and alienation is understood as the presence of salvation among the unsaved and as revelation of the divine righteousness. In God's raising of Jesus from the dead, the Easter witness recognized the exaltation of Jesus as Lord of the end time, and they saw in this action the anticipation of the coming resurrection of men and the new creation of the world.

The memory of Jesus and the Christ-event (cross and Resurrection) is the basis of primitive Christian eschatology in its many variations. Its object is the Parousia (Coming Presence) of the crucified one in the glory of the coming God. The Parousia brings the visible presence and indwelling of God in his new creation. It brings the freedom of the "children of God" and thus the new humanity. It brings the liberation of the whole creation, which "waits with eager longing" (Rom. 8:19) for its release from the power of death to participation in the glory of God and the participation of the new man in peace and freedom.

Eschatology, in primitive Christianity, developed within the process of the interaction between the memory of Christ and new experiences in the world. A primary controversial issue was the *particularity* or *universality* of hope for salvation. The primitive congregation in Jerusalem understood itself as a messianic community, but remained in the sphere of Judaism. Only after the awakening of Israel—according to its beliefs—would the Gentiles be allowed to come into the community of faith and hope. In the Hellenistic congregations (Antioch, Corinth), on the contrary, the Gentiles came to faith without having first become Jews. "Spirit and freedom" were understood as already universally existent.

Early controversial Christian eschatological concerns

A second controversial issue was the *presentness* or *futurity* of salvation. Primitive Christian enthusiasm viewed salvation as already present in the possession of the Spirit. The Gospel of John also emphasizes the present eschatological possession of salvation in faith. The theology of Paul, however, is dominated by the tension between the presence and the future of salvation. Faith gives a ground for hope and prevents it from being superfluous for the reconciled are not yet redeemed; they wait together with the unredeemed creation for the "day of Christ," the freedom from death, and the unhindered participation in the presence of God ("that God may be everything to every one"; I Cor. 15:28). The letters of John, Hebrews, and the Revelation of John also emphasize the future, the universality, and the bodiliness of salvation.

ESCHATOLOGY IN MODERN TIMES

Times of hope and times of fulfillment have alternated with each other in the history of Christianity. During the first three centuries the eschatological views of the martyrs stood in the foreground. But since the time of the Roman emperor Constantine (died 337) the political recognition of Christianity has been understood as a realized hope in the Kingdom of Christ. Futuristic eschatology continued to exist in the suppressed underground sects. The consequent history of state Christianity and of the established churches is interlaced with an abundance of adventistic and millenarianistic sects (those awaiting the coming Kingdom of God). Such sects appeared among the leftwing groups of the Reformation. They were also active in the English Revolution. They were conspicuous in the awakening and revivalistic movements of the "Christian century" (19th century), particularly in so-called "Christian nations." The more established Christian churches excluded eschatology as meaningless or as an insignificant myth. The more these hopes and visions emigrated out of Christianity the more they were invested in revolutionary movements. The history of revolution in Europe, Russia, and North America is dominated by the motives and images of a secularized eschatology. Here eschatological expectation radically transforms itself into revolutionary action, as it sometimes did during the Reformation.

Revolutionary eschatology

Since the exegetical (interpretive) works of Johannes Weiss and Albert Schweitzer around 1900 (school of "consistent eschatology") and dialectic theology (Karl Barth, Rudolf Bultmann), eschatology has again become a principal theme of academic Christian theology. The crises in the Western countries have led to a renewed activization of eschatological hopes. In church struggles, this was expressed in terms of distinctions between Christianity as a state religion and congregations with eschatological orientations. On its margins, Western civilization contains a series of mystical and apocalyptical "anticultures." Initial attempts to combine eschatology and philosophy, hope, and social practice, and thus overcome the difference between the church and the sects, as well as the church and the modern age, are found in Ernst Bloch's philosophy of hope (*Das Prinzip der Hoffnung*, 1959), the writings of P. Teilhard de Chardin, and in the "theology of hope" (J. Moltmann, W. Pannenberg, H. Cox, L. Dewart, etc.). Eschatology is one of the main focuses of Christianity. Therefore, its most important theological decisions about theory and practice must take eschatological concerns seriously, as seriously as the many and varied revolutionary groups (both religious and secular) have done in the 20th century.

BIBLIOGRAPHY

History of religions: J.F.H. GUNKEL, *Schöpfung und Chaos in Urzeit und Endzeit* (1895); W. SCHMIDT, *Der Ursprung der Gottesidee*, 12 vol. (1912–55); J.G. FRAZER, *The Belief in Immortality and the Worship of the Dead*, 3 vol. (1913–24); M. ELIADE, *Le Mythe de l'eternel retour* (1949; Eng. trans., *The Myth of the Eternal Return*, 1954), an internationally recognized standard work; E. SARKISYANZ, *Russland und der Messianismus des Orients* (1955); V. LANTERNARI, *Movimenti Religiosi di Libertà e di Salvezza dei Popoli Oppresi* (1960; Eng. trans., *The Religion of the Oppressed*, 1965); T. ROSZAK, *The Making of a Counter Culture* (1969), a discussion of the formation of secular messianic movements in modern society.

Biblical: H. GRESSMANN, *Der Ursprung der israelitisch-jüdischen Eschatologie* (1905), somewhat antiquated, but still worth reading; R.H. CHARLES, *Eschatology: The Doctrine of a Future Life in Israel, Judaism and Christianity* (1899, reprinted 1963), a comparison of ideas; MARTIN BUBER, *Königtum Gottes* (1932), a standard work of Judaic studies; G. VON RAD, *Theologie des Alten Testaments*, 2nd ed., 2 vol. (1958; Eng. trans., *Old Testament Theology*, 1962–65), the standard work of German Old Testament research; S.O.P. MOWINCKEL, *Han som kommer* (1951; Eng. trans., *He that Cometh*, 1956); H.H. ROWLEY, *The Relevance of Apocalyptic* (1944); C.H. DODD, *The Parables of the Kingdom* (1935), a present-oriented eschatology; O. CULLMANN, *Christus und die Zeit*, 3rd ed. (1962; Eng. trans., *Christ and Time*, 1962); R.K. BULTMANN, *History and Eschatology* (1957), an existentialist interpretation of eschatology; P.S. MINEAR, *The Christian Hope and the Second Coming* (1954), an integral biblical view.

Theological and philosophical: F.H.C. HOLMSTROM, *Das eschatologische Denken der Gegenwart*, (1936) for the period 1900–30; P. ALTHAUS, *Die letzten Dinge*, 6th ed. (1957), a Lutheran representation of eschatology; M. SCHMAUS, *Von den letzten Dingen* (1948), the best Catholic treatise on eschatology; H.E. BRUNNER, *Das Ewige als Zukunft und Gegenwart* (1953; Eng. trans., *Eternal Hope*, 1954), a personalistic eschatology; N. BERDYAEV, *The Beginning and the End* (1957, orig. pub. in Russian, 1947), Russian Orthodox philosophy of religion; J.A.T. ROBINSON, *In the End, God* (1968), an introduction to this subject; P. TEILHARD DE CHARDIN, *Le Phénomene humain* (1956; Eng. trans., *The Phenomenon of Man*, 1959), a discussion combining eschatology with the theory of evolution; E. BLOCH, *Das Prinzip Hoffnung*, 3 vol. (1954–59), a neomarxist philosophy of hope and standard work in Europe; J. MOLTMANN, *Theologie der Hoffnung*, 8th ed. (1964; Eng. trans., *Theology of Hope*, 1967), on the beginning of the ecumenical theology-of-hope movement.

(J.D.M.)

Eskimo-Aleut Languages

Eskimo and Aleut, once neighbour languages on the Alaska Peninsula, are related but quite distinct; together they form the Eskimo-Aleut language family.

Eskimo is spoken in Greenland by a native population of about 40,000, in Arctic Canada by 15,000 or more, in Alaska by some 25,000 native inhabitants, and on the coast of the Chukchi Peninsula in northern Siberia by fewer than 800 of the Siberian Eskimos (1970). In Green-

land, Eskimo is used in the schools, in the church, and on the radio; in Canada and in Alaska the language is employed mainly as a means of religious education. It is the language of instruction in the first grade of primary school in the Soviet Eskimo area. There are numerous Eskimo dialects; between them there is a more or less clear-cut break at the Bering Strait and at Norton Sound in Alaska. Within each of the dialect chains the neighbouring dialects are mutually intelligible, but the cumulative differences impede or prevent understanding between the dialects at the geographical extremes.

Intelligibility of dialects

Aleut, now greatly reduced in number of speakers, is the smallest branch of the family. It is spoken in the Aleutian Islands and in the Pribilof Islands in the Bering Sea, settled in the 1820s, by fewer than 1,000 Aleuts; on the Soviet Commander Islands, settled in 1826, there are fewer than 100 Aleut speakers (1970). Three mutually intelligible Aleut dialects are left, although one is nearly extinct. The Eskimo language is unintelligible to the speakers of Aleut, and vice versa.

Classification. The Eskimo dialects are usually classified into two main branches—an Inuk, or Inupik, branch, which spreads eastward from the Bering Strait and the Norton Sound, and a twofold Yuk, or Yupik, branch, consisting of Asian (Siberian) and Alaskan varieties. This grouping is based on the dialectal forms of the term for people or real people that most Eskimos use to refer to themselves. The exact interrelationship of the dialects remains to be clarified.

The term Aleut, introduced by Russian traders from Kamchatka in northeastern Siberia after 1741, is commonly used both for the Aleuts proper, who call themselves Unangan (or, on the island of Atka, Unangas), and for the Eskimos of former Russian Alaska. The linguistic relationship between Eskimo and Aleut, sensed as early as the 1820s by the Danish linguist Rasmus Rask, could be proved only through a fairly advanced analysis of both languages. The proposed relationship of Eskimo-Aleut with other language families, such as Chukchi-Kamchadal, Uralic, and/or Indo-European, remains speculative.

Relationship between Eskimo and Aleut

Alphabets and orthography. The first book in Eskimo was published in 1742 by Hans Egede, a Dano-Norwegian missionary to Greenland. It was printed in the Roman alphabet, with the letter *r* used to represent the Eskimo uvular fricative (a sound involving friction in the airstream, made at the back of the mouth with the appendage known as the uvula). The Greenlandic orthography was systematized in 1851 by Samuel Kleinschmidt, a German of the Moravian Brethren. He introduced a small capital ʁ to distinguish the uvular stop, made with stoppage of the airstream by contact of the uvula and the back of the tongue (also written *q*), from the velar sound *k*, made with the back of the tongue touching the velum, or soft palate (like the English *k*). Three accents were introduced to differentiate long sounds from short ones; *e.g.*, *anâna* (equivalent to *anaana*) "mother," *mâna* (equivalent to *manna*) "this," *mãna* (equivalent to *maanna*) "now." That is, the circumflex, ^, marks a long vowel, the acute accent, ´, indicates a long consonant following the marked vowel, and the tilde, ~, notes a long vowel and a long succeeding consonant. The Kleinschmidt orthography is still used today, and similar orthographies have been employed by the Moravian Mission in Labrador since 1800.

In Russian Alaska, an adequate Cyrillic alphabet was designed for the Aleut language by the Orthodox missionary Ivan Veniaminov in about 1830. In 1848 it also became used for southern Eskimo. Books in both languages were published by the Russian Orthodox Church in America until 1903.

Adaptation of the Cree syllabary

In 1878 the syllabic characters originally designed for the Cree Indians were introduced to the Eskimos of central Canada by the Rev. Edmund J. Peck of the Church Missionary Society. Since 1953 this form of writing has also been used in an Eskimo periodical sponsored by the Canadian government, alongside a new, tentative orthography in the Roman alphabet.

The ordinary Roman alphabet was used for Alaskan Yupik by Moravian missionaries from Pennsylvania in the 1920s. From 1948 on, Protestant missionaries in northern Alaska developed an alphabet with seven additional letters. In 1961 a program was started at the University of Alaska, with the active participation of Yupik Eskimos, for working out a systematic Eskimo orthography in the Roman alphabet to be used in Alaskan public schools from 1971 on. In the Soviet schools for Asian Eskimos a Roman alphabet with two additional letters was introduced in 1932, but it was replaced by the Russian Cyrillic alphabet in 1937.

Phonological characteristics. Eskimo and Aleut have relatively simple systems of distinctive sounds. The accent (stress) depends upon the length of the syllables and never has independent value as in English.

In Eskimo there are four distinctive vowels. Three of them, usually written as *a, i, u,* vary considerably in pronunciation according to the consonants that follow or precede. These three vowels occur both short and simple, and combined into long vowels and diphthongs. The vowel combinations result primarily from the loss of former intervocalic consonants; *e.g.*, Greenlandic *ûvoq*, the equivalent of *uu-vuq* "is burnt," is related to the Asian Yupik form *ugu-*, which has a fricative *g* sound between the two vowels. The fourth Eskimo vowel is represented in Yupik by a short *e* and is pronounced somewhat like the vowels in "but" or "bird." In the Inupik area it has become identical with the vowel written *i*. The Aleut language has only *a, i,* and *u* in both short and long varieties. The fourth vowel has become identical with one of the first three or has been dropped from the language.

The consonantal system

Of consonants, Eskimo has from 13 to 21, depending on the dialect. The stop sounds include the uvular *q*, the velar *k*, the dental *t*, made with the tip of the tongue touching the back of the upper teeth, and the labial *p*, made with the lips; in Alaskan Yupik there is also a palatal *c* (like English *ch*), to which an *s* corresponds in the other dialects. In parts of Canada this has changed to *h*. The nasal sounds, made with the breath passing through the nose, include ŋ (as in "sing"), *n*, *m*, and, in Asian Yupik, also a voiceless *n*, made without vocal cord vibration. Voiced and voiceless varieties of the continuant consonants *r, g, l,* and *v* are distinctive sounds in the western dialects but in eastern Inupik they are only variants. In addition to *j*, pronounced as *y* in English "year," some dialects have sounds similar to English *r* or *z* or to *sh* (in Greenlandic written *ss*). Corresponding to *ss* in Aleut there is a fricative *d* (pronounced as the *th* in "that"); *e.g.*, Aleut *da-* "eye" is related to Eskimo *izi* and *ii* and to Greenlandic *isse*. Aleut shares with Eskimo most of the consonants articulated with the tongue, including *c* (the English *ch* sound) and *s*, but has *p* and labial fricatives (*f* and *v*) only in loanwords from Russian or English. Aleut *m* corresponds with both Eskimo *m* and *v;* to Eskimo *p* corresponds the Aleut *h* (in initial position) and the Aleut aspirated nasal sound *hm* (in noninitial positions)—*e.g.*, Aleut *hum-* "to swell" corresponds to Eskimo *puvi-*; Aleut *ahmat-* "to ask" is cognate with Yupik *ap(e)t-*. (An aspirated sound is one pronounced with an accompanying puff of air.)

In initial position, Inupik uses only a single stopped consonant (*i.e., p, t, k,* or *q*), or *s, n,* or *m* (rarely *l*); between vowels it employs at most two consonants, including double ones. In contrast, Yupik and Aleut have initial consonant clusters, resulting from the loss of a vowel in the first syllable from an older historical form; *e.g.*, Yupik and Aleut *sla* "weather," Inupik *sila.*

Grammatical characteristics. Eskimo has a great number of suffixes but practically no prefixes or compounds. In Aleut the word forms are simpler, but the syntactic constructions (the way the words are arranged in sentences) are more complicated. Suffixes are often accompanied by changes in the stem, especially in Eskimo; *e.g.*, *nanuq* "polar bear," dual *nannuk* "two polar bears," plural *nannut* "several polar bears"; *inuk* "person," dual *innuk,* plural *inuit; umialik* "owner of boat (*umiaq*), chief," dual *umiallak,* plural *umialgit* (North Alaska dialect). Note the doubled consonants in the dual and plural

forms of some terms, and the other consonant and vowel alternations.

Grammatical numbers—singular, dual, plural—combine with suffixes for person—*e.g.*, *ulu-ga* "my knife," and *ulu-t-ka* "my knives," in which *-t-* means "several" and *-ga* or *-ka* "my." The possessor of someone or something occurs in the so-called relative, or subordinative, case—*e.g.*, *umialgum pania* "the chief's (his) daughter," in which *pani-a* means "his or her daughter"; this is distinguished from the reflexive *panni* "his or her own daughter" and the stem *panik*.

Personal pronouns

The personal pronouns consist of a stem and person suffixes—*e.g.*, Aleut *ti-*ŋ "I, me," with the ending -ŋ as in *ada-*ŋ "my father"; Eskimo *uva-*ŋa "I, me," with a demonstrative stem (*uva*) as in *uva-ni* "here." Corresponding to the separate personal pronoun forms in Aleut are the Eskimo suffixes for the 1st, 2nd, and reflexive 3rd persons (himself, etc.). These indicate the subject of an intransitive verb or the object of a transitive verb—*e.g.*, *-*ŋ*a* means "I" in *uqaqtu-*ŋ*a* "I said," and "me" in *uqautigaa-*ŋ*a* "he told me." A noun that would be the subject of "said" (the intransitive verb) or the object of "told" (the transitive verb) is in the absolute case (it is not marked with a suffix), whereas the subject of a transitive verb (*e.g.*, "tell") occurs in the subordinative case. For example, compare the unmarked form for "chief," *umialik*, in *umialik uqaqtuq* "the chief said" with its corresponding form *umialgum* in the subordinative case in *nukatpiarzuk umialgum uqautiga-a* "the young man (object), the chief told him" (*-a* "he - him"). With a 3rd person object ("him, them"), a transitive verb has largely the same suffixes as a noun—*e.g.*, the *-t-* and *-ka* in *ulu-t-ka* "my knives" and *tigu-gi-t-ka* "I took them" (*-gi-* is indicative) and the corresponding Aleut forms, *-ni-* and -ŋ in *ukina-ni-*ŋ and *su-ku-ni-*ŋ.

Eskimo nouns and pronouns have six adverbial cases, expressing relations such as "in," "to," from," "along," "with," and "like"—*e.g.*, *iglu-mi* "in the house," *iglu-pti*ŋ*-ni*, "in our house." In the Yupik dialects there are only five such adverbial cases. These are reduced to two (in/to, from/along) in Aleut and are limited to pronouns and relation words—*e.g.*, *ula-m nag-a-n* "of house (*ula-m*) in its (*-a-*) interior"; this corresponds to Eskimo *iglu-m ilu-a-ni* "in the house."

Verbal modes include indicative ("he goes"), interrogative ("did he go?"), imperative ("go!"), optative ("may he go"), participles ("going, gone"), and other forms corresponding to subordinate clauses in English—*i.e.*, clauses beginning with "if," "when," etc. Other modal relations and tenses are specified by derivational suffixes, and in Aleut also by auxiliary verbs; *e.g.*, *haqa-l sara-nar* "coming he slept" is equivalent to "he came yesterday." An Eskimo derivative form may also correspond to an English complex sentence—*e.g.*, $tikit_1$-$qaar_2$-$mina_3$-it_4-ni_5-ga_6-a_{7-8} is "he $(A)_8$ $said_{5-6}$ that he $(B)_7$ would not_4 be $able_3$ to $arrive_1$ $first_2$," or, in exact Eskimo order, "to arrive first be able would not said him he."

Vocabulary. A remarkable feature of the vocabulary is the great number of demonstratives, about 30 in western Eskimo and in Aleut. For example, in Aleut there is *hakan* "that one high up there" (as a bird in the air), *qakun* "that one in there" (as in another room), and *uman* "this one unseen" (heard, smelled, felt).

The vocabulary naturally has its local particularities, the various groups having lived under very different conditions. The Eskimo word that means "meat" from Greenland to Siberia means "fish" south of Norton Sound and also in the Aleut language. Word taboo has also played its part, as in East Greenlandic, in which the general Eskimo-Aleut word for "eye" (*da-*, *izi*, see above) has been replaced by *uitsatai* (*ui-sa*+*uta-i*) "those by which he keeps gazing."

Word formation by derivation

Eskimo-Aleut derived words (*e.g.*, similar to English "winter-ize" or "anti-dis-establish-ment-ari-an-ism") correspond quite often to simple, nonderived English words. The possibility of derivation is virtually unlimited in the languages, and the number of word stems is comparatively small; *e.g.*, there are fewer than 2,000 in well-known West Greenlandic. Examples of derivatives are: *nalu-voq* "is ignorant," *nalu-vâ* "does not know it," and *nalu-na-er-* "make not (*-er-*) to be (*-na-*, *-nar-*) ignored," which is equivalent to "communicate." *Nalunaer-asuar-ta-ut* "that by which (*-ut*) one communicates habitually (*-ta-*, *-tar-*) in a hurry (*-asuar-*)" is the form for "telegraph," a term coined in the 1880s.

In Greenlandic there are four loanwords from medieval Norse; from the colonial period after 1721 there have been surprisingly few borrowings until recently. In Aleut and in the Eskimo of former Russian Alaska there are many borrowings from Russian, and there are several in Asian Eskimo from English, and many from Chukchi, a Paleo-Siberian language. Notable Eskimo contributions to the vocabulary of English and other European languages are "igloo" (or *iglu*) and "kayak" (*qayaq*).

BIBLIOGRAPHY. For bibliography, demographic data, and literacy programs, see M.E. KRAUSS, "Eskimo-Aleut," in T.A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 10 (in prep.). Supporting evidence is given in: L.L. HAMMERICH, "The Western Eskimo Dialects," *Proc. 32nd Int. Congr. Americanists, Copenhagen 1956*, pp. 632–639 (1958); KNUT BERGSLAND, "The Eskimo Shibboleth *Inuk/Yuk*," in *To Honor Roman Jakobson*, pp. 203–221 (1967), and "The Eskimo-Uralic Hypothesis,"*Journal de la Société Finno-Ougrienne*, vol. 61 (1959). S.P. KLEINSCHMIDT, *Grammatik der grönländischen Sprache* (1851; restated by MORRIS SWADESH in HARRY HOIJER *et al.*, *Linguistic Structures of Native America*, pp. 30–54, 1946), is a classic in the field. See also C.W. SCHULTZ-LORENTZEN, *Dictionary of the West Greenland Eskimo Language* (1927).

(K.B.)

Espírito Santo

The State of Espírito Santo (Estado do Espírito Santo), a small coastal state of eastern Brazil, is bounded on the north by the State of Bahia, on the east by the Atlantic Ocean, on the south by the State of Rio de Janeiro, and on the north and west by the State of Minas Gerais. It has a total area of 17,605 square miles (45,597 square kilometres), including the uninhabited offshore islands of Trindade and Martin Vaz; the state is about 250 miles from north to south and about 75 miles from east to west. At the start of the 1970s the population was over 1,600,000, almost double the figure of 860,000 recorded in 1950. Vitória, the port and capital city, stands on an island in the Baía do Espírito Santo (Bay of Espírito Santo), the only bay on the coast.

History. Espírito Santo formed part of one of the captaincies given by the Portuguese crown to the explorer Vasco Fernandes Coutinho, who, on May 23, 1535, first landed on the future site of Vitória. In 1799 Espírito Santo became an independent captaincy, in 1824 a province of the Brazilian Empire, and in 1889 a state of the newly formed federal republic.

Physical geography. The relief is characterized by the low mountain ranges of the Serra dos Aimorés on the western border and by isolated groups of hills on the eastern coastal plains. The most important rivers—the Doce, São Mateus, and Itapemirim—flow eastward from Minas Gerais across the state to the ocean; navigation on these rivers is hampered by their irregular rate of flow, as well as by falls, rapids, and sandbars.

The country is today sparsely forested, as most of the forests that formerly supplied Brazil with a substantial proportion of its exports of rosewood and other cabinet woods have now been destroyed.

Soils and climate

With the exception of the sandy plains and swamps of the coastal strip, the soil is generally fertile. Repeated plantings of the same crops have, however, exhausted some soils, and many fields have been converted to pasture land.

The climate of the coastal zone and the valleys is hot and humid. In the highlands the temperature is lower, and the climate more comfortable. Rainfall, which averages about 50 inches a year, is heavier from October to March.

Population. The population is predominantly white, but Negroes, mulattoes, and caboclos (of mixed European and Indian ancestry) are also found in high proportions. The population is unevenly distributed.

The largest and most populous cities and towns are Vitória (122,000), Itaquari (65,000), Cachoeiro de Itapemirim (59,000), Colatina (46,000), and Vila Velha (43,000).

The overwhelming majority of the population are Roman Catholic, although Protestants of all denominations and Spiritualists are also represented. Portuguese, the mother tongue of almost the entire population, is generally spoken.

Administration and social conditions. Legislative power is held by the state assembly, while the executive power is vested in the governor. The state is divided into 36 judicial districts and 53 municipalities.

Campaigns against infectious diseases have been successfully pursued. Yellow fever and smallpox have been eliminated, and malaria has been virtually eradicated. The inroads of tuberculosis have also been greatly reduced. Between 1950 and the early 1970s, the death rate dropped sharply.

Primary education is administered by the state but is largely subsidized by the federal government. The adult literacy rate is about 72 percent. In the early 1970s the state embarked on an extensive school-building program, as well as on a vigorous literacy campaign. Secondary schools are distributed throughout the state; there are two medical schools in Vitória (one privately run, the other affiliated with the Federal University of Rio de Janeiro).

Agriculture remains the leading economic activity; coffee, cocoa, sugarcane, rice, and corn (maize) are the principal products. Cattle raising is important, with beef cattle generally being raised in the north and dairy cattle in the south. Poultry farming is becoming increasingly important, and industry, especially food processing, is making progress.

Transportation

Espírito Santo has two railways—one connecting Vitória with Rio de Janeiro to the southwest and one that moves iron ore from the mines of Minas Gerais for export through the port of Tubarão near Vitória. The state is crisscrossed by an extensive road network, and Vitória airport is capable of handling small jet aircraft.

There are 11 radio stations and one television station in the state; Vitória has several cultural institutions. Local festivals usually have a religious meaning, and local folklore is deeply rooted in Portuguese and African traditions.

BIBLIOGRAPHY. FERNANDO DE AZEVEDO, *A Cultura Brasileira*, 2nd ed. (1944; Eng. trans., *Brazilian Culture*, 1950), is by far the most valuable work for understanding the manifold aspects of Brazilian culture. There are three important books for an adequate study of the history of Brazil: PEDRO CALMON, *História da Civilização Brasileira* (1933); J.F. DA ROCHA POMBO, *História do Brazil*, 10 vol. (1905); F.A. DE VARNHAGEN, *História Geral do Brazil*, 2nd ed., 2 vol. (1877). JOSE TEIXEIRA DE OLIVEIRA, *História do Estado do Espírito Santo* (1951), is the best historical work on Espírito Santo; while MARIA STELLA DE NOVAES, *História do Espírito Santo* (1970), is a highly informative study. AUGUSTE DE SAINT-HILAIRE, *Segunda Viagem ao Interior do Brasil, Espírito Santo* (1936), is a narrative of travels in the state of Espírito Santo by the French naturalist.

(D.N.daC.)

Este, House of

The House of Este was an Italian princely family that ruled in Ferrara from the 13th to the end of the 16th century and in Modena and Reggio from the later Middle Ages to the end of the 18th century. It played a significant role in Italian history, and its influence on the political and military events of the peninsula is second only to that of the House of Savoy, which produced the Italian kings of the 19th and 20th centuries.

The Estensi were of ancient Lombard descent. They had first settled in Lombardy as a branch of the great 10th-century dynasty of the Obertenghi, which held power and wealth in Lunigiana, Genoa, and Milan and which also gave rise to the feudal houses of the Malaspina, the Pallavicini, and the margraves of Massa and Parodi. Subsequently, after various vicissitudes, the members of this dynasty removed to the lands of the Venetians, where they had estates at Este, Monselice, Rovigo, and Friuli. The Estensi took their name from the township and castle of Este, 17 miles (27 kilometres) southwest of Padua, and the true founder of the family was the margrave Alberto Azzo II (died 1097). From his son Welf IV, duke of Bavaria, there began a related branch that gave origin to the dukes of Bavaria, Brunswick, and Lüneburg, as well as the electors of Hanover. Another son, Ugo, tried without success to establish in France, while a third son, Folco I (died *c.* 1136), became second in line in the House of Este. Neither he nor his successor, Obizzo I (died 1193), however, achieved any great distinction, beyond the offices and titles that fell naturally to the upper feudal families; but it was during the lifetime of Obizzo I that the Estensi first acquired political importance in Ferrara, through his marriage to a member of one of the two great and rival families of Ferrara. His successor, Azzo VI, acquired considerable authority in the city, though his premature death in 1212 left the family temporarily weakened. Not until 1240 did a descendant, Azzo VII, return to power in the city, in alliance with the Guelf league formed by Pope Gregory IX. This marked the true beginning of Este rule in Ferrara.

Founding of the House of Este

In 1264 Azzo's heir, Obizzo II (1264–93), was created perpetual lord by the people of Ferrara under the pressure of Guelf strength. The Pope, lawful lord of the Ferrarese territory, at first did not oppose this action, but afterward began to contest the Estensi government. Obizzo II's power was growing, however, and he had himself chosen lord of Modena in 1288 and of Reggio in 1289. In the 14th century the House of Este went through difficult, stormy periods, not only because of its controversies with the Pope but also because of domestic dissensions, sometimes very hazardous. The house succeeded, nevertheless, in strengthening its position, and, under Nicolò II (1361–88), called the Lame, there was built the famous Este Castle, the work of the architect Bartolino da Novara, which became a symbol of the power of the duchy and a sure defense against external dangers. To the brother and successor of Nicolò II, Alberto V (1388–93), is due the erection of the University of Ferrara, destined for lasting fame; it was obtained by Pope Boniface IX as a concession in 1391.

Involvement in Italian politics

The reign of Nicolò III (1393–1441), son of Alberto, marked the strengthening of Estensi domination in Ferrara and the introduction of Estensi influence generally in Italian politics. After having defeated an attempt by the Paduans to achieve hegemony in Ferrara, the Estensi duke became intermediary in the political and military contests in the Italian states and extended his dominions. Personally, Nicolò was known for his sensuality; a Ferrarese saying runs, "On both sides of the River Po they all are Nicolò's sons." He had his son Ugo and his young second wife Parisina Malatesta beheaded because they were found guilty of adultery together. But he devoted himself to the exterior manifestations of the religious faith—going on a pilgrimage to the Holy Sepulchre and to Vienna's Saint Anthony, and playing host to the ecumenical council in 1438 that represented a fruitless attempt to bring together again the Western and Eastern churches. (This council was afterward transferred to Florence.) He even seems to have come close to obtaining the succession of an Estensi heir to the Milanese states, but he died suddenly, perhaps poisoned, on December 26, 1441.

Whereas Nicolò III raised the Estensi state to high position in Italian politics in spite of its territorial and financial limits, his bastard son and chosen successor Leonello (1441–50) gave Ferrara considerable distinction in the fields of art and culture. Leonello had been educated by the humanist Guarino Veronese, called to Ferrara by his father; and the period of his reign was one in which Ferrara represented a lively centre of culture and humanism, filled with painters (Pisanello, Jacopo Bellini, van der Weyden, Mantegna), architects (Leon Battista Alberti), and scholars (centring on Guarino Veronese). Leonello's brother and successor, Borso (1450–71), notwithstanding some military failures, not

only maintained his state and increased its aesthetic and cultural prestige but also received from the Holy Roman emperor Frederick III the title of duke of Modena and Reggio (1452) and from Pope Paul II the title of duke of Ferrara (1471).

The long rule of Leonello's and Borso's half brother Ercole I (1471–1505) marked one of the most important periods for the history of the House of Este and of Ferrara. He succeeded in obtaining considerable political support with his marriage to Leonora, the daughter of the King of Naples. These were troublous times, however. Ercole had to defeat the attempt of a nephew, Nicolò, son of Leonello, to usurp the throne; and then he had to face the hostile coalition of Venice and Pope Sixtus IV, which brought war nearly to the walls of the city (1482–84). The subsequent Peace of Bagnolo, however, though not entirely satisfactory, did free Ferrara from immediate dangers.

Ercole's crucial problem became one of consolidating his own political position through marriages that would bind him to the principal Italian powers: of his three daughters, Lucrezia was married to Annibale Bentivoglio (of Bologna), Isabella to Francesco Gonzaga (of Mantua), and Beatrice to Ludovico Sforza (of Milan). His eldest son, Alfonso, was married first to Anna Sforza (of Milan) and then to the famous Lucrezia Borgia, the daughter of Pope Alexander VI. In spite of these difficult affairs of state, Ercole was able to continue his dynasty's patronage of the arts, taking the poet Boiardo as his minister, extending his favour to the poet Ariosto, espousing the theatre and musical arts, and enlarging and beautifying Ferrara to such an extent as to make it one of the first cities of Europe.

His son Alfonso I (1505–34), rough and rude when he was young, proved wise and sure of himself once he had taken the reins of government. First he foiled a plot of a stepbrother, Giulio, and another brother, Ferrante, against him and sentenced them to perpetual imprisonment. Then his attention was completely attracted by the war against Venice (1509), in which his skill in mechanics and artillery design was proved. He was victorious in the naval battle of Polesella and won back the Polesine of Rovigo (which had been lost by Ercole I). At the same time, however, papal ambitions of territorial expansion became threatening. By consistent adherence to the French interest in Italy, Alfonso came into collision with Pope Julius II and was deprived of Modena (1510) and Reggio (1512) and was excommunicated. The Medici popes, Leo X and Clement VII, were both determined on the destruction of the Estensi, but the first-mentioned pope was frustrated by death, the second by political weakness, and Alfonso was able to recover Reggio in 1523 and Modena in 1527. He died in 1534. His succession was assured not only by his legitimate children but also by the issue of his lover Laura Eustochia Dianti, from whom derived the future dukes of Modena and Reggio.

During the reign of his son and successor Ercole II (1534–59), the military events proved less interesting (though the wars of 1557–58 were difficult) than the personal ones. He married Renée, daughter of King Louis XII of France; and in Ferrara she came to embrace the Lutheran religion, becoming its ardent defender and establishing at her court a meeting place for the most famous heretics and liberal thinkers of the day. Ercole, who was the pope's vicar in Ferrara, tried restraining her, even to the point of temporary imprisonment, but to no avail. Next to rule was his first-born, Alfonso II (1559–97), the fifth and last duke of Ferrara. He also tried, vainly, to be elected king of Poland and to organize a crusade against the Turks. More important for the dynasty, however, was the fact that, though he had three marriages, he had no children; and Pope Pius V in 1567 expressly forbade having illegitimate children rule in ecclesiastical lands. Alfonso was so disappointed and discouraged that he let the conditions of his state decay. At his death he bequeathed the duchy to his cousin Cesare (1562–1628), but Pope Clement VII refused to recognize the settlement, declaring Cesare illegitimate; in 1598 direct papal rule was established in Ferrara. The main branch of the Este family had come to an end.

End of the main line of Estensi

Cesare kept Modena and Reggio, but with him the Estensi ceased to play so important a part in Italian politics, and the court was culturally inferior to its brilliant predecessors. Among the several Modenese dukes who followed in the 17th century, Francesco I (1629–58), who came to the throne during the stormy period of the Thirty Years' War, was perhaps the most important. His people were able to survive the famous plague of 1630. In the wars he was first allied to Spain, then to France, whose alliance he thought would best sustain his claims to Ferrara. Later on, he attempted reconciliation with Spain, but ironically it was on the scene of battle, fighting the Spaniards, that he died of malaria. He was a man with enormous aspirations, and, though inclined toward treachery in politics, he gave art his patronage, favouring men of letters and collecting works of art (there is an extraordinary portrait of him by Diego Velasquez and a beautiful bust by Gian Lorenzo Bernini).

Among his successors, the one most deserving to be remembered is Rinaldo I (1694–1737), whose marriage to Charlotte Felicitas of Brunswick-Lüneburg reunited the long-separated branches of the House of Este. Throughout his reign he engaged in imperial politics. His son Francesco III (1737–80), known as a libertine, received the governorship of Lombardy from Empress Maria Theresa of Austria. Ercole III (1780–96), gentle and affable, abandoned Modena in 1796 when the Revolutionary French Army invaded it.

End of the male line

After the Napoleonic Wars, Duke Francesco IV (1814–46), son of Maria Beatrice d'Este (the only surviving daughter of Ercole III) and of Archduke Ferdinand of Asburgo-Lorena, son of Maria Theresa, came back to Modena. He founded the Austro-Este line in Modena, which, however, ended with his son Francesco V (1846–59) when Modena revolted in order to join Sardinia-Piedmont and then Italy.

Childless, Francesco V selected as his universal heir Francis Ferdinand, heir to the Austrian throne, who was murdered at Sarajevo in 1914. Today, the rightful holder of the surname and heraldic bearings of the Estensi is considered to be Archduke Robert, the second-born of Charles of Asberg and Zita of Borbona-Parma. In 1953 he married the princess Margherita of Savoy-Aosta.

BIBLIOGRAPHY. The classic work on the house of Este is L.A. MURATORI, *Delle antichità estensi ed italiane*, 2 vol. (1717–40), which is still consulted. Others are: A. FRIZZI, *Memorie per la storia di Ferrara*, 2nd ed., 5 vol. (1847–50); E.G. GARDNER, *Dukes and Poets in Ferrara* (1903); and E. NOYES, *The Story of Ferrara* (1904). About particular figures of the House of Este: J. CARTWRIGHT, *Beatrice d'Este, Duchess of Milan, 1475–1497* (1899), and *Isabella d'Este, Marchioness of Mantua, 1474–1539*, 2 vol. (1903). About the Signoria of Estensi in Modena and Reggio: O. ROMBALDI, *Gli Estensi al governo di Reggio dal 1523 al 1859* (1959); M.C. NANNINI, *Gli Estensi di Modena* (1959); and L. AMORTH, *Modena capitale* (1961). The most recent study, provided with rich bibliography, is L. CHIAPPINI, *Gli Estensi*, 2nd ed. (1970).

(L.Ch.)

Estonian Soviet Socialist Republic

The Estonian Soviet Socialist Republic, one of the 15 republics of the Soviet Union, was proclaimed on July 21, 1940, and became a member of the Soviet Union on August 6, 1940. On the north and west the republic is bounded by the Baltic Sea and on the east by Lake Peipus (Chudskoye Ozero) and the Narva River; in the south it borders on Latvia and the Russian S.F.S.R.

Estonia has an area of 17,400 square miles (45,100 square kilometres), of which 9 percent is made up of some 800 islands and islets. The larger islands are Saaremaa (1,048 square miles) and Hiiumaa (373 square miles). The capital is Tallinn. For the history of Estonia, see BALTIC STATES, HISTORY OF THE.

The landscape. *Relief.* The Estonian landscape bears traces of glacial activity; the south is covered with moraine (glacial deposit) hills, and the central part abounds in elongated hills with flat tops. They are usually ar-

Glacial landforms

ranged in groups in the direction of glacial movement, for the most part from the northwest to the southeast. The northern part of Estonia is characterized by long narrow swells consisting of deposits left by glacial rivers that formed during the melting of ice. Extensive sandy areas mark what was once the glacier's edge.

The Estonian relief is thus generally undulating, with small hills and numerous lakes, rivers, and forests lending the scene mildness and picturesqueness, particularly in the south.

The mean absolute altitude is 160 feet, only one-tenth of the territory lying at altitudes exceeding 300 feet above sea level. In the southeast is the Haanja elevation containing Mt. Munamägi (1,042 feet [318 metres]), which is the highest point in Estonia.

During the postglacial period a considerable part of Estonia was flooded by the Baltic Sea. Later, large areas emerged from under water as a result of the elevation of the mainland—a process still observable, particularly in the northwest, where a rise of approximately five feet in a hundred years has been recorded.

In northern Estonia is a low limestone plateau that falls abruptly to the sea. The sheer cliff stretches along the southern shore of the Gulf of Finland from Tallinn and continues beyond the republic's borders. It rises to its maximum height (544 feet, or 166 metres, above sea level) at Pandivere.

Rivers and lakes

Estonia abounds in rivers, which carry their water to the gulfs of Finland and Riga and into Lake Peipus. The longest rivers are the Pärnu (88 miles, or 142 kilometres) and the Pedja (81 miles). Lakes occupy about 5 percent of the republic's territory. The largest lake is Peipus, with a surface area of 1,370 square miles (3,548 square kilometres).

Climate. The temperate and humid climate of Estonia differs sharply from the climates of many regions of the Soviet Union situated at the same latitude but in the depth of the continent. The republic lies in the path of air masses brought in by cyclones born in the northern Atlantic that carry warm air in winter and cool air in summer. The mean temperature in February is 23°–43° F (–5° to +6° C) and in July 61°–63° F (16°–17° C). Annual precipitation is 24–28 inches (610–711 millimetres), which, coupled with negligible evaporation and plain relief, leads to waterlogging. The Estonian climate is generally favourable to agriculture.

Flora and fauna. Mixed forests, with about 90 species of trees and shrubs, cover 3,404,000 acres (1,378,000 hectares), or 30.5 percent of Estonia's territory. Most widespread are pine, fir, birch, and aspen; less common are oak, lime, maple, elm, and ash. Meadows occupy a considerable area, as do marshes and swamps.

About 60 species of mammals live in Estonia. The elk is the largest; roe deer, red deer, and wild pigs are also found. In the deep forests of the northeast, bear and lynx are encountered. Foxes, badgers, otters, rabbits, hares, and—along the riverbanks—mink and nutria (coypu) are fairly common. Among the sea animals, seals and fishes (cod, salmon, eel, plaice, and others) are of commercial importance. Birds are very numerous; 295 species have been identified, of which 60 are year-round residents.

Minerals. The most important mineral is bituminous shale; the output of Estonia accounts for 80 percent of the shale produced in the Soviet Union. Reserves and production of peat also are substantial, and large deposits of high-quality phosphorites, limestone, dolomites, marl, and clay exist.

The people. Estonians belong ethnically and linguistically to the Baltic-Finnish group of Finno-Ugric peoples. The population was 1,429,000 in 1975; at the census of 1970 Estonians comprised 68.2 percent, and 24.7 percent were Russians, 2.1 Ukrainians, 1.4 Finns, and 1.4 Belorussians. The average density of population was 82 persons per square mile. Town dwellers comprised 67 percent of the population.

The development of Estonia's economy has been attended by a considerable internal migration. The population in the northern, industrially advanced part of the republic has increased appreciably since the mid-1950s at the expense of the southern and western regions, which are basically agrarian. There has also been extensive immigration from other territories of the Soviet Union, particularly the Russian S.F.S.R.

Towns

There are 33 towns and 26 urban settlements in Estonia. The capital—Tallinn—has a population of 392,000. The other major towns are Tartu, Kohtla-Järve, Narva, and Pärnu, with populations ranging from 25,000 to 100,000.

The economy. Estonia is basically an industrial region, with agriculture also making a contribution.

Fuel and power. About 20 percent of the industrial workers of Estonia are employed in producing bituminous shale and phosphorites or in the power industry. Estonia, the only area of the Soviet Union that has a large shale processing industry, produces about 75 percent of the Soviet Union's artificial gas, much of which is transported by pipelines extending from Kohtla-Järve to Leningrad.

More than 90 percent of the electricity produced in the republic is generated by thermal power plants fired with bituminous shale. One station near Narva, the Estonskaya, which began to operate in 1968, accounts for more than 60 percent of the electricity produced in the Soviet Baltic republics. Other major power plants include a hydroelectric station at the Narva falls and a peat-fired plant at Ellamaa. The electrical power industry has great significance both for the economy of Estonia and for the whole northwest Soviet Union.

Manufactures. Estonia's industry uses both local resources and imported raw materials, those utilizing imported raw materials accounting for 50 percent of the gross industrial output. About 20 percent of the industrial labour force is engaged in engineering and metalworking activities that provide oil-refining equipment, agricultural implements, mining machinery, gas pipes, and excavators. Technical and scientific instruments and electronic apparatus are also produced.

The chemical industry

Shale processing underlies a developing chemical industry centred in Tallinn and Kohtla-Järve. Such products as benzene, adhesives, tanning agents, resins, formaldehyde, and detergents are made.

Estonia's natural resources provide a base for the production of building materials, including cement, mural blocks, and panels made either from shale ash or reinforced concrete. The main centres of this industry are Tallinn, Kunda, Tartu, and Aseri.

Timber and woodworking make up one of the oldest industries of Estonia. As a result of a disorderly exploitation of woodlands in former times the total wooded area decreased more than 15 percent. Many thousands of acres of new trees have been planted, but in order to satisfy its needs Estonia is nevertheless obliged to import from the northwestern regions of the Soviet Union about one-fourth of its requirements. Paper, pulp, plywood, matches, and furniture are among the republic's wood products. The main centres of production are Tallinn, Tartu, Narva, Pärnu, Kehra, Kuressaare (Kingissepp), and Viljandi.

Among consumer goods industries, textiles are the most developed. Seventy percent of all cotton cloth produced in the Soviet Baltic republics is manufactured in Estonia. The republic also produces wool, silk, and linen. Knitted and woven garments and shoes are also produced.

Agriculture. After 1945 Estonia's agriculture was collectivized. Instead of the former 120,000 small peasant farms there were by the mid-1970s more than 300 collective farms and about 150 state farms. Agriculture accounts for 20 percent of the gross national product and engages 26 percent of the labour force.

The land is generally fairly difficult to farm. Large areas have to be cleared of stones, and 70 percent of the natural pastures require draining. Cattle and pigs account for about two-thirds of the Estonian S.S.R.'s agricultural production.

Fodder crops for animal husbandry account for more than 45 percent of total crop production. Grain crops, legumes, and potatoes represent most of the balance.

An increased level of mechanization and the application

of more advanced methods of work in agriculture have resulted in a considerable increase in labour productivity and in agricultural incomes.

Transport and communications. Transport systems in Estonia include railways, roads, rivers, and pipelines. Most of the republic's freight (68 percent) is carried by railways; the share of road transport is 30 percent, that of sea transport 2 percent. River transport is of local significance only.

Estonia is connected by air with Moscow, Leningrad, the capitals of the union republics, and the Black Sea resorts.

Administrative and social conditions. *Constitutional and political framework.* Estonia has a state emblem, flag, and national anthem. It is a socialist republic, and its constitution (1940) declares it to be a sovereign entity, though in fact it has no power to carry on relations with foreign countries. The highest organs of state power are the Supreme Soviet, elected for a four-year term, and the Presidium, selected by the Supreme Soviet. The highest executive and administrative organ of government is the Council of Ministers. In the provinces government is effected through local soviets.

Communist Party of Estonia

All political life in the republic is under the direction of the Communist Party of Estonia, which in 1974 had some 79,000 full and candidate members. The party concerns itself with the political and ideological education of the population, establishes policy for the development of the economy, and directs the activities of the Young Communist League of Estonia (Komsomol; created in 1921), which in 1974 had a membership that totalled 128,000.

Trade union membership totalled some 618,000 in 1970. Unions do not, as in the West, represent the workers in negotiations with management; their efforts are directed more to providing incentives for increased production, maintaining labour discipline, and serving as an instrument of Communist Party policy.

Welfare and culture. As a result of the development of industry and agriculture, the economic and social welfare of the population has increased substantially compared to the period immediately after World War II. Over the decade 1960–70 the national income doubled, permitting advances to be made in such areas as pensions, education, medical care, and housing.

Scientific research in the republic is centred on the Academy of Sciences of the Estonian Soviet Socialist Republic (founded in 1946), which has three departments: physics, mathematics, and technical sciences; chemistry, geology, and biology; and social sciences. In the 1970s the academy supervised about 12 institutes and 11 other establishments employing approximately 900 scientific workers.

More than 40 newspapers are published in Estonia, some with a circulation of more than 1,000,000 copies. Book publishing also flourishes, the output in 1973 comprising 1,966 titles and 13,120,000 copies.

Estonian radio broadcasts diversified programs. The system possesses a mixed chorus, a symphony, and a variety orchestra. The republic has had a television centre since 1955, and television broadcasts from Moscow and Leningrad are also received in most areas.

Literature

The scope and importance of Estonian literature has steadily increased since the period of national awakening in the 19th century. The greatest achievements in the first half of the 20th century were made in the genres of realistic novel (Anton Tammsaare) and imaginative poetry (Marie Under, Betti Alver). Open to cultural and literary influences of western Europe, Estonian literature developed a diversity of styles, ranging from Neoclassicism to bold experimentation. Since World War II, the dominant theme has been Soviet reality, expressed in the form of Socialist Realism especially in the prose writings of Soviet Estonian authors (Juhan Smuul). During the 1960s, a younger generation of exceptionally talented poets (Paul-Eerik Rummo, Jaan Kaplinski) brought about a revival of Estonian poetry that continued into the '70s (Jüri Üdi). A new and sophisticated genre, the so-called short novel, deals primarily with psychological problems (Enn Vetemaa, Mati Unt), sometimes projecting them into history (Jaan Kross). New developments can also be found in drama and short story. Both Estonian classics and the works of contemporary authors have been translated into many languages.

The beginning of professional theatrical art in Estonia is closely connected with the creation of the Vanemuine Theatre in Tartu in 1870. Tallinn has five theatres, including an opera and ballet theatre, a drama theatre, a youth theatre, and a puppet theatre. There are also professional theatres to be found in Pärnu, Viljandi, and Rakvere.

Tallinn is famed for its song festivals. There is an enormous stage for popular mass performances, and tens of thousands of singers and up to 200,000 spectators come to take part in these events.

Prospects. Natural resources, investment funds, and a qualified body of workers and specialists are the foundation of the future development of the national economy of the republic. The ninth Soviet Five-Year Plan, adopted in 1971, laid down the following directives for the Estonian economy: major emphasis was to be given to those branches of machine building that do not consume much metal, particularly the electrotechnical and the instrument-building industries; the light industries and meat and dairy production were also to be given increased emphasis.

During the five-year period, gross industrial production was planned to increase 36–38 percent. The output of shale was to reach 26,000,000 tons annually, necessitating the building of associated plants.

The output of electricity was expected to increase 1.4 times, construction of the Estonskaya power station was to be completed (it became fully operational in 1973), and work was to start on building the Tallinn thermoelectric station.

The fish catch was expected to increase, and the production of the fish-processing industry was expected to expand almost one and a half times.

Estonia's agriculture, it was anticipated, would become still more specialized, with attention focussed on meat and milk production, pig breeding, and potato growing. There were plans to drain 519,000 acres (210,000 hectares) of swampy regions; to further improve land quality over 370,000 acres (150,000 hectares); and to create productive meadows and pastures covering some 320,000 acres (130,000 hectares).

(A.A.Ke./V.J.T./Ed.)

Estuaries

An estuary is a partly enclosed body of water that forms where river water is mixed with and diluted by seawater. In a general sense, the estuarine environment is one that is defined by salinity boundaries rather than geographic boundaries.

Many coastal features that are designated by other names are, in fact, estuaries. The lower reaches of rivers such as the Hudson, Thames, and St. Lawrence are estuaries. Other river systems with large deltas are actually estuaries at their seaward end; the Mississippi is an example. Certain coastal embayments, such as Chesapeake Bay and Galveston Bay, also are estuaries because fresh and salt water undergo considerable mixing. In addition, most of the fjord systems in the world are estuaries (*e.g.*, Oslofjorden, Hardangerfjorden, and fjord systems in the Pacific Northwest of North America). Large, semi-enclosed tidal flat regions and coastal marshes, such as Pamlico Sound, the Waddenzee area of The Netherlands, and those of New England, also may be termed estuaries.

The adjective estuarine can be applied to aqueous environments where seawater has become somewhat diluted by fresh water. For example, the northernmost Pacific Ocean in the vicinity of the Bering Sea and Gulf of Alaska has estuarine characteristics because of the great quantity of fresh water that flows into this region. The same holds true for the Baltic Sea, Gulf of Bothnia, Gulf of Finland, the southeastern section of the North Sea, and the coastal zone around Norway. In these regions

high amounts of precipitation and runoff dilute the coastal seawater until it has estuarine characteristics. Ocean areas off the mouths of the Mississippi, Amazon, and Columbia river systems also become estuarine during the rainy seasons when the river discharge increases and dilutes the coastal waters.

Importance to man

Estuaries have long been important to man as harbour sites and centres of commerce. Some of the oldest continuous civilizations have flourished in such estuarine environments as the lower region of the Tigris and Euphrates rivers, the Po River Delta region of Italy, the Nile Delta, the Ganges Delta, and the lower Hwang Ho Valley. As civilizations developed, it was logical that the seaports were founded at the seawardmost point of the major river systems. In this way commerce could move cheaply and relatively safely down the rivers and be loaded onto ships to be moved to more distant ports. Cities including London (Thames River), New York (Hudson River), Montreal (St. Lawrence River), Baltimore (Chesapeake Bay Estuary), New Orleans (Mississippi River), Houston (Galveston Bay), San Francisco (San Francisco Bay), Hamburg (Elbe River), Liverpool (Mersey River), and Bordeaux (Gironde Estuary) have developed on estuaries and have become important centres of commerce in the modern world.

The world's fishing industry is somewhat dependent on the estuarine environment, either directly as an area of fishing or indirectly as a region where the species spend a portion of their life cycle. In the United States alone, the Pacific salmon, anadromous trout, shrimp, and menhaden fisheries provided the majority of the total weight of fisheries' products landed by United States fishermen. These species are all estuarine dependent for part of their life cycle.

This article treats the formation and maintenance of estuaries, their general hydrology, and the ecology of estuarine life. For further information on relevant physical factors, see WATER WAVES; TIDES; RIVERS AND RIVER SYSTEMS; and HYDROLOGIC CYCLE; see also OCEANS AND SEAS; GULFS AND BAYS; COASTAL FEATURES; RIVER DELTAS; SWAMPS, MARSHES, AND BOGS; and AQUATIC ECOSYSTEM for a description of the nature and the biota of neighbouring environments.

THE FORMATION OF ESTUARIES

Changes of sea level. The geologic processes that form an estuary are extremely complex and varied, but it is clear that the existence of an estuary or estuarine environment is largely dependent on the position of sea level relative to the freshwater discharge. The estuarine brackish water zone is relatively narrow in width and is located at the margins of the continents. If sea level were lowered, the estuarine zone would migrate seaward at the interface of the marine water and the edge of the newly exposed land area. Such migration has occurred in the past as a consequence of the Earth's several continental glaciations. For each glaciation the primary source of moisture has been the oceans. Whenever sea level fell, the estuarine environment at the continental margin was forced to migrate in a seaward direction. Areas of marine water became brackish and supported fauna that were typically estuarine rather than marine. Empty shells of the oyster *Crassostrea virginica* are found in two bands at 70-metre and 50-metre depth on the Atlantic shelf of the United States, for example. Today these oysters are only found in estuarine waters shallower than ten metres, in which the salinities are much lower than in the open ocean. In addition, bottom sampling of the continental shelf off New England often reveals past deposits and intertidal sediments that are indicative of a previous estuarine environment.

Relation to former ice ages

Approximately 18,000 years ago the Wisconsin Glaciation attained its maximum, and glacial melting began. The seas rose, forcing the estuarine environment to migrate back up the continental shelf. During the period of lowered sea level, some rivers had become entrenched in the continental shelf and deepened their valleys. These valleys were soon flooded by the rising marine waters, forming a typical drowned river estuary. In areas such as Norway and parts of the coast of British Columbia, valley glaciation had served to deepen river valleys. These narrow drowned glacial valleys became the modern fjord estuaries as sea level rose.

When sea level rose, the coastal plains and deltas near the outer edge of the continental shelf were the first to be submerged. The shoreline moved over the shelves into the valleys of river systems that previously had transported sediments far across the continental shelf. The physical boundaries of the estuary then became more complex as brackish water filled the former tributaries of recently drowned river systems. In some cases entire broad valleys were filled by the rising of sea level, and in this way some large, shallow estuaries, such as Chesapeake Bay, were formed.

Geomorphology of estuaries. The geomorphology, or form, of an estuarine basin is usually developed by fluvial erosion, fluvial and marine deposition, glacial erosion, or tectonic activity. These processes are associated with four types of basins, namely, fluvial, depositional, glacial, and tectonic.

Fluvial basins. Drowned fluvial valleys are probably the most common type of estuarine basin. Where the valleys were simple and straight with few tributaries, drowning has simply raised the level of water on the valley walls. This is true of the Hudson and St. Lawrence rivers, for example.

In other river systems, especially those that were developed on softer, more unconsolidated rocks, the rivers were quite complex and have an elaborate tributary system. When these regions are flooded, the resulting estuary exhibits an extremely crenulated border that follows the contours of the former tributary valleys. An example of this type is Chesapeake Bay on the eastern coast of the United States (Figure 1).

Depositional basins. Some estuarine basins are formed when barriers that separate the ocean and river waters develop. This has occurred off Galveston and Corpus Christi bays, where barrier islands have, in effect, created an estuarine environment sandwiched between the inner Texas coast and the offshore islands. The Vadehavet, the Danish tidal area, and the Wadden Zee of The Netherlands also have offshore barriers that in part act as a permeable wall behind which the estuary has formed.

Natural levees, spits, and bars also can restrict inlets and thus permit the establishment of estuarine environments in coastal indentations that were formerly open to the sea. Spits commonly form at seaward ends of large estuarine systems and further restrict free exchange between the open ocean and the river water. This has happened to the Swan Estuary of Western Australia.

Fjords as estuaries

Glacial basins. Probably the most picturesque of all estuaries are the fjords. Fjords are deep, narrow, U-shaped gorges that were cut in bedrock by tongues of glacial ice during former ice ages. Many of the fjords were former river valleys that were deepened and widened by ice action. Because of the thickness and weight of the ice, the glaciers were able to erode far below sea level. In addition, the seaward end of the fjord is usually marked by a bedrock or depositional sill that forms an internal basin. This means that the fjords are extremely deep and narrow with a shallow restrictive entrance. The shallow depths of the sills prevent free exchange of the oceanic water outside the sill and the water in the basin. Because of this lack of renewal of oceanic waters to the basin, many fjords have become oxygen deficient and will not support life in their deeper layers.

Tectonic basins. Continental tectonics, or movement of the land surface, may also create estuaries. A common form is the graben, or downfaulted block; Tomales Bay, off the northern California coast, is a graben-controlled estuary.

Broad local subsidence of portions of a coast also can create estuarine environments because its effect is equivalent to a rise in sea level. The San Francisco Bay region, which has had a complex geologic history, is in this category.

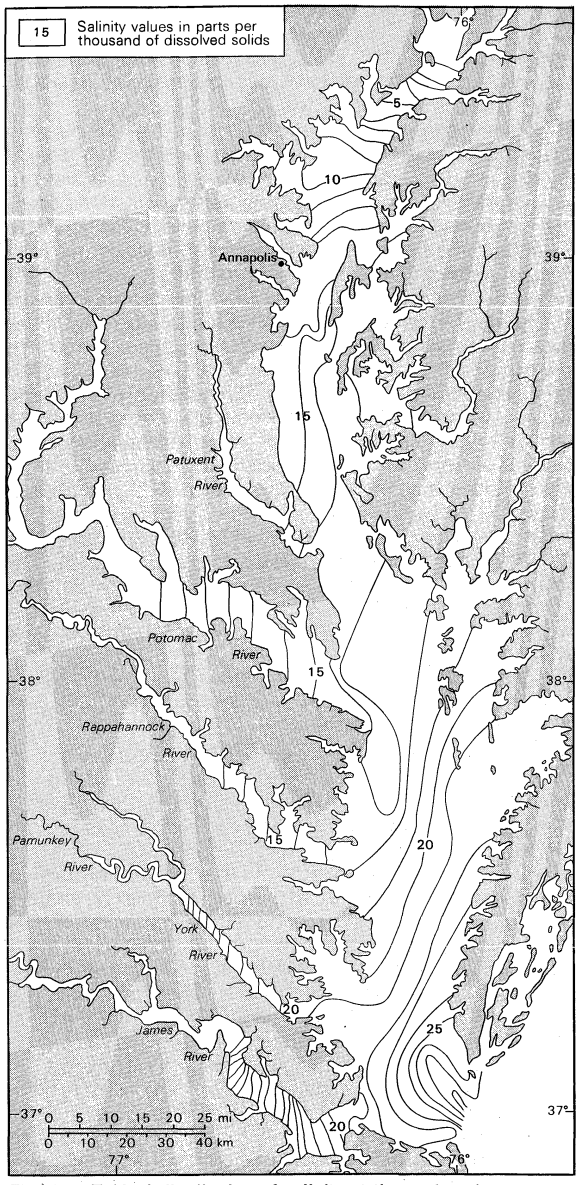


Figure 1: Typical distribution of salinity at the surface in Chesapeake Bay.

Adapted from J. McHugh, "Estuarine Nekton," *Estuaries*, pub. no. 83; © 1967 by the American Association for the Advancement of Science

Factors tending to destroy estuaries. There are four major factors that tend to destroy the estuary: these are sediment influx, vegetation encroachment, relative changes in sea level, and the activities of man.

Sediment influx. The most natural enemies of the estuary are the river and marine sediments that tend to fill it. In the river environment the current velocity is directed downstream and sediment is constantly moving along; the movement is rapid during flood season and periodical during the quieter water of low discharge periods. When river sediment reaches the estuary, the coarser fraction tends to settle out, often forming deltas. Even the finest clays are affected by the brackish water of the estuary and may begin to flocculate and accumulate. The net effect of this sedimentation is a gradual loss of estuarine area and a decreasing mean depth of the estuarine system. Rates of sediment deposition are variable, but one to three metres per thousand years is probably an average figure.

Infilling by sedimentation and vegetation

Vegetation encroachment. As a natural consequence of sedimentation, vegetation will develop in the estuary. The increase of the vegetable cover depends on the local climate, salinity of the estuary, river drainage, and available types of vegetation. In a broad, shallow estuary that is undergoing deposition in a temperate climate, *Spartina* grasses are one of the first and hardiest of the forms of vegetation to begin filling the estuary. Shoal areas that are exposed during the tidal cycle initially become colonized, and once the plants gain a foothold they act as anchoring devices to stabilize the sediments. Vegetation also acts as a dampening agent to currents and waves, creating areas where finer sediments can settle out. As the vegetation spreads, more sediment is trapped and deposited, and soon the tidal flat is a series of tidal channels separated by expanses of vegetation. The vegetation thickens and grows upward, often creating peatlike deposits below the surface. As the vegetation and peat continue to build up, the region is less influenced by the daily tidal cycle and before long gives way to species of plants that utilize fresh water. The former tidal flat is then relatively divorced from the marine environment.

In tropical estuaries, the vegetation sequence is dominated by mangrove trees, whose densely interwoven systems of stems and roots rapidly build up a swampy shore. With time, the substrate may build up above the reach of the tide, and the mangrove swamp may give way to oak, mahogany, and other hardwood trees. One of the greatest mangrove swamps in the world exists in southwestern Florida where the brackish estuarine water forms ideal environmental conditions. Although the estuaries may lose area, valuable land is often built up by mangrove growth. In the last 40 years in the Biscayne and Florida Bay regions, some 1,500 acres of new land have formed as a direct result of mangrove growth.

Relative changes in sea level. Relative changes in sea level also can modify or shift the estuarine environment. Certain areas of the world that were depressed by ice loads during glacial times, for example, are now rising out of the sea. Some Scandinavian shorelines have risen as much as 300 metres since deglaciation in an attempt to re-establish equilibrium. The brackish water estuarine environments of the Gulf of Bothnia, Gulf of Finland, and even the Baltic Sea are in part related to the former depression of the Earth's crust in these areas by glacial ice. In time these regions will become less marine as they are uplifted relative to world sea level.

Activities of man. Man's effect on estuaries can be dramatic. Dams or locks built on rivers may prevent some estuarine deposition, but upstream deposition will be increased. In the three years following World War II, when the locks of the Rochefort Basin of France were destroyed, some seven metres of sediment accumulated in the estuarine portion of the Charente Estuary. As a result of cultivation, plowing, and deforestation, more sediment is now transported into the estuaries than ever before. In Madagascar, on the Betsiboka Estuary, deforestation has increased the sediment load and deposition to such an extent that an entire harbour built in World War II is now completely filled.

Effects of dams and locks

In general, estuaries are fragile environments that can respond dramatically to minor changes in the sediment load, vegetation, and sea level. True estuarine sedimentary rocks are rare occurrences, and this suggests that estuaries in the past also have been only temporary environments that responded to minor fluctuations of various processes and factors.

HYDROLOGY AND SEDIMENTATION IN ESTUARIES

Factors controlling and affecting estuarine circulation. *The tides.* The tides have a very important effect on the circulation within an estuary. The volume of water within the estuary between high and low tides is known as the tidal prism. The currents generated by this volume of water flowing in and out of the estuary depend on the geometry of the tidal basin and the tidal range (distance sea level rises from maximum ebb tide to maximum flood tide). In estuaries where the tidal range is large, the tidal currents have the ability to mix both the inflowing river water and ocean water in the estuarine basin. In estuaries with smaller tidal ranges and lower current velocities, the river water may remain relatively unmixed within the basin and form a low-density layer that floats on the surface of the denser salt water.

River discharge. The quantity of freshwater inflow is important in determining the nature of the estuary. During periods of high river discharge, fresh water may force most of the salt water out of the estuary. During other seasons when river discharge is low, the estuary may be more akin to an extension of the ocean rather than an estuarine environment. Moderate river discharge may stratify, or layer, the estuary, with the low-density brackish waters on top and the denser, more saline waters below.

Wind. Wind can strongly affect the mixing of estuarine waters by creating local waves that break down water stratification. In addition, through stress on the surface, the wind can cause a net transport of surface water and thus reinforce or retard the normal tidal flow.

Ice. The formation of ice within estuaries can act in many ways to affect the estuarine circulation. If large portions of the estuary are covered by ice, local wind-generated waves will have little effect in mixing the surface and bottom waters. In contrast, restricted channels caused by ice blockage may locally increase the tidal currents and mixing of the waters. Melting of estuarine ice in the spring also provides water that is of lesser density than normal estuarine water and may lead to stratification.

Physical dimensions. The physical dimensions of an estuary are extremely important with respect to the hydrology. Broad, shallow estuaries like Great Bay, New Hampshire, are much more susceptible to the effects of wind mixing by local waves than are deeper estuaries. Narrow, constricted entrances in some estuaries create a Bernoulli effect, wherein the tidal current increases as the water passes through the narrow region. The currents under Golden Gate Bridge in San Francisco Bay are an example. Certain fjords with restrictive sill depths at the seaward portion of the estuary inhibit circulation and mixing to such an extent that the lower waters in the estuary become devoid of oxygen.

Fresh and brackish water and the saltwater wedge

Types and examples of estuarine circulation. In most estuaries there is a certain amount of to-and-fro motion of the waters as the tides flush in and out the basin. Besides this, however, there is a net circulation pattern within estuaries that is a function primarily of the mixing of the fresh and salt water in this complex area. This is termed estuarine circulation

Highly stratified estuary. Most highly stratified estuaries occur in regions where the tidal range is low and river discharge into the estuary is high. In these cases, the less dense surface water of the river will float over the salt water, forming stratification, or layers of water, within the estuary. The fresh river water that forms the surface layer quickly becomes brackish as it mixes with the estuarine waters.

It is possible that internal waves develop at the interface between brackish surface water and the underlying salt water. The formation of these waves would cause some of the underlying salt water to mix with the overlying, lighter brackish water. This means that some salt water would be transported seaward in the upper layer and, as a net result, means a slow movement of salt water along the base of the estuary up the river. This type of estuary often is called a salt-wedge variety because of the wedge of salt water that intrudes at depth (Figure 2). The lower end of the Mississippi River is an example of this type of estuarine circulation.

Conditions leading to highly stratified estuaries are enhanced in regions where there is high river discharge and low tidal flow. In addition, highly stratified estuaries are more likely to develop in deep, narrow, drowned river valleys.

Moderately stratified estuary. As tidal flow increases, or as river discharge decreases, more mixing takes place between the upper and lower waters in an estuary. The sharp salinity gradient, called a halocline in highly stratified estuaries, is more diffuse in a moderately stratified estuary. The greater amount of tidal mixing adds a larger volume of salt water to the upper layers than in the salt-wedge type of estuary. The net flow in this system is still primarily two layer, with the surface water moving seaward and the bottom water moving landward. Now, however, because of mixing, circulation is much greater and there is more flow both in the upper and the lower salt waters. The volume of water circulation in this type of flow far exceeds the salt-wedge type. For example, the seaward flow in the surface layer may be 30 times the river discharge whereas the inflowing water at the base of the estuary is 29 times the river flow (Figure 3). Moderately stratified estuaries include the James River, the Thames River, Chesapeake Bay, and portions of the Strait of Juan de Fuca.

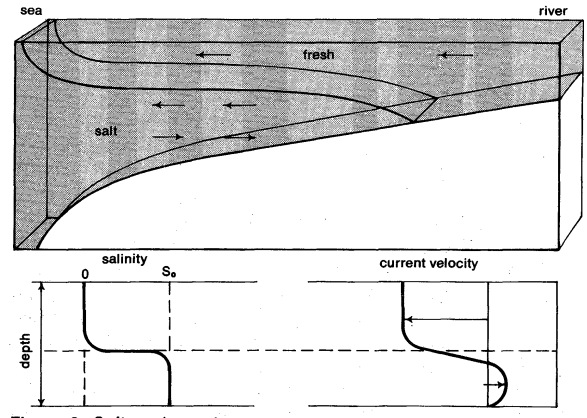


Figure 2: *Salt wedge estuary.* (Top) Section along estuary. (Bottom) Typical salinity and velocity profiles.

From K. Bowden, *Estuaries*, pub. no. 83, p. 17; © 1967 by the American Association for the Advancement of Science

From K. Bowden, *Estuaries*, pub. no. 83, p. 18; © 1967 by the American Association for the Advancement of Science

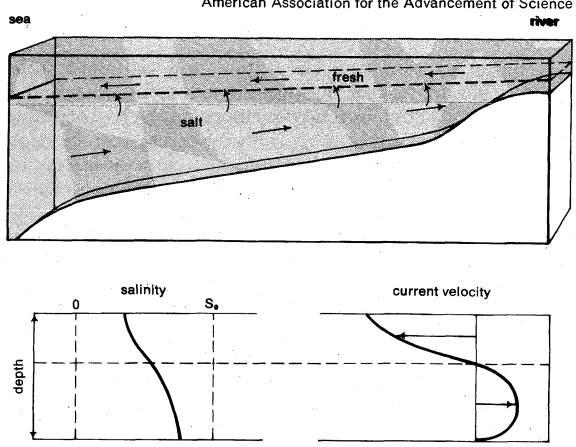


Figure 3: *Partially mixed estuary with entrainment and mixing.* (Top) Section along estuary. (Bottom) Typical salinity and velocity profiles.

Vertically homogeneous estuary. As mixing increases due to tides and wind, the water becomes less stratified and ultimately reaches a point at which its properties from top to bottom are relatively constant at any one location within the estuary. Although the vertical gradient is erased, a strong horizontal gradient extending from brackish water near the head of the estuary to salt water near the mouth still exists.

Coriolis effect and circulation

In certain broad, shallow estuaries there is a lateral salinity gradient across the estuary as a function of the Coriolis effect. In the Northern Hemisphere lower salinity water moves on the right side of the estuary whereas oceanic water enters the system on the left. The net effect is still to maintain vertical homogeneity, but a counterclockwise circulation is established within the estuary. Some estuaries of this type are the relatively wide portions of Delaware Bay and the Raritan Estuary in New Jersey.

In smaller estuaries of this type, the lateral gradients and cyclonic circulation may not exist. In these estuaries there may be net flow seaward at both the surface and the bottom. The exact mechanism of this water move-

ment is poorly understood, however. Perhaps the closest natural example is provided by parts of the Severn Estuary of England.

Fjords. Fjords are somewhat unique estuaries, and it is difficult to classify them according to the above scheme. The freshwater source is usually concentrated at the head, and there is a strong halocline that exists within the fjord during most seasons. The brackish water flows out in the upper layers and in the process entrains some salt water. The salt water enters the estuary over a restrictive sill and fills up the basin with waters of fairly constant salinity. Most of the mixing and entrainment occur at a fairly shallow depth; consequently, the lower waters of the deep, narrow fjords are sluggish and rarely mixed. With time, decaying organic matter falling into the basin uses up the oxygen, and the bottom waters become anoxic (without oxygen) and toxic to life-forms. Many of the fjords found in Norway and British Columbia are of this variety.

Other estuaries. Many other estuaries tend to defy classification. Certain bar-built estuaries have very restrictive inlets with extremely shallow basins, for example. A wind-driven circulation may be the most important factor involved in water transport within these estuaries.

In other estuaries tidal mixing occurs only in portions of the estuarine system where land barriers, islands, or similar features cause currents to increase as they move across or through the barriers. In this case the estuary changes type both up the estuary and through time within the tidal cycle.

Other estuaries change through time as rivers are dammed up, harbours become dredged, or siltation occurs. Some estuaries may become nonestuarine as basins are blocked by hurricanes or by man-made efforts to create freshwater environments.

Relation of estuarine circulation and sediment transport. The motion of the tides have a pronounced effect on the river flow into a typical estuary. In the lower reaches there is a to-and-fro motion of the surface water over the tidal cycle. Proceeding upstream, however, the river becomes more dominant until a point is reached at which the tides no longer can reverse the river current. At this point the water usually is fresh from top to bottom. From this zone seaward the salt content in the lowermost waters increases. If tidal mixing is not too great in this region, the river water, being less dense than the seawater, rides up over the salt water and may actually lose contact with the riverbed. At this ideal stage, a two-layer system exists with the lower water acting as a barrier or cushion to the overlying "fresh" water. This zone is not fixed but will migrate to-and-fro with the tides. If the effect of the tides is removed by averaging the to-and-fro motion, there will be a boundary between the surface and lower waters along which there is little actual motion. This region where the downstream estuarine circulation is balanced by the upstream circulation has been called the level of no motion.

Tidal versus fluvial transport

A river may introduce sedimentary particles of several sizes, including sand, silt, and clay. A sedimentary particle of sufficient size and density may settle down through the fresher surface water moving seaward. As it settles into the saltwater layer, however, its direction of travel will be changed, and it may be transported back up the estuary along the bottom. Assuming the particle is carried in suspension, it may even be entrained in the surface water again, and again continue on its seaward journey. The process could be repeated again and again before the particle finally came to rest or was carried out to sea.

Because of the cyclic nature of this suspended-sediment transport, there exists a turbid zone near the head of the estuary where these fine particles are recycled in suspension. The sizes of the particles contained in the maximum turbidity zone is not known in great detail, but fine silt and some clay probably are present. If a particle is too fine, it will not settle through the upper layer to the lower zone. Instead the particle will be transported seaward and finally escape from the estuary. On the other hand, if a particle is too coarse, it will settle through both the surface and lower layers and consequently it will form a bottom deposit.

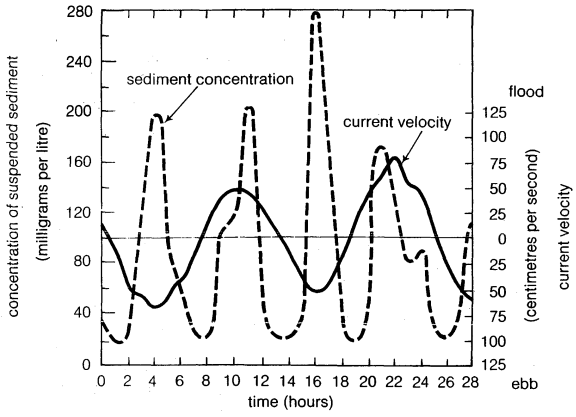


Figure 4: Relationship between suspended sediment near the bottom of an estuary and the tidal current velocity.
From J.R. Schubel, *Science*, vol. 161, p. 1014, © 1968 by the American Association for the Advancement of Science

Sources of estuarine sediments. In many estuaries the main sources of sediment are the associated rivers. In the Rhône Delta region of southern France at least four-fifths of the sediment is derived from the Alps. The Vigo River Estuary, Spain, contains estuarine sediments primarily derived from the erosion of continental weathering products. In the upper Chesapeake Bay region, over 80 percent of the total suspended sediments are thought to be due to discharge from rivers emptying into the bay. Most of this sediment is carried in by the Susquehanna River during high river discharges in the early spring. In the Gulf of Mexico, the Apalachicola and Mobile estuaries contain clays that are typical of those found inland from the coast, and this suggests that the rivers are major sources of sediment.

Estuarine sediment also can be derived offshore and then transported landward in the intruding salt water. Investigations of the Lek Estuary of The Netherlands shows a displacement of sand from the offshore region into the estuary. Work on the Wadden Zee has shown that most of the sediments were derived from marine erosion and transported inshore. Investigations on the Schelde Estuary at the Belgian–Netherlands border have shown that most of the fine silts have originated from the southern North Sea region. There is also evidence from the lower Seine Estuary in France that most of the sediments are of marine origin. This is supported by the fact that in a single year more than 9,500,000 cubic metres of sediment were deposited in the Seine, whereas only 1,000,000 cubic metres were brought down by the river.

Marine derivation in the North Sea area

The ability of tidal currents to erode and resuspend the amount of organic detritus in the suspended matter within the estuary also can be an important factor. In estuarine marshes off Georgia, some 90 percent of the total material in suspension is of organic origin. Most of this is from the breakdown of *Spartina* species. In the upper portions of the Chesapeake Bay region approximately 25 percent of the suspended material is organic.

In summary, then, there are many sources contributing sediment to an estuarine environment. Many of these sources change with the seasons. For example, during the spring and other rainy seasons high river discharge may be the most important source of sediment to the estuary. In the summer, local plant productivity may cause a tremendous input of "suspended sediment" into the estuary. In the fall storms may stir up the bottom and increase the turbidity over short periods of time. The sources of sediment are almost as varied as the types of estuaries themselves.

The concept of flocculation. Fine-grained clay sediment remains in suspension in river waters if the current velocity and turbulence are high enough to keep the par-

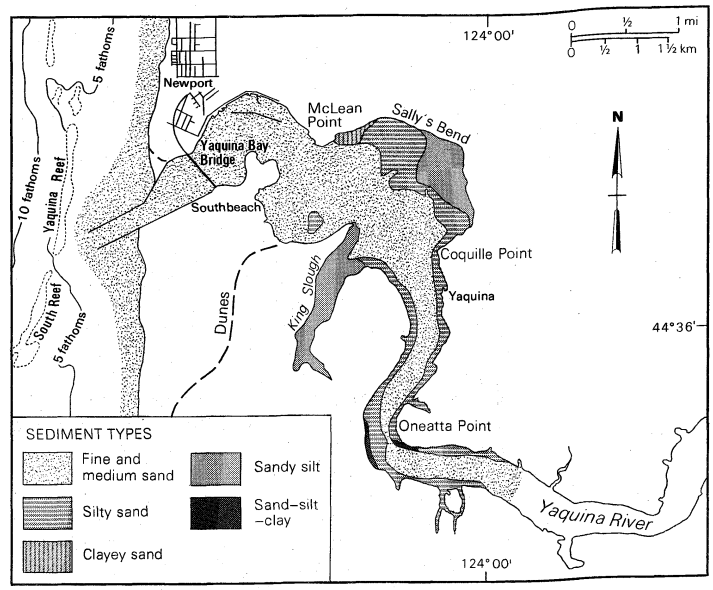
ticle from settling. Each of these very small, flat clay particles settles through a water column at a rate proportional to its density, size, and shape; the molecular viscosity of the water; and other factors. Clay particles in fresh water have a negative electrical charge that attracts and holds dipole water molecules around it. The clay particle then is surrounded by a "hydrated" shell of water that prevents contact with other clay particles.

Creation of larger sedimentary particles

As the clay particle moves from the river to the estuarine environment, it begins to encounter measurable quantities of sea salt. Sea salt cations, especially magnesium and calcium, tend to selectively replace the dipole water molecules surrounding the clay particle. The net effect is that the "hydrated" shell thickness is reduced and particles may come into contact and stick to one another. These aggregated clay particles are called flocs and the process is called flocculation. In sedimentation, the main effect of flocculation is to create larger sized particles, the aggregate weight of which will permit them to settle in the estuarine environment. The process of flocculation is probably reversible; particle aggregates carried back into fresh water by the underlying saltwater intrusion will tend to disassociate.

Distribution of bottom sediments in an estuary. The most landward portion of the estuary is characterized by dominance of fluvial or river sediments. In many estuaries, the point where the river first enters the estuarine basin is marked by deltaic deposits, primarily sands and silts. Farther seaward, where the estuarine circulation patterns are better developed, finer grained deposits of silts and clays commonly occur on the sides of the estuary and coarser sand deposits in the tidal channels. The finer grained material on the sides of the estuary reflect quieter zones of deposition whereas the estuarine channels are regions of higher current velocities that keep all but the coarsest sediment from settling out. The shoreward portion of the tidal flat regions of the estuary are often marked by marsh growth and encroachment of vegetation into the estuary.

Farther seaward, the marine influence becomes increasingly dominant. Quite often tidal deltas are found at the mouth of the estuary extending both seaward and landward from the restricted entrance. The higher tidal currents associated with the entrances are reflected in the coarser sediments found in this region. In addition, coarse-grained materials from outside the estuary from beaches and offshore deposits are often washed into this zone and dominate the bottom sediments. Figure 5 shows a typical estuary and the type of sediments that are found in the various portions. (F.E.A.)



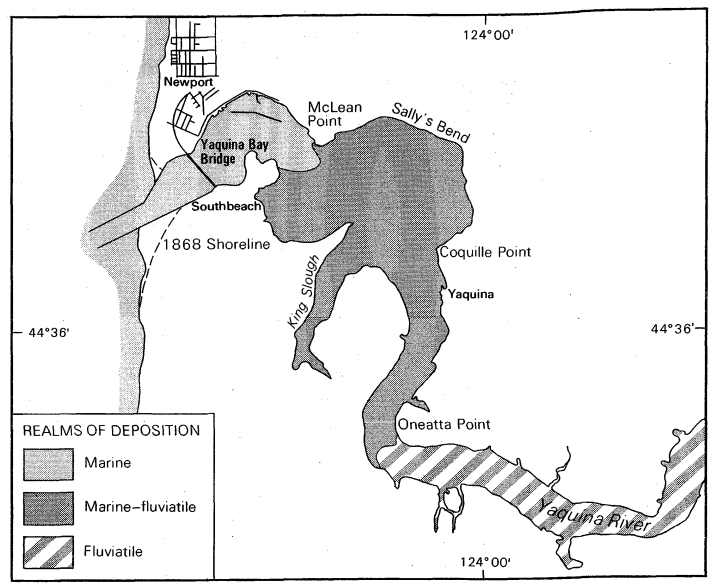


Figure 5: Sediment types and realms of deposition in Yaquina Bay, Oregon.

Adapted from L. Kulm and J. Byrne, "Sediments of Yaquina Bay, Oregon," *Estuaries*, pub. no. 83; © 1967 by the American Association for the Advancement of Science

LIFE IN ESTUARIES

The estuary as a biological environment. Estuarine organisms must cope with the most variable environment on Earth. While other environments may be more extreme in terms of heat or cold, or even salinity, none changes as rapidly, or in so many ways, as an estuary. The salinity at any one point may vary from that of fresh water to that of seawater during the course of a single tidal cycle. At one moment a tidal bore may rush into the estuary; a few hours later the water may be so calm that fine particles of mud can settle. Where the tidal range is great, large areas exposed at low tide may be subjected to the full effects of the sun at midday and to frosts at night.

Environmental gradients

Within an estuary are two major environmental gradients that result from periodic tidal flushing. The first is a longitudinal gradient from fresh water to the sea, the major variable being salinity, subsidiary variables being wave action and current speed. The second is a transverse gradient, which is best shown in estuaries with a large tidal range. The transverse gradient is more complex than the longitudinal: salinity, currents, exposure, temperature, and deposits all vary along the transverse gradient. Low on the estuarine shore the salinity is more variable than it is higher up. When the tide comes in over the lower part of the shore, the outflow of the river is first pushed back, rising slightly over the shore, and is then replaced by water of higher salinity as the sea penetrates into the estuary. High on the shore the salinity is less variable, as it is covered only for a short period, either by fairly high salinities, if the estuary is not stratified, or by fairly consistent low salinity, if the estuary is of the type in which a layer of fresh water flows out over the top of a penetrating tongue of seawater. The lower parts of the shore are also subjected to stronger currents, and this in turn influences the fineness of the deposits. The smaller the particles, the longer the time required for them to settle. Coarse sand settles within a few seconds, but fine mud may take several hours to settle. The finest muds are usually deposited fairly high on the shore, where the currents are slowest.

Habitats of plants and animals

The two gradients, longitudinal and transverse, produce a variety of habitats within the estuary. Each habitat has its own characteristic plants and animals. Low on the shore, where there is bare sand or mud, the majority of animals are burrowers. The bare surface is too exposed for most animals and not stable enough for the attachment of seaweeds, so that there is no shelter from predators. The only escape is to burrow beneath the surface. Various worms, mollusks, and crustaceans make burrows differing in shape and complexity. For instance

the lugworm (*Arenicola marina*) makes a U-shaped burrow with an open tail shaft, but the head shaft is filled with sand that slowly sinks down as the worm eats it from below for suspended food. This sinking of the sand down the head shaft leaves a shallow conical depression in the surface of the sand. At intervals the worm moves backward up the tail shaft and defecates onto the sand above, leaving the sandy casts that are often abundant on estuarine beaches.

For many estuarine animals the burrow is a part of the feeding mechanism. A current of water is passed through the burrow, and small particles in suspension are trapped and eaten, often by elaborate mechanisms, which may involve hairy fringes and mucus nets. The current of water flowing through the burrow also serves for respiration and carries away stagnant water.

The bivalve mollusks use tubular siphons to make contact with the water above the mud. *Scrobicularia plana*, for example, lives several inches below the mud surface. Its two siphons can be extended to ten inches or can be retracted into the shell if danger threatens. This is basically the same as living in a U-shaped burrow, but instead the body of the animal has become modified to enable water to flow through and supply food and oxygen.

Filamentous algae and various unicellular forms, such as *Euglena* and diatoms, are found on bare mud and sand. The filamentous algae help to stabilize the surface and provide food for a number of microscopic animals. Both *Euglena* and the diatoms move up to the surface when uncovered by tides, accounting for the green and yellow patches on bared estuarine muds.

Wherever rocks or stones provide shelter, the estuarine fauna changes. Barnacles can withstand quite low salinities but need something solid to which they can attach themselves. Some of the amphipod crustaceans spend a large part of their time hiding under stones. Mussels, also, are influenced in their distribution and abundance by the presence or absence of a sufficiently firm substratum for attachment. Once a few mussels become attached, they in turn provide surfaces for the attachment of other mussels and shelter for a whole host of smaller animals. The mussel attaches itself by secreting a bundle (byssus) of numerous tough threads that adhere to solid objects. Oyster reef communities also provide a stable bed upon which other organisms can attach.

Colonizing plant communities

The upper parts of the estuarine shore often support rooted vegetation, the most characteristic form in temperate regions being the salt marsh. This flourishes best above the high watermark of neap tides, so that the plants of the marsh have periods of several days every two weeks when they are not covered by water.

Several different plants may act as primary colonizers on estuarine shores. Contrasting examples are *Spartina* and *Salicornia*. *Spartina* is a tall grass with a creeping rootstock from which a series of stout, vertical roots descend and another series of horizontal roots spreads out into the surrounding mud. A rooting system of this sort is very effective in binding soft mud. Once the plant has become established, its dense tussocks slow water currents, allowing the settlement of more mud. Areas colonized by *Spartina* are thus rapidly built up above surrounding areas. Where the water is fresher, *Spartina* tends to be replaced by tall reeds, such as *Phragmites*.

Salicornia, a small plant compared with *Spartina*, has no distinct leaves but only swollen leaf bases around the stem, and it has shallow roots. It is a rapid colonizer of regions with slow currents.

Behind the primary colonizers a range of plants may form a dense sward or general salt marsh, with its characteristic animal life, including both terrestrial and aquatic forms. The latter must tolerate periodic exposure to air, while the former must either leave the marsh at high tide or tolerate submergence for several hours at a time. In the tropics the salt marsh is replaced by mangrove communities, in which the most important trees are the red mangrove (*Rhizophora*) and the black mangrove (*Avicennia*), whose tangled roots and branches provide shelter for numerous animals.

The open waters of estuaries have their own characteristic animals. Floating animals (zooplankton) are usually restricted to relatively few species but can be quite abundant and widespread. Some copepods, such as *Eurytemora velox* and *Acartia tonsa*, are found in estuaries over a large part of the Northern Hemisphere. Sometimes one species replaces another seasonally. For instance, *Acartia tonsa* tends to be abundant in summer, while *Acartia clausi* is generally more abundant in winter; such differences can be related to the interaction of temperature and low salinity on the survival and reproduction of the two species.

The mysids, or possum shrimps, can sometimes be seen as a dense band about a yard wide at the edge of the water as it rises over estuarine mudbanks. They feed by filtering out small particles from the water and in turn act as food for a variety of fish.

Adjustments to estuarine living. An animal living in an estuary must reproduce in such a way that its young can remain in the estuary. An excellent example is given by the ragworm (*Nereis diversicolor*), which belongs to a group of worms that normally produce larvae that float in the plankton. Such a mode of life would be hazardous in an estuary: the larvae would get washed out to sea, away from the habitat to which they are best adapted. The ragworm adapts by producing larvae that do not float but remain on the bottom among the surface layers of the mud.

A high proportion of estuarine animals are crustaceans of the orders Amphipoda and Isopoda, which do not lay their eggs freely into the water but carry them in pouches on the undersides of the females. The young hatch as miniatures of their parents and so avoid the dangers of the floating planktonic life. The burrowers, such as *Corophium*, often liberate their young inside the burrow, and the young then form their own subsidiary burrows.

Adaptation to the stress of salinity

Variation in salinity imposes the greatest stress upon the organisms living in estuaries. A marine animal normally has blood as salty as seawater. If such an animal is placed in fresh water, its tissues take up water by osmosis and the animal swells and dies. Estuarine animals have developed means of controlling such effects and are able to withstand sudden changes in salinity. Some animals develop a tolerance to changes in the salt concentration of their blood. This method is used by the sea slater (*Ligia oceanica*), which can, if necessary, allow the salt concentration of its blood to double.

In general, however, it is more efficient for an aquatic animal to develop a less permeable skin and thus reduce the stress caused by variation in external salinity. The general principles can be illustrated by a consideration of the Chinese woolly-handed crab (*Eriocheir sinensis*), which has been introduced into Europe and is now widespread in many of the larger rivers. This species is capable of living in both fresh water and seawater. When in seawater, the blood has the same osmotic pressure as the surrounding water. When in fresh water, the concentration of the blood is about half that of seawater but still well above that of fresh water. The crab regulates the amount of salt in its blood by taking up salts in its gills from the small amounts present in fresh water. *Eriocheir* is one of the most impermeable crabs. When it is in fresh water, the osmotic inrush of water is severely restricted, and it produces about the same amount of urine in both fresh water and seawater. The common shore crab (*Carcinus maenas*), however, is somewhat more permeable. When in seawater, it produces urine at a rate of 4 percent of its body weight per day. If *Carcinus* is transferred to a mixture of half seawater and half fresh water, the rate of urine production increases up to about 20 percent of body weight per day. This increase is adequate to cope with the osmotic inflow of water that would otherwise dilute the blood.

Biological productivity. Although the estuarine environment is variable and difficult, it is also very productive, for several reasons. One, the river constantly brings down fresh supplies of nutrient salts, especially by pollution from towns built near the head of the estuary. The constant, repeated invasion by the sea also improves the nutrient situation. The basic source of production in estu-

aries lies in the salt marshes. The amount of material that can be harvested from a *Spartina* marsh, for example, is considerably greater than that produced by an average rice field, three times that of the best hayfields, and 50 times that of short grassland areas with low rainfall.

The food-supply chain

The economy of an estuary depends on the major supply of food from the salt marsh. Organic debris from this source supports the growth of bacteria, and the combination of debris and bacteria supports huge populations of invertebrates, which in turn are preyed upon by migratory populations of fish and birds. In most estuaries the production of floating plants (phytoplankton) in the open water is not very important because the water is often too turbid with silt to allow for photosynthesis more than a short distance into the water. Plants growing in the marsh, however, are covered by the water for a few hours at most each day and may remain uncovered for several days at neap tides.

The salt-marsh plants eventually die and decay, the bits of debris being dispersed over the estuary. Certain bacteria aid in the decomposition and add to the nitrogen content of the plant debris. This combination of debris and bacteria forms the basis of the estuarine food web. Nearly all the invertebrates that occur in large numbers in estuaries feed on fine debris and its associated bacteria collected either from suspension in the water or from deposits on the surface of the mud.

A single square metre of estuarine mud may contain a thousand burrowing bivalves and several thousand ragworms. These immense populations of invertebrates provide food for large numbers of fish and birds.

Many fishes move into the estuary on the high tide, feed on invertebrates that live in the mud, and then retreat on the ebb tide. Some estuarine fishes eat plants, and a surprising number include organic debris in their diet. Gray mullets (*Mugil cephalus*) in particular are great consumers of mud, from which they strain out edible material. They also browse on algae growing on stones or on mangrove trees in the tropics. Some of the smaller fishes, such as the gobies, are permanent residents in estuaries, living in creeks and salt-marsh pools where they feed on the rich invertebrate life.

Estuarine shores provide ideal feeding grounds for vast numbers of birds. The rich supply of invertebrates is combined with open expanses that give no cover for bird predators. The most characteristic estuarine birds are the waders, which often occur in flocks of thousands and feed particularly on the fauna of sand and mud, in some places competing with man. Pelicans, cormorants, herons, egrets, mergansers, and kingfishers subsist mainly by fishing. Geese and ducks show a variety of feeding habits. Some of the geese are almost entirely plant eaters, while some of the ducks feed largely on invertebrates. The shelduck (*Tadorna tadorna*), for instance, includes a high proportion of the small snail *Hydrobia ulvae* in its diet.

In a quantitative study of the amounts of energy involved in the various links of the estuarine food web, it was found that *Spartina* produced a net amount of 6,580 kilocalories per square metre per year. The most important consumption of this accumulated energy was through the decomposing activities of bacteria, which accounted for 3,890 kilocalories per square metre per year. Of the total energy accumulated in the salt marsh by the growing plants, about 55 percent was consumed by organisms within the marsh, leaving 45 percent available for utilization in the estuary itself. This underlines the importance of salt marshes in the economy of estuaries, particularly those that are turbid and poor in phytoplankton. Any measure that reduces the area of adjoining salt marsh will reduce the productivity of an estuary.

(J.Gre.)

MAN AND ESTUARIES

Land reclamation. The Netherlands has always been a region faced with a shortage of agricultural land and an expanding population. The rise of sea level since the last glaciation, combined with a gradual sinking of the general area, has produced broad, shallow estuarine waters near or slightly below sea level along much of the coast.

In the Middle Ages land reclamation consisted primarily of marsh drainage, but as population increased land became scarce. Dikes were erected to push back the seas, and former tidal areas were reclaimed chiefly as pastureland and farmland. These broad, flat zones are known as polders and are usually crisscrossed by the canals for which the Dutch are so famous.

Creation of the IJsselmeer

The largest land reclamation project was the conversion of the Zuiderzee into the freshwater lake IJsselmeer by the construction of a great barrier dam between the Wadden Zee and the IJsselmeer. The barrier dam was completed in 1932 and isolated approximately 2,700 square kilometres of shallow estuarine waters. Almost at once the salinity of the former estuary began to drop to that of fresh water. Of course, with the drop in salinity many estuarine species of organisms were no longer able to survive. This included herring, anchovies, shrimp, edible mussels, soft-shelled clams, some crabs, etc. Although the former estuarine fishing industry was eliminated, new species of organisms, including pike, perch, bream, and roach, took over the IJsselmeer environment. In addition, a few estuarine species survived; the eels, for example, form one of the principal commercial species today.

The creation of the IJsselmeer had several advantages. The lake provides the northern Netherlands with a supply of fresh water that is important during dry periods. The land that borders the IJsselmeer is no longer in danger of flooding as when it was open to the Wadden Zee. Creating land from a freshwater lake is much easier than from a former tidal region, where the land must be desalted before it becomes completely productive. Finally, although the net fisheries production decreased by approximately 9,000,000 kilograms when the IJsselmeer was created, over 70,000,000 kilograms of pork was raised on the land reclaimed from the former estuarine environment.

The Netherlands is also studying the possible conversion of other estuarine areas into freshwater environments; plans include the damming of portions of the Haring Vliet, Grevelingen, and Ooster Schelde estuaries and their eventual conversion to fresh water.

Most estuaries bordering population centres have been used in minor land reclamation projects. Some sanitary landfill projects have eliminated vast acreage of wetland estuaries and at the same time created valuable coastal real estate. Boca Ciega Bay, on the west coast of Florida, is an example of an estuary where extreme dredging and filling has reclaimed land to such an extent that scientists are afraid that the estuary may no longer be a suitable breeding and nursery ground for the fisheries population of the region. The Florida region has long been used for reclamation sites primarily for real estate developments in coastal regions. Of Florida's 796,000 acres of estuarine water less than six feet deep, some 60,000 of these valuable acres of nursery and breeding areas have been lost to landfill and dredging.

Estuaries as recreation sites. As technology increases, more and more leisure time will be available to man. Estuaries will provide sites for two major forms of recreation: boating and sport fishing. Many boaters prefer the estuary for both sailing and motorboating primarily because the semi-enclosed basin usually has relatively calmer seas than the open ocean. Puget Sound and Chesapeake Bay in the United States, both major estuarine systems, may support the world's largest flotillas of small pleasure boats.

Estuarine fishing is also an important form of recreation. The great abundance of food sources for fish in estuarine waters make them relatively productive fishing grounds for the recreational angler.

Because of the great quantity of suspended matter in estuaries and the abundance of soft mud near shore, estuaries are less important as swimming areas; however, where high energy environments exist in the estuary and where artificial beaches have been developed, swimming is an important pastime.

Ebb and flow in estuaries

Tidal power. The first record of tide mills is from England and Wales around the 11th century. These early mills were crude affairs consisting of a floodgate that let water in on the rising tide but closed on the ebbing tide. The water then trapped in the tidal basin was forced to drive a waterwheel, usually an undershot type. Today, the tidal flow in estuaries can be used as a clean, cool source of hydroelectric power. In November 1966, the world's first major hydroelectric plant that uses tidal energy was opened on the Rance River Estuary in northern France. This plant can produce over 500,000 kilowatts of electrical power.

The Rance River Estuary was chosen as a site primarily because of the high tidal range (nine to 14 metres) that exists in this region. The power is generated both when the tide is flooding and ebbing through the water turbines. Although the initial cost and operating expenses are somewhat greater for a tidal power plant, in the long run the cost disadvantage is outweighed by low environmental pollution and the cheap power source.

Tidal power sites, however, are limited to those regions with a high tidal range and relatively close proximity to the consumer. In Great Britain the Severn River Estuary has been studied for a possible power station. In North America, the Passamaquoddy Bay region of Maine seems the most likely spot for a tidal hydroelectric plant. In the world, there are nearly 100 estuarine sites that could be harnessed for tidal hydroelectric power.

Estuarine pollution. Many of the world's largest metropolitan areas border estuarine environments; *e.g.*, Buenos Aires, Osaka, Los Angeles, Tokyo, London, New York, and Shanghai. Total population of these seven areas alone approaches 90,000,000 people. These cities are highly industrialized, and all make use of the estuary as a waste disposal site to a greater or lesser extent.

There are many ways man can pollute estuaries. Some are subtle, with no apparent short-term effects, while others are immediately disastrous with large wildlife kills.

Siltation, the natural enemy of harbours and estuaries, is a form of the subtle pollutants. Each year rivers wash more and more of the tilled soil away from the continents, and estuaries become filled with these silts and clays. In some cases, construction and farming have so increased the estuarine sedimentation rates that natural estuarine biologic populations are "smothered" by layers of silt and clay.

Thermal and inorganic pollutants

Thermal pollution can change the ecology of the estuary. Both fossil fuel and nuclear power plants require vast amounts of cooling water in their power output. In the United States alone it is estimated that some 200 × 10^9 gallons of water per day will be required by the electric generating plants. Many of these plants will be constructed on estuaries, where a large steady supply of water is available.

Pesticides that man has been using to control insect population both on inland crops and around marshes will eventually wash off the land into the rivers, then into the estuaries, and eventually into the ocean. Human and animal wastes are another principal form of estuarine pollutants. The high organic content of wastes can use up oxygen in estuarine waters as the wastes decay and thus cause the estuary to become anoxic and essentially devoid of life.

The list of pollutants in an estuary is as varied as the industries that surround it. Heavy metal concentrations, especially zinc, copper, and lead, have been shown to be particularly harmful to some estuarine organisms in high concentrations. Oysters and clams have been studied that can concentrate zinc in their systems over 100,000 times that found in the natural water.

Although estuarine pollution is one of man's more serious problems, there is an indication that he is now aware of the problem and taking steps to combat it. Effects of pollution can be reversed. San Diego Bay, off the coast of California, was one of the most polluted estuaries on the western coast of the United States. The waters were a brownish gray and practically devoid of life. Coliform bacteria counts were high, and the waters were considered a health hazard to bathers. After a decade of concentrated efforts by state and local officials to eliminate raw sewage and discharge effluents into the area, San Diego Bay's waters are again blue, and fish and other forms of wildlife are beginning to return. As man becomes increasingly aware of this environment, he will better understand the role of estuaries in coastal ecology. With this understanding he will be able to preserve this unique and geologically ephemeral environment. (F.E.A.)

BIBLIOGRAPHY. A. GUILCHER, *Coastal and Submarine Morphology* (1958), a useful summary of the agents responsible for transporting sediments in the estuary and forming its geomorphology; J. WILLIAMS *Oceanography*, ch. 13 (1962), an excellent description of the hydrology of estuarine types and a discussion of the factors causing the estuary to change from one type to another; G.L. PICKARD, *Descriptive Physical Oceanography*, ch. 8 (1963), on coastal oceanography and estuarine processes, with figures of typical salinity–depth profiles and longitudinal salinity sections that make estuarine hydrology easy to understand; G.H. LAUFF (ed.), *Estuaries* (1967), a collection of papers from a conference on estuaries that contains the best and most up-to-date information available; however, many of the papers are technical and would be difficult to understand without a background in the subject.

Estuarine life: J. GREEN, *The Biology of Estuarine Animals* (1968), a general introduction to estuarine life, with particular emphasis on animals; R.M. INGLE, "The Life of an Estuary," *Scient. Am.* 190:64–68 (1954).

(F.E.A./J.Gre.)

Ethics

Ethics is the branch of philosophy that is concerned with what is morally good and bad, right and wrong; a synonym for it is moral philosophy. Traditionally, ethics has undertaken to analyze, evaluate, and develop normative moral criteria for dealing with moral problems. In recent years, however, pervasive current in Anglo-American ethics has come under the influence of a new conception of the proper methods and capacities of philosophy in general, viz., that its task is the analysis of language; and on this basis it has come to hold that the concern of the moral philosopher should be restricted to the analytic task, or to "meta-ethics." This task consists in the logical analysis of the meanings of moral concepts and of the methods of supporting moral judgments and thus stands in contrast to the more traditional view, which combined such analytic studies with normative ethics, the view that authentic standards exist. The relation between these two conceptions, to be discussed more fully following the outline below, is an example of the controversies that have always divided moral philosophers on a large variety of basic issues arising not only from within their subject matter but also from opposed views about their very study of it.

The article is divided into the following sections:

I. Nature, types, and relations of ethics
  Types reflecting intra-ethical differences
    Meta-ethics and normative ethics
    Different analyses of the concept of morality
  Types reflecting relations to other disciplines
    Relations to the philosophical disciplines
    Relations to the humanistic and social disciplines
II. Meta-ethical theories
  Intuitionistic cognitivism
    Moore's axiological intuitionism
    Prichard's deontological intuitionism
    Ross's distinction between prima facie and actual duties
    Critical evaluation
  Naturalistic cognitivism
    Ancient Greek and medieval naturalisms
    Varieties of contemporary naturalism
    Ontological difficulties of naturalism
    Valuational difficulties of naturalism
  Noncognitivism
    Concern with language and verifiability
    Moral statements as "commending"
    Moral statements as "commanding"
    Function of moral judgments
    Cognitive undecidability of moral issues
III. Normative ethical theories
  Principal issues of normative ethics
    Three traditional questions
    The central question

Deontological theories of normative ethics
Material deontological ethics
Formal deontological ethics
Teleological theories of normative ethics
General considerations
Egoism
Utilitarianism
IV. Good reasons theories

I. Nature, types, and relations of ethics

Meaning of ethics

The word ethics signifies not only a certain branch of philosophic study but also the object or subject matter of that study. "Ethics" and "ethical" are often used synonymously with "morals" (or "morality") and "moral," as when reference is made indifferently to the ethics or to the morality of a person or a group, to their ethical or moral virtues or qualities. This synonymy has a clear etymological basis, for the Greek word *ēthos*, from which "ethics" is derived, and the Latin word *mores*, from which "morals" is derived, both mean habits or customs. The ethics or morality of persons or groups, however, consists not merely in what they habitually or customarily do but in what they think it is fitting, right, or obligatory to do. Men's actions are often, but not always, a sign of what they believe: their actions may diverge from their beliefs, and both actions and beliefs may differ from what men say they ought to do or believe. Morality contains an ineluctable normative element. Whereas a person may engage in habitual and customary conduct without any reflective thought, ethics always involves reflective evaluation or prescription concerning the conduct in question. Even when "customary morality" is spoken of, the reference of the term is not merely to the customs as such—in the sense of regular, repeated sequences of behaviour—but also to the view, at least implicitly held by the participants, that what they regularly do is in some way right; it is not merely what is done, it is also what is to be done.

The central concern of moral philosophy, then, has always been the double task (1) of meta-ethics, of analyzing the meaning and nature of this normative moral element in man's action, thought, and language, and (2) of normative ethics, of evaluating this element by presenting and appraising criteria for justifying rules and judgments of what is morally right and wrong, good and bad.

Logical relation between normative and meta-ethics

Because in meta-ethics the logic of normative ethics in its various phases is studied, these two parts of moral philosophy are intimately connected. Contemporary philosophers have disagreed, however, about the precise logical nature of this connection—particularly on whether meta-ethics is or can be neutral with respect to normative ethics. This question, in turn, raises two subquestions: whether holding a specific meta-ethical doctrine logically commits one to holding a specific normative ethical doctrine, and vice versa.

Some philosophers have given negative answers to these subquestions, pointing out that various sorts of combinations have, in fact, occurred among moral philosophers: some intuitionists (stressing the direct grasp of moral qualities) in meta-ethics have been teleologists (stressing goals) in normative ethics, others have been deontologists (stressing obligations), and so forth. These philosophers have also contended that there is a difference in logical type between meta-ethical and normative ethical statements, that the former are logical and conceptual analyses of the latter, and that, hence, the meta-ethical statements do not, as such, make any normative moral assertions.

Other philosophers have given affirmative answers, pointing out that the meta-ethical doctrines of cognitivists —those who affirm that ethical statements are a kind of knowledge—are concerned to bring out just those meanings of "good" or "ought," and just those methods of justifying moral judgments, that fit their normative ethical doctrines. Hence, there is a direct logical sequence from their normative ethical doctrines to the meta-ethical doctrines derived from the analysis of their normative doctrines. Proponents of an affirmative answer have also contended that, although the noncognitivists, qua philosophers, present no normative ethical doctrines, their noncognitivist meta-ethical positions sometimes have logical implications for normative ethical judgments concerning the psychological or other conditions required for moral goodness.

Although this issue cannot be gone into further here, it can be argued that each side may be partly right and partly wrong. The negative answer is at its strongest when the most abstract possibilities of the relations between the two fields are at issue, and the affirmative is strongest when specific philosophers or doctrines are being considered.

Moral philosophers have pursued this double task of meta-ethics and normative ethics in a large variety of ways and with a large variety of emphases, a variety that has two main dimensions—one contained within ethics itself, the other deriving from the relation of ethics to other branches of philosophy, to science, or to human experience in general. Although these two dimensions are in some respects distinct, the question of the precise or proper relation between them itself reflects important divergences among moral philosophers.

TYPES REFLECTING INTRA-ETHICAL DIFFERENCES

Cognitivists versus noncognitivists

Meta-ethics and normative ethics. The first, the intra-ethical dimension consists primarily of the various doctrines that have been upheld in both meta-ethics and normative ethics. In meta-ethics, philosophers have differed over the logical status of moral concepts and judgments. Some philosophers, the naturalistic descriptivists, have held that moral concepts describe natural or supernatural entities in the world. Others, the nonnaturalistic descriptivists, while agreeing that moral concepts are descriptive of entities in the world, have held that these entities are entirely unique in kind. Still other philosophers have been nondescriptivists, holding that moral concepts function primarily not to describe anything but rather to express attitudes or emotions, to command or prescribe, or to commend or condemn. These various groups of philosophers have held corresponding views about the logical status of moral judgments: either the judgments, as in cognitivism, are capable of being true and false and of constituting a kind of knowledge, natural or nonnatural, or, as in noncognitivism, they are not capable of being true or false or of constituting knowledge but function rather to express attitudes or give commands. There have also been differences over whether moral judgments are objective or subjective, absolute or relative—concepts the various meanings of which are best developed in contexts that allow elucidation with greater precision.

Deontologists versus teleologists

In normative ethics the main division has been between deontologists and teleologists, who differ on what constitutes the primary criteria by which actions are to be evaluated as morally right and wrong, and the ground of their difference derives from what they take to be the basic concepts of moral discourse and justification. The deontologists (Greek: *deontos*, "of the obligatory") take as basic such concepts as right and ought, and they hold that other moral concepts are definable in terms of these concepts or, at least, that moral judgments using other moral predicates are to be justified by means of judgments based on these deontic concepts. The teleologists (Greek: *teleios*, "brought to its end or purpose"), on the other hand, take as basic such axiological concepts (Greek: *axios*, "worthy" or "valuable") as good and value; they hold that the deontic concepts are to be defined in terms of these concepts and that moral judgments are to be justified by reference to the goodness of the purposes or consequences achieved by the actions. This difference largely determines their views of what makes certain kinds of actions morally right. Deontologists typically hold that certain kinds of actions are inherently right or right as a matter of principle because of their being the kinds of actions that they are or because of their conforming to some formal principle. Teleologists, on the other hand, typically hold that actions are right because of the goodness of their consequences. Another group of philosophers, related to but distinct from teleologists, are called axiologists; they hold that certain actions

are right because of the value or goodness that they inherently contain, such as happiness or pleasure, not merely because of the goodness of their consequences. Both the axiologists and the teleologists take goodness and value as the basic moral concepts, but the axiologists emphasize intrinsic goodness itself, whereas the teleologists emphasize the calculation of consequences in order to ascertain where, among various alternative actions, the preponderance of intrinsic goodness lies.

Different analyses of the concept of morality. In addition to these differing views about the logical status of moral judgments and the criteria for justifying them, philosophers have been concerned to analyze the concept of morality itself (in the sense in which moral is contrasted with nonmoral rather than with immoral). Thus, they ask how moral rules and judgments differ from nonmoral ones and what the difference is between something's being morally good or right and its being good or right in some other respect, such as prudentially, legally, aesthetically, intellectually, or religiously. And social scientists sometimes ask how moral feeling or how the institution of morality differs from other feelings of persons or from other institutions of society. Insofar as the question concerns the difference between moral and nonmoral rules, judgments, and criteria, however, it falls within ethics. Though the question seems in many respects to belong to meta-ethics rather than to normative ethics, some of the main considerations in answering it have derived from the normative distinction between deontological and teleological criteria.

Formal versus material conditions

The main division has been between philosophers who hold that, for a judgment, rule, or consideration to be a moral one, it must fulfill only certain formal conditions and those who hold that it must fulfill a certain material condition. The formal conditions that have been proposed include that the proponents of the judgments or rules must regard them as prescriptive, as universalizable, or as having overriding importance or authority; the material condition is that the rules, judgments, or considerations must bear on the interests or welfare either of society as a whole or at least of persons other than the judge or agent. It can be seen that the upholders of the formal conditions are closer to the deontologists, because they emphasize that what makes them moral are certain features of the judgments themselves or of the judges' views concerning them. The upholders of the material condition, on the other hand, are closer to the teleologists and axiologists, because they emphasize the bearing of the judgments on basic human goods or purposes.

The impact of the distinction between deontological and teleological criteria on differing interpretations of the concept of morality can also be seen in a related distinction that goes back to the ancients—that morality is constituted either by laws, rules, or commands or by the pursuit of human happiness or well-being. In the former, deontological view—found in the Ten Commandments and in Kant's categorical imperative—morality is regarded primarily as controlling and even repressing natural human tendencies. In the latter, teleological view—found in Aristotle and in the Utilitarians, beginning at least with David Hume, an 18th-century Scottish Skeptic—morality is regarded primarily as developing and fulfilling natural human tendencies. These two views were variously combined in Plato and in the New Testament (and also, in one form, in Kant); they held (like the law and rule view) that certain rules are unconditionally obligatory, even though they may lead to this-worldly disaster, but (like the happiness or well-being view) that obedience to the rules will lead to well-being in a future life.

What virtues or values are moral

One upshot of this difference between definitions of morality emerges in divergent conceptions of what virtues and other goods or values are moral (as against nonmoral) ones. Kant declared the deontological view that only goodness of will consisting in obedience to duty is "good without qualification," as against the various components of "happiness," consisting in "talents of the mind" and "gifts of nature or fortune," which may all become morally bad in the absence of a good will. On the other hand, Aristotle, in upholding the axiological view, listed as "moral" virtues many traits of character that, from at least one deontological point of view, would not count as moral virtues at all: *e.g.*, cheerfulness, wittiness, and "magnificence." These traits are moral virtues for Aristotle because they contribute to individual or social happiness. Belonging to this same tradition, Hume insisted that to try to distinguish moral "virtues" from mere "talents" is to engage in a merely "verbal dispute," because all of these form part of "personal merit." On this ground Hume, too, listed as moral virtues a large number of qualities that strike the contemporary reader, influenced by the deontological-law view, as dubiously moral, such as pride, tranquility, and wealth.

The disagreement between these two views of morality is both conceptual and normatively moral. It concerns the conceptual question of what is to count as moral (in opposition to nonmoral). But this distinction is itself of normative moral significance, for, when certain qualities are regarded as moral, they are viewed as being of greater importance for human life than if they were regarded as nonmoral.

TYPES REFLECTING RELATIONS TO OTHER DISCIPLINES

A further variety of doctrines arises, not within ethics itself but from the relations of ethics to other branches of thought and experience.

Relations to the philosophical disciplines. As a branch of philosophy, ethics has both influenced and been influenced by other parts of philosophy.

Linguistic Analysis, Existentialism, Phenomenology

Recent philosophical schools. In the 20th century, three such influences have been especially important. One has been the development of Linguistic Analysis, which has led to a new sensitivity to logical and conceptual issues in ethics. The methods of careful definition and clarification espoused by this school are traceable, however, back to the dialectical explorations presented in Plato's dialogues and have been utilized in many other eras of ethical inquiry. A second influence has been that of the emphasis placed by such Existentialists as the Danish Christian Søren Kierkegaard and the French atheist Jean-Paul Sartre on authenticity, autonomy, and decision. The Existentialist emphasis focusses primarily on the standpoint of the agent who is actually confronted with a moral problem, rather than, as with the Linguistic Analysts (until recently), on the standpoint of the critic or spectator who makes judgments about the agent or his actions. Although this focus is of the greatest importance, a serious difficulty is posed by the Existentialists' disavowal of general moral rules or principles. This denial of a rational basis of morality echoes the voluntarisms (will-centred views) of such late medieval philosophers as Duns Scotus and William of Ockham, as well as earlier skeptical positions. A third influence has been that of the Phenomenological approach of the Göttingen ontologist Nicolai Hartmann and the French philosopher Maurice Merleau-Ponty, which emphasized the concrete data of moral experience. Though in this position the focus of the Linguistic Analysts on moral judgments has been combined with that of the Existentialists on moral action, the main emphasis has been on the accurate description of man's preconceptual, pre-analytic involvements with morality. Antecedents of this approach can be found in Joseph Butler, an 18th-century bishop of Durham, in his description of the compelling authority of conscience.

Closely related to these philosophic influences on ethics has been the strong concern of philosophers, newly manifested in recent years, with the theory of human action. Ethics is largely concerned with how men ought to act, and this raises questions about the nature and limits of action. Insofar as moral judgments are addressed to persons, it is assumed that the persons addressed can in some sense act voluntarily or intentionally. Questions then arise as to just what such action consists of, how it differs from mere bodily movements and reflexes, and, in the light of the determinism that has often seemed to ground the natural sciences, how men can be held to act voluntarily. For, if determinism is universally true, then all of men's actions are the effects of causes outside themselves;

and what men take to be their autonomous reasons for their behaviour are then simply causes that move them in certain ways. Some of the ramifications of these questions will be mentioned later in relevant contexts.

Relation to metaphysics

Fundamental areas of philosophy. Many ethical doctrines bear a close relation to metaphysics, the study of the general nature of being or reality. The question of determinism just mentioned is an obvious example, deriving, as it does, from the impact of the nature of things on the possibilities of human action within that nature. In a more specific way, the various meta-ethical positions discussed above reflect metaphysical theses about what kinds of entities there are in the world. The naturalistic and nonnaturalistic cognitivists both assert that the world contains entities referred to by moral terms, whereas the noncognitivists deny such entities. Many philosophers, beginning with Plato, have held that moral criteria of what is good or right are to be based on subsistent metaphysical objects or relations, which range from Plato's Idea of the Good to Plotinus' One, the Stoics' right reason, Augustine's God, and Thomas Aquinas' nature and natural law. These last two in particular have had a long subsequent influence, and all of them raise meta-ethical issues (see below *Naturalistic cognitivism: Ancient Greek and medieval naturalisms*).

Moreover, a Materialistic metaphysics, like that of Epicurus, an ancient Greek moralist, and Thomas Hobbes, a 17th-century political and moral philosopher, has had ethical implications of two sorts. One is negative: the philosophers deny that the objective world of matter in motion has any place for moral goods; and they often tend also to deny, though sometimes with qualifications, that this world leaves any place for freedom of the will. The other implication is positive: moral concepts are held to express men's reactions to the impact of material particles on those parts of their sense organs that cause them to experience desires and aversions. Hence, good and evil in this second view are not in the world per se; they are rather men's internal and variable reactions to the ways in which they are stimulated by the world. As Hobbes put it:

> Whatsoever is the object of any man's appetite or desire, that is it which he for his part calleth *good;* and the object of his hate and aversion, *evil;* and of his contempt, *vile* and *inconsiderable.* For these words of good, evil, and contemptible are ever used with relation to the person that useth them: there being nothing simply and absolutely so; nor any common rule of good and evil, to be taken from the nature of the objects themselves.

Relation to logic and epistemology

Moral philosophy has also had a close relation both to logic and to epistemology (theory of knowledge). The meta-ethical enterprise consists in certain applications of these disciplines to moral concepts, judgments, and arguments. In addition to the direct applications (considered in detail below), there are other developments of both logic and epistemology that are closely relevant to ethics. One such development, work on which has greatly increased in recent years, is the elaboration of deontic logic and the logic of imperatives (see LOGIC, APPLIED: *Deontic logic* and *The logic of commands*). In deontic logic the philosopher studies the logical relations between the concepts of the obligatory, the prohibited, and the permitted and tries to set up calculi for inferring from certain axioms the formulas that predicate these concepts. In the logic of imperatives he tries to show how inference can proceed from premises containing at least one imperative sentence to an imperative conclusion; it is particularly relevant to those meta-ethical theories that hold that moral judgments either are or entail imperatives. In one respect, the lineage of these attempts goes back to Aristotle's work on the practical syllogism, whereby, from certain empirical and valuational statements, a conclusion is held to follow consisting in an appropriate action. Like other parts of logic, these theories have incurred various paradoxes.

Insofar as moral philosophy is a quest for ethical knowledge, whether conceptual or normative, it bears a close relation to the more general views elaborated by philosophers about the nature of genuine knowledge and the methods of attaining it. The specific bearings of this relation for cognitivist and noncognitivist theories in meta-ethics will be considered in detail below (II. *Meta-ethical theories*).

There have also been important historical influences. For example, the Cartesian doctrine of representative perception, which from the 17th century on has led philosophers to hold that men cannot directly perceive the external world, had strong repercussions for ethics. Some can be seen in Hume, for example, when he defined moral qualities not as existing or perceived in the world per se but rather as resulting from an impartial spectator's feelings of approval. Just as Hume held in his general epistemology that sense qualities cannot be known to exist in the absence of a perceiver, so he held in his moral epistemology that moral qualities require, for their very existence, the presence of a spectator. Without the spectator, men's characters and actions could not rationally be held to have any moral qualities at all. Many moral-sense theories in addition to Hume's have derived from this epistemological doctrine that qualities cannot rationally be held to exist unless they are in some way sensed or perceived.

Relations to the humanistic and social disciplines. Such humanistic disciplines as religion and history, as well as psychology and the social sciences, have significant interrelations with ethics.

Religion and history. Moral doctrines have been central to religion. All of the great religions prescribe obedience to the Golden Rule, and all set forth many other precepts about the way men ought to live and the kinds of characters they ought to have. The main difference between the religionists' prescriptions of ways of life and those given by moral philosophers—from Socrates through the Stoics and Epicureans to such more recent thinkers as Kierkegaard—consists not in the adherence of the religions to supernatural entities (lacking in Buddhism but present in Kierkegaard) but rather in the degree of conceptual and rational elaboration set forth by the philosophers. On the question of how religion bears upon moral criteria, the dominant view, presented in Plato's *Euthyphro* and in Kant, has been that reference to God is logically dispensable. The relation between God and moral rightness, according to these philosophers, is that God approves of certain actions because they themselves are morally right, not that they are morally right because God approves of them. Hence, the moral validation of the rightness or wrongness of actions is logically independent of any appeal to God. This view does not, however, deny that belief in God may be of practical moral significance, and it is a view, moreover, with which the medieval voluntarists mentioned above strongly disagreed. But the alternative that they upheld leaves moral criteria contingent on the arbitrary will of God.

God and morality

Insofar as philosophy is part of the culture of a civilization, its history belongs to the history of civilizations. In moral philosophy this historical dimension is especially prominent, because men's conceptions of moral character and right action have often reflected the dominant institutions of their society. This reflection has not always, however, been positive; it has sometimes consisted in reactions against severe problems faced by those institutions. In any case, it is possible to trace many correlations between the history of moral philosophy, especially on its normative side, and the history of society. In the ancient world of the West, this history ranged from the probing dialectic of Socrates questioning the unexamined beliefs of the Athenians through the aristocratic doctrines of Plato and Aristotle focussed on the values of small city-states, to the quietistic concern with moderate pleasures among the Epicureans faced with severe social instability, and to the Stoics' universalist emphasis on a universal law of right reason attributing a certain equality of dignity to all men, whether slaves, such as Epictetus, or emperors, such as Marcus Aurelius. The medieval concern with theological ethics, epitomized in Augustine, was followed by the rediscovery of Aristotle in the 13th century, which thus led Thomas Aquinas and others to reinstate a modest naturalism, as against the antecedent su-

pernaturalistic criteria based on God. In the modern world the developments from the Renaissance and Protestant Reformation to the Industrial Revolution and the various political revolutions have gone hand in hand with deontological and teleological emphases in normative ethics, which vary from egoism to Utilitarianism, individualism to Socialism, and elitism to egalitarianism.

Psychology and the social sciences. Psychology and the social sciences have had two main kinds of relations to ethics: explanatory and normative. In fulfilling their explanatory role, they have made empirical studies of moral phenomena in which they have pursued such questions as the nature and origins of men's moral beliefs and the ways in which such beliefs function in individual and social life. These studies have been concerned, among other things, with the differences and similarities between morality and other phases of human life and institutions; but they have focussed not primarily on the conceptual aspects of moral rules, judgments, or criteria as such but rather on their empirical genesis and effects. Among the familiar theories that have been elaborated in this sphere are the Darwinian view, continued by the anthropologists, that moral ideas serve to foster evolutionary or social adaptation; the Marxist doctrine that moral ideas are parts of the ideological superstructure that reflects and subserves the interests of the economically dominant class in any society; and the Freudian conception of moral ideas as the results of parental and societal conditioning aiming at the repression of vital impulses. Some scholars feel that theories such as the last two, though their lineage goes back to the ancient Greeks (as in the first two books of Plato's *Republic*), often view the moral ideas to which they refer in a rather restricted way. Both Marx and Freud, they argue, were so strongly attached to such moral ideas of their own as those urging the removal of exploitation and dishonesty, that they exempted their own moral ideas from the general diagnosis they had made of those of other persons.

Normative role of the social sciences

In their other, normative relation to ethics, the social sciences have tried in various ways either to set forth or to implement norms or criteria for moral actions, characters, and institutions. These norms have ranged from the psychologists' concern with mental health to the doctrines propounded by economists, political scientists, and legal theorists on the requirements and possibilities of justice in the economic, political, and legal spheres. In such endeavours the respective scholars are themselves engaged in moral discourse. Whereas the empirical findings on which they base their moral recommendations are often necessary conditions of morally right policies in their respective areas, they are not sufficient conditions in the absence of independent moral scrutiny. The logical or meta-ethical problems that confront such endeavours are not different in principle from those that face the metaphysical doctrines mentioned earlier. The ways in which moral philosophers have dealt with these problems will be considered in the subsequent sections of this article.

II. Meta-ethical theories

As was noted above, the cognitivist and the noncognitivist are two main varieties of meta-ethical doctrines on the meanings of ethical terms and judgments and the methods of supporting the judgments. The cognitivist doctrines affirm and the noncognitivists deny that moral terms signify qualities in the world and that moral judgments are a kind of knowledge.

Cognitivist theories fall into two groups: intuitionism (or nonnaturalism) and naturalism. The basic difference between them is that the intuitionists affirm and the naturalists deny that the objects that moral terms signify and the knowledge that moral judgments constitute are unique to ethics and hence irreducible to the objects and knowledge found, for example, in the natural sciences.

INTUITIONISTIC COGNITIVISM

According to the intuitionists, the basic moral term or terms signify a nonnatural, indefinable quality of things. G.E. Moore, one of the fathers of Linguistic Analysis, held that the basic term is "good"; two of his critics, the Oxford philosophers H.A. Prichard and W.D. Ross, held that it is "ought"; thus, their views on this point were, respectively, axiological and deontological. All three held that the meaning of the basic term is grasped by direct inspection; *i.e.*, the term's meaning consists in its denoting a certain objective property, that is itself grasped by immediate intuition.

Moore's axiological intuitionism. Moore distinguished three different meanings of the question "What is good?" One meaning is that the question asks: "What is the nature of the property goodness?" On this question Moore held that goodness is a simple, indefinable property, to be grasped by intuition.

Meanings of "What is good?"

A second meaning of the question is: "What things are good in themselves or intrinsically good?" This question, in turn, means: "To what things does the simple property good immediately or directly attach?" This question, too, Moore held, could be answered only by immediate intuition; for, because goodness is indefinable and unanalyzable, there is no way to answer the question by analyzing goodness to see whether it attaches to something. And, presumably, if one had to analyze the objects to see whether goodness attaches to them, then goodness would not attach to the object directly or to the whole of the object but only to a part of it.

The third meaning of the question "What is good?" is: "What things are instrumentally good or good as means to other things?"; *i.e.*, "What things are the causes of other things to which goodness attaches directly?" If, for example, personal affection, aesthetic enjoyment, and knowledge are good in themselves (as Moore held), then what can bring these goods into existence? Moore maintained that this third question is identical in meaning with the question: "What ought one to do?" Hence, he held that "ought" is definable in terms of "good": to say, "I ought to do x," is to say, "My doing x will bring about more intrinsic goodness—that is, the existence of more things to which goodness directly attaches—than will my doing anything else that it is open to me to do." Ought judgments, then, according to Moore, are capable of argument and proof, as against judgments of intrinsic goodness. On this point, he is an ideal Utilitarian: a Utilitarian insofar as he holds that one's duty is to bring about as much good as possible; ideal insofar as the goodness in question is quite general—as against, for example, views in which goodness is regarded as consisting in pleasure or happiness.

Indefinability of good. It can be seen that, of Moore's three questions, the first, on the nature of goodness, is the most basic, for the other two both depend on his answer to the first question: that the property goodness is simple, hence, indefinable and unanalyzable. This answer and the one main argument that he gave for it were of crucial importance for the subsequent development of meta-ethics.

The argument involved an appeal to what Moore called the naturalistic fallacy. He sometimes said that the naturalistic fallacy consists in confusing a nonnatural object, such as good, with a natural object, which must exist in space and time and be empirically observable. But he also insisted that, even if good were a natural object, any attempt to define it would still commit the naturalistic fallacy. The main point of this fallacy consists, according to Moore, in making two other confusions. One is to confuse a simple, indefinable object with a complex, analyzable object; since good is simple, any attempt to define it commits this fallacy. But how is it known that good is simple and hence indefinable? To prove that it is, Moore indicated still a third version of the naturalistic fallacy, according to which any attempt to define good confuses the "is" of identity (which figures in a definition) with the "is" of predication (which figures in a proposition asserting one thing of another). In other words, persons who claim to give definitions of "good" are confusing Moore's first and second questions, namely, "What is the nature of goodness?" and "What things are good?" The first question is to be answered by a definition, expressed in a sentence the copula (equational verb) of which is the "is" of identity; the second question is to be answered by

Naturalistic fallacy

a proposition, expressed in a sentence the copula of which is the "is" of predication. As Moore put it, when a person says that an orange is yellow, he is not saying that yellow is identical with orange or is the nature, meaning, or definition of orange; he is saying, rather, that yellowness is predicated of orange, that yellowness pertains to orange but is not identical with it. To give a definition of "good" as meaning, for example, pleasant or conducive to survival is to make this confusion between the "is" of identity and the "is" of predication; it is to assume that, because what is good is also pleasant or conducive to survival, therefore, goodness is identical with pleasure or conduciveness to survival.

Moore also held, however, that this confusion is logically inevitable no matter what definition one tries to give of "good"; the problem is thus not merely that goodness is not definable by some other concepts but rather that it is not definable by any other concepts. His argument consisted in the appeal to what he called the open-question test: no matter what concepts or qualities a man uses to try to define "good," one can always significantly ask, concerning those qualities, whether they are themselves good. Because it is thus an open question whether goodness attaches to them, those qualities cannot be identical with goodness nor constitute its definition. For, if they were thus identical, it would not be an open question whether those qualities were good. If, for example, good is defined as meaning pleasant, it can always be meaningfully asked whether what is pleasant is good, which is an open question. But, on the proposed definition, this question would mean the same as the question whether what is pleasant is pleasant; and this is not an open but, rather, a closed question: for there is no point or significance in asking it.

Open-question test

Stated somewhat more formally, Moore's argument is as follows: If a person considers any proposed definition of good (G) in terms of some property (P); that

(1) $G =_{df} P$,

in which $=_{df}$ means "is defined in terms of," then he can always ask, with significance,

(2) Is P G?

And this question does not mean the same as

(3) Is P P?

But question (2) would have to mean the same as question (3) if (1) were a genuine definition. As it is, (2) is an open question, whereas (3) is not. Hence, (1) is not a genuine or correct definition.

Bearing on the nature of definition and antinaturalism. This famous and influential argument raised many questions about the nature of meaning and definition. It assumed that every definition of some term must be completely intersubstitutable with that term, not only with no loss of truth but even with no loss of meaning, because the definition must mean the same as what it defines. This assumption ignores, however, an important aspect of definitions. According to Moore, when one term is predicated of another in a question, only two differences between the terms can cause that question to be open and not closed. One is a difference of extension: one of the terms does not denote or refer to the same entities or objects as does the other term. The second and more important difference is one of intension: one of the terms does not connote the same property or properties as does the other term. Moore held that the fact that an open question could always be asked about any proposed definition of good proves that there is always this difference of intension and that the definition is thus not a genuine one. For in a genuine definition the definiendum (to be defined) and the definiens (defining expression) must have the same intension as well as the same extension.

Extension and intension

There is, however, a third possible difference between terms, which Moore overlooked: a difference in clarity between the term defined and its definition. The term defined may, as such, be a vague general idea, and the definition then makes that idea more definite and precise, without thereby changing either its extension or its intension. In the definition of circle, for example, as "a plane figure each point on which is equidistant from a fixed point at the centre," there is no difference either of extension or of intension between the definition and the term that it defines, but there is a difference of clarity, since the definition articulates or explicates what was only vaguely understood in the original idea of the term circle. Because of this difference in clarity, the open question can be asked of this definition, "Is a circle a plane figure each point on which is equidistant from the centre?" But the fact that this question can be asked, with significance, about circles—the fact that this question does not mean the same as the question, "Is a plane figure each point on which is equidistant from the centre a plane figure each point on which is equidistant from the centre?"—does not at all prove that the predicate of the former question does not constitute a definition of its subject.

Moore assumed that all definitions must constitute or be reducible to explicit tautologies of the form "A is A." In such a tautology, however, the predicate simply repeats the subject without bringing in the conceptual analysis or clarification of the subject that a definition provides. It seems, then, that Moore's argument, if it were valid against proposed definitions of good, would rule out most, if not all, definitions.

Another important issue about Moore's contention that good is indefinable concerns its bearing on his antinaturalism; *i.e.*, on his doctrine that goodness is a nonnatural quality. Despite his efforts to establish the indefinability of good, there is no necessary logical connection between that contention and antinaturalism. The indefinability of some concept or quality does not mean that it is nonnatural, for, according to Moore, natural qualities, such as yellow and pleasant, are also simple and indefinable. Indeed, it might even be held that Moore's argument for the indefinability of good does not establish its simplicity. For good might have parts and yet be indefinable because the specific combination of the parts might be unique, so that to enumerate them would not constitute a definition of good; good might consist in the unique, indefinable way in which the parts are combined. In any case, Moore's antinaturalism is rather the result of a further interpretation that he makes concerning the indefinability of goodness: that intrinsic goodness is *sui generis*, uniquely different from all other concepts, and, hence, *a fortiori*, from all natural concepts. This interpretation is the basis of Moore's view that ethics is an autonomous study, different in kind from the natural sciences, and that to apprehend goodness requires a nonempirical intuition.

Prichard's deontological intuitionism. As Moore was an intuitionist with regard to the definition of good and to the question of what things are intrinsically good, so Prichard, an Oxford philosopher of the first third of the 20th century, was an intuitionist with regard to the apprehension of ought. Prichard's position is especially important because he tried to provide for the autonomy of normative ethics conceived not primarily axiologically or teleologically, as with Moore, but rather deontologically. He held that judgments about what one morally ought to do are not analyzable or provable in terms of any other considerations, including self-interest or general utility.

Underivative nature of obligation. In his most famous essay, "Does Moral Philosophy Rest on a Mistake?," Prichard answered his question affirmatively: the mistake consists in the attempts by moral philosophers to give an answer to the question "Is there really a reason why I should act in the ways in which hitherto I have thought I ought to act?" Such attempts are mistaken because the question's demand for a reason is improper. No reason can be given for a person's moral duty or obligation to perform some action. The reasons to which philosophers have appealed have all been some sort of goodness: either the selfish good of the agent or some general good consisting either in the consequences of the action or in the motive of the action. But none of these appeals to goodness can ground any obligation. One's selfish good rests on one's desires or inclinations, and these are obviously different from moral obligations, which operate rather to constrain inclinations. Nor can consequences for the general good provide the ground for moral obliga-

Mistaken quest for reasons

tions. The most basic reason Prichard gave for this denial is a cardinal statement of the irreducibility of deontological considerations to considerations of any other sort: "An 'ought,' if it is to be derived at all, can only be derived from another 'ought.' " This principle of nonderivability at once cuts the ground from under all attempts to justify obligations by means of any other considerations.

Another of Prichard's arguments against grounding the obligation to perform some action on its consequences for the general good also had important ramifications for the theory of justice. His argument was that the obligation to act justly is indifferent to and may even be inimical to good consequences. As for the intrinsic goodness of actions, consisting in the motive from which a person acts, this cannot be the ground for moral obligation, because a person may fulfill his moral obligation to perform some action even if he acts from bad motives. One's moral obligation is simply to do a certain act, not to do it from a certain motive.

The upshot of these considerations, according to Prichard, is that man's sense of obligation to do some action of a particular kind is "absolutely underivative or immediate." Certain preliminary inquiries may be needed in order to appreciate that an action is obligatory or right, but these entail simply getting clear about the precise nature of the action: one may need, for example, to trace out certain consequences of a proposed action in order to see that it will involve hurting someone's feelings. But once a person sees that it is an act of this kind, his perception of his obligation with respect to it is direct or immediate; *i.e.*, it is not mediated by any other consideration, such as goodness.

Objections and rejoinders. Prichard was aware of the apparent dogmatism of his position and tried to answer the resulting objections. To the criticism that obligations cannot be self-evident because men disagree as to what is obligatory, he replied that persons differ in their degree of moral development, that they sometimes fail to carry out the required preliminary inquiries because of lack of thoughtfulness, and that even the best men are therefore "blind" to many of their obligations. To the objection that obligations may conflict, so that one self-evidence will be in opposition to another, Prichard replied that obligation does indeed admit of degrees, so that one has to decide on which obligation is greater. Prichard did not, however, indicate how apart from intuition this decision should be made, although he insisted that, in any case, it is not a matter of calculating which alternative produces the greater amount of good.

Charge that the question of obligation is tautologous

One aspect of what Prichard took to be an improper question might seem obvious. If the question of whether there is any reason for acting the way one thinks he ought is interpreted as meaning, "Why ought I to do my duty?" or "Why should I carry out my obligation?," an obvious objection to it would be that the question is tautologous. Because "one's duty" or "one's obligation" means what one should or ought to do, to ask "Why ought I?" or "Why should I?" would be the same as asking "Why ought I to do what I ought to do?" and to ask this is like asking "Why is *A A*?" (which is a senseless or, at least, a pointless question). In this interpretation, Prichard's point would be analogous to Moore's: he would be saying that a certain question is mistaken because tautologous, just as Moore said that a certain purported definition is mistaken because it is not tautologous.

Two considerations must be noted in this connection. First, the basis of Prichard's own objection to the question that he regarded as improper was not that it is tautologous in the way just indicated. His version of the question was not, "Why are certain acts that are my duty my duty?" but rather, "Why are certain acts that I have believed to be my duty really my duty?" Prichard regarded this question as analogous to that raised by the classical theory of knowledge, "Why are certain things that I believe I know things that I really know?" For each question the point was to obtain some criterion, reason, or guarantee that what a person believes, he is fully justified in believing; that apparent or believed duty or knowledge is real or genuine duty or knowledge. And, in each instance, Prichard held that this quest for a justifying reason was in principle mistaken. But the mistake did not consist in asking a merely tautologous question; it consisted rather in asking for a reason for what was self-evident and, hence, admitted of no reason or of no justification.

"Ought only from ought" versus noncontradiction

A second, closely related consideration concerns an implication of Prichard's argument. Holding that the statement that one ought to do something is not provable or derivable from any statement about what is good, Prichard went on to argue that a statement about ought is derivable only from another statement about ought. But also holding that statements about ought are self-evident, he concluded that no derivation of them from any other statements is needed. This line of reasoning, however, overlooked at least one important alternative. Particular ought statements might be derivable from some general ought statement, some universal moral principle that would serve to ground and explicate all particular moral duties. This principle, in turn, would be derived not from some still more general moral principle but rather from certain principles of reason, involving especially the principle of noncontradiction. Kant's categorical imperative, "Act only on that maxim through which you can at the same time will that it should become a universal law," is the most famous example of such a universal moral principle; but there are other, perhaps more viable possibilities. The advantage of such an alternative over Prichard's position may be that it would show the rational justification of men's moral obligations in a way that avoids both his anti-Utilitarian strictures and his own dogmatic intuitionism. It might also clarify the problem of conflicting obligations, which he left unresolved.

Ross's distinction between prima facie and actual duties. Prichard's doctrines were carried on in amended form by another Oxford moral philosopher, W.D. Ross. The most important amendment was Ross's distinction between prima facie duties and actual duties. An action is a prima facie duty if it tends to be a duty but is not necessarily one's actual or full-blown duty, if it results from some one component of the total nature of the action—*e.g.*, that the action is an instance of keeping a promise or telling the truth. The whole nature of the action, however, may be such that to perform that action is not one's actual duty—*e.g.*, because telling the truth in the given situation will bring harm to some innocent persons. Whether some action is one's actual duty, then, depends on the total nature of the action; whether it is a prima facie duty depends only on some salient part of the total nature of the action.

Conflicts of duties and self-evidence

By this distinction between prima facie and actual duties, Ross was able to cope somewhat more adequately than Prichard with some of the difficulties confronting their common doctrine. One such difficulty concerns conflicts of duties: how they arise and how they are to be resolved. Ross's distinction explains how a certain action may be a duty and yet also not be a duty; the prima facie duty of keeping a promise may be defeasible (not an actual duty) because the action generates another prima facie duty opposed to the former duty, that of preventing harm to innocent persons. The question then arises as to how the conflict between these prima facie duties is to be resolved; *i.e.*, how in such a case a person is to determine what his actual duty is. Ross recognized that his position, like Prichard's, provides no principle that would resolve the conflict; but he also said that Utilitarianism is in no better position and that in any case there is no basis for anticipating that every duty is a duty for the same reason. Thus, Ross repeated Prichard's pluralism of duties.

Another difficulty on which Ross's distinction sheds some light is that of how duties can be self-evident and yet not be acknowledged by many persons. He held that prima facie duties are self-evident to all "thoughtful and well-educated people" who "have reached sufficient mental maturity and have given sufficient attention" to them. With respect to propositions expressing prima facie duties, such as that a person ought to keep his promises, to return services rendered, to effect a just distribution of good, or to promote the good of others, Ross wrote:

> The moral order expressed in these propositions is just as much part of the fundamental nature of the universe (and, we may add, of any possible universe in which there were moral agents at all) as is the spatial or numerical structure expressed in the axioms of geometry or arithmetic.

On the other hand, judgments of actual duty, because of the complexities they involve, are neither self-evident nor logical conclusions from what is self-evident.

Critical evaluation. The crucial difficulties of intuitionism are both epistemological and ontological (on the kind of being or existence). Epistemologically, the claim that men have intuitions of right or ought just as they have sensations of colours has seemed implausible to many scholars, both in itself and in its failure to account for conflicting intuitions. After all, moral conflicts are far more prevalent than disagreements about colours. To other philosophers, however, intuitionism has seemed to provide an indispensable basis for moral argument. They have held, for example, that to engage in moral argument with some person is impossible if that person does not even see the relevance of certain considerations, such as that the action would cause pain, to the desirability or undesirability of actions. For some philosophers, such considerations are part of the very definition of morality. In this way, intuitionist views continue to exert influence as bearing on conceptual or definitional issues of morality. There still remain problems, however, of conflicting intuitions and of rational principles for resolving the conflicts.

Ontological difficulties

Intuitionism also encounters severe ontological difficulties. Basic moral words have been held to signify indefinable moral qualities that are part of the world and yet not empirical or natural. To some philosophers, such a Platonism is difficult enough in itself, but it also raises the question of how these nonempirical moral qualities are related to the empirical actions to which they are attached, as when a person says, for example, that keeping one's promise or telling the truth is right or a duty. The moral qualities have been said to be consequential on the empirical qualities, but the nature of this relation has been left quite unclear. Critics, such as P.F. Strawson, an Oxford "ordinary-language" philosopher, have pointed out that the relation could not be definitional in the sense that the empirical qualities define the moral ones; for the intuitionists held that the latter qualities are indefinable. Nor could the relation be causal, such that the empirical qualities cause the moral ones. For, in order to ascertain a causal relation between two entities, one must be able to have an independent awareness of each; but the intuitionists held that there is no independent awareness of the moral qualities because they are always seen as dependent on the empirical qualities.

NATURALISTIC COGNITIVISM

Naturalism as a meta-ethical theory is, in important respects, an attempt to surmount the difficulties of intuitionism. It accords with intuitionism in its view that moral words or concepts signify real qualities in the world, that moral judgments may be true of the world, and that such judgments thus constitute knowledge. But naturalists seek to avoid the dogmatism and seeming obscurantism of the intuitionists' appeal to nonempirical intuitions and nonnatural qualities by directly identifying moral attributes with empirical qualities. There is no separate ontological realm inhabited by moral qualities, they hold, nor is a separate kind of cognition required to ascertain them. On the contrary, moral judgments are empirical statements verifiable by the same methods of natural science or experience as are other empirical, scientific statements. It follows also that moral knowledge is a part of natural science and not sharply separate from it, as the intuitionists hold.

Naturalism versus nonnaturalism and supernaturalism

In the sense in which Moore distinguished naturalism from his own theory, naturalism may include any position holding that moral qualities are not completely unique in kind. On the basis of this definition, even a theory that held moral qualities to be supernatural and that, for example, defined good as what God commands would still be naturalistic. According to the interpretation most widely accepted in the 20th century (the one to be followed for the most part here), naturalism is opposed, however, to supernaturalism as well as to nonnaturalism or intuitionism. In this more specific interpretation, naturalism holds that moral words signify only natural, empirical qualities, not also supernatural ones.

Ancient Greek and medieval naturalisms. The above characterizations of naturalism have used words like natural, scientific, and empirical. It is important to recognize, however, that the reference imputed to these words in any period will vary with the philosophic and scientific views prevalent in that period. The view of nature that obtained in the ancient and medieval worlds differed in important respects from that of the modern world since Galileo and Descartes (died mid-17th century). The most obvious difference is the widespread removal of final (or purposive) causes, either transcendent or immanent. In the Platonic and Christian doctrines, all things are products of supernatural causes and are fashioned to achieve purposes imposed on them from without by those causes.

Plato, Aristotle, and Aquinas. In important respects, however, the Platonic doctrine is closer to intuitionism than to naturalism, as so far defined. For Plato distinguished between intelligible reality and sensible appearance. The former was the realm of the eternal, unchanging Ideas or Forms, and especially of the Idea of the Good, the ultimate principle of all other Ideas. The sensible, on the other hand, was the realm of changing, empirical, natural objects, which are real and intelligible only insofar as they participate in the eternal Ideas. Hence, just as the intuitionists distinguished between nonnatural moral qualities and the natural empirical qualities to which they are somehow attached but which do not as such constitute moral realities, so Plato distinguished between the evanescent empirical qualities and the unique Idea of the Good, from which alone, by participation, they derive whatever goodness they might achieve.

Transcendent versus immanent forms

Aristotle, on the other hand, rejected this distinction between transempirical Ideas and changing empirical objects. He held, like Plato, that there are forms that give intelligibility and fixity to things; but, unlike Plato, he regarded these forms as being in the things themselves, not separate from them. These forms are ultimately one with the ends or final causes that constitute the good of each kind of thing, in the sense of its perfect functioning and complete development as a member of its species. Indeed, the nature of each kind of thing is identical with its form and end and hence with its good; and natural science attempts to ascertain these natures and their consequent properties by an empirical, inductive process that culminates in an intuitive grasp of the defining essence of each kind of thing. Hence, Aristotle's ethical doctrine is thus far naturalistic in at least two respects. First, goodness is an objective characteristic that belongs to each kind of natural thing; it is of the same kind as other natural characteristics—although in one respect it is more general than they, since objects in each category of things (substance, quantity, relation, and so forth) have their characteristic good. Second, the goodness of each kind of thing can be ascertained by empirical, inductive methods similar to those of all other scientific inquiries.

Thomas Aquinas, living in the 13th century, took over this basic Aristotelian position. As a Christian he believed, of course, in a transcendent God; but, unlike Augustine and his followers who had dominated much of medieval thought, Thomas did not simply absorb all goods and all realities into God's creative power as the first cause of all things. Instead, Thomas emphasized the importance of what he called second causes: the specific, immanent forms that constitute each kind of thing making it what it is. Hence, Thomas declared that good and being, or reality, are the same in each thing, differing only in that the goodness of the thing consists in the perfection that it strives to attain. To the objection that some realities are bad, Thomas replied that they are bad only insofar as they lack some part of their full perfection.

Difficulties in the doctrine of final causes. Some scholars have discerned severe difficulties in this Aristotelian–Thomistic doctrine of immanent final causes and perfections. It may be granted that for frogs and dogs it is easy to ascertain whether they have achieved their full development, the criteria of which are certain kinds of biological functioning. But how does this judgment apply to human individuals? In what respect does a Hitler or a Stalin not represent the complete development of the human form, whereas a Schweitzer or a Gandhi does? In what way does the concept of complete development account for criteria of moral goodness?

Ontological versus practical moral philosophy

To examine how Aristotle might deal with this difficulty, it must be recognized that there is a difference between his general ontological account of goodness (summarized in the preceding section) and the account of moral goodness or virtue that figures in his ethical philosophy. The difference derives from the sharp distinction he drew between theoretical and practical philosophy. Theoretical philosophy—which includes physics, biology, psychology, and metaphysics—deals with objects whose natures or essences are invariant and independent of human action or production. Practical philosophy, which includes ethics, deals with objects that vary because they depend for their existence and salient characteristics on human action. Such varying objects consist not of fixed essences but of human habits or dispositions and of institutions that are the effects of human contrivance.

Hence, human goodness or moral virtue, which is a habit or disposition, cannot be read off from the essence of man in the way in which the goodness of natural entities, which are independent of human contrivance, can be read off from their essences. On the contrary, since human goodness or virtue depends on human action, it will have an important variable feature. The matter of which the moral virtues are composed is human actions and emotions; this matter—*i.e.*, men's actions and passions—develops into virtues when a certain rational pattern is regularly imposed on them. This pattern appears as a mean between extremes and is determined by rational appraisal and ultimately by a kind of intuitive perception. It is in this way that the man of practical wisdom determines what is the right amount of fear to have in each dangerous situation, what is the right amount of food to eat in each gustatory situation, and so forth. In each case the formula is: neither too much nor too little but just the right intermediate amount, and this varies from one person to another and from one time to another. Aristotle also applied this formula, although with some modifications, to justice.

Intuitionist and naturalist strains in Aristotle

It is not entirely clear how Aristotle conceived the relation between this double determination of moral virtue or goodness (1) by a rational pattern or principle and (2) by intuitive perception and his more general ontological conception of goodness as consisting in subsistent final causes. In the sense in which he used "nature" in the *Nicomachean Ethics*, moral virtue for the most part is not natural but results rather from habituation. Hence, Aristotle was not a naturalist in the sense in which a naturalist holds that moral qualities exist as natural qualities and are knowable by the empirical methods of natural science. His position is in one respect closer to intuitionism; for, although he elucidated the rational, moral grasp of the means between excess and defect by analogy with the processes of the arts, he also insisted that the moral virtues are unique qualities quite different from those found in works of art. There was, nonetheless, a deep naturalistic strain in Aristotle insofar as he held that the nature of man, in the sense of his final cause, is to live in society and to have the corresponding virtues, especially that of justice. Thus, he referred, at one point, to natural justice, and in the *Politics* he declared that "the nature of a thing is its end" and that "man is by nature a political animal."

Varieties of contemporary naturalism. In the modern era, the idea of immanent final causes or perfections as part of the natures of things, for the most part, has been relinquished. The consequence for the theory of goodness or value can be seen in the statement of George Santayana, a 20th-century aesthetic Materialist, that "apart from ourselves, and our human bias, we can see in such a mechanical world no element of value whatever." Nevertheless, in some modern naturalisms, analogues can be found of the doctrine that goods are part of the essential natures of things.

An initial broad division can be made between objective and subjective naturalism. According to objective naturalism, goodness belongs to certain objects simply by virtue of their own natures or relations, regardless of whether the objects are known or desired. Such goodness depends upon the natural constitution of the objects and consists in what is required for those objects to achieve the perfect functioning or development of which they are capable. Santayana referred to such an objective goodness when he said that there are "natural relations that make one thing good for another." Food, for example, is good for an animal independently of whether it is known or desired by the animal.

According to subjective naturalism, on the other hand, goodness belongs to things in virtue of some sort of knowledge, desire, or interest. To call such naturalisms "subjective" does not mean that they hold the goodness or value of things to be merely apparent or unreal; rather, the point is that they regard value as pertaining to objects by derivation from the knowing or valuing subject, so that the value consists in or results from a relation between the object and the subject. But, once this relation obtains, the value genuinely exists; it is not mere illusion or appearance.

Epistemic versus nonepistemic subjective naturalism

There are two main kinds of subjective naturalism, epistemic and nonepistemic. Epistemic naturalism holds that goodness or value arises from the application of certain cognitive procedures to objects or experiences. Thus, according to John Dewey, "Without the intervention of thought, enjoyments are not values but problematic goods . . . enjoyment becomes a value when we discover the relations upon which its presence depends." Nonepistemic naturalism holds that goodness or value arises rather from certain desires, interests, or emotions of the organisms. There is an important distinction, however, between those nonepistemic naturalisms that are normative and those that are positive.

The normative variety imposes certain restrictions on the kinds of interests that generate goodness or value. Thus, Santayana recognized that "there are desires for what is bad," but he insisted that such desires are merely "incidental phases of consciousness" and, hence, represent only "momentary and partial interests." Opposed to these are "constitutional interests," which derive from the basic natures of human animals; these interests are "more comprehensive" than the former, and it is they that determine what is really good. He wrote:

> It is in reference to such constitutional interests that things are "really" good or bad; interests which may not be fairly represented by any incidental conscious desire. No doubt any desire, however capricious, represents some momentary and partial interest, which lends to its objects a certain real and inalienable value; yet when we consider, as we do in human society, the interests of men, whom reflection and settled purposes have raised more or less to the ideal dignity of individuals, then passing fancies and passions may indeed have bad objects, and be bad themselves, in that they thwart the more comprehensive interests of the soul that entertains them.

The positive variety of nonepistemic naturalism, in contrast, imposes no restrictions or requirements on the desires or interests that generate goodness or value. Thus, Ralph Barton Perry, a Harvard Realist philosopher, defined value as "any object of any interest"; he said that "any object, whatever it be, acquires value when any interest, whatever it be, is taken in it."

Realistic and descriptive status of naturalistic ethics

For all these varieties of naturalism, goodness and other valuational and moral qualities pertain to things themselves, either directly or in virtue of their relation to some condition of persons or other organisms—this condition being cognitive, conative, or emotional. Moral judgments, which predicate a moral quality of some subject, are hence descriptive, empirical propositions that are either true or false. It has already been argued that nat-

uralism does not incur the logical fallacy attributed to it by Moore and other intuitionists. Whether it incurs a different kind of logical fallacy, however, will be discussed below (*Noncognitivism: Moral statements as "commending"*). Nevertheless, many scholars hold that naturalistic theories face some serious difficulties. They are of two interrelated kinds, ontological and valuational.

Ontological difficulties of naturalism. The ontological difficulty pertains most directly to the objective variety of naturalism. How can goodness or value, one may ask, belong to things themselves irrespective of any desires or knowledge on the part of anyone? It must be emphasized that, in much ordinary thinking and talking, such objectivity of value is assumed. A distinction is sometimes drawn, for example, between men's preferences and their well-being; what men prefer is often not what is really good for them. Obvious examples are smoking and ingestions of heroin or of excessive liquor, in contrast to a moderate amount of food, drink, and rest. This distinction is held to be an objective one, in that what is and what is not for the good of an organism can be ascertained independently of its or anyone else's desires or preferences; the good consists in what contributes to the organism's effective and prolonged functioning as an organic whole.

For this distinction to be tenable as an ontological one, however, it would have to apply indiscriminately to all organisms. Thus, the food that is good for humans is not good for the beasts and plants that are destroyed, as the distinctive organisms that they are, in the process of becoming food for men. Germs are not good for human beings, but similarly human beings are not good for the germs they try to destroy. It seems, then, that what is good can be ontologically good in the way upheld by objective naturalism only at the cost of being made completely relative to each kind of organism. This much was already recognized and accepted by Aristotle in his immanent specific teleology. But Aristotle also upheld a natural hierarchy according to which some organisms are inherently superior to others in virtue of the superior functions they can perform. The criteria of such superiority he held to be quite objective: they include especially the ability to know more things and to know them more deeply in respect to their causes and the ability to engage in moral relations. But these criteria obviously raise many problems; as they stand, they suggest an anthropocentric approach that begs at least part of the question at issue.

Valuational difficulties of naturalism. The valuational difficulties faced by naturalism revolve around the question whether the definitions of goodness or value that its various versions claim to provide can account for moral goodness or value. It is significant that most 20th-century naturalists, like many earlier ones, concern themselves more directly with the general notion of value than with specific moral notions, especially deontic ones. The reason derives from three interrelated factors. First, naturalists see continuities between human and nonhuman nature, and corresponding continuities between their respective sources of value, especially in natural desires or impulses. Since, however, moral considerations do not pertain to nonhuman nature, the naturalists' general approach to value is, as such, nonmoral. Second, the naturalists regard value as affirmatively related to human nature rather than as negatively related—as concerned with the development rather than with the repression of natural tendencies. Insofar, then, as morality involves restraints on natural impulses rather than their fulfillment, the naturalistic approach to value does not, as such, provide for morality. Third, the naturalists consider value as pertaining to each organism or to each human being separately. But, insofar as morality concerns the interrelations among persons, it is not accounted for by naturalism.

Reasons for general concern with values

James's and Perry's ethics of demands and interests. With respect to each of the last two points. however, it is important to note that many naturalists supplement their general, nonmoral definitions of value by further criteria of moral value that include social and interpersonal relationships. The Pragmatist William James, for example, having said that "the essence of good is simply to satisfy demand," recognized that demands may conflict with one another. His solution was a simple, aggregative one, reminiscent of the Utilitarians:

> The guiding principle for ethical philosophy [must be] simply to satisfy at all times *as many demands as we can.* . . . There is but one unconditional commandment, which is that we should seek incessantly, with fear and trembling, so to vote and to act as to bring about the very largest total universe of good which we can see.

James did not, however, deal with the difficulties of this purely quantitative or aggregative doctrine, despite the fact that earlier in the same essay he had recognized the specific problem of justice.

Perry, having defined value as "any object of any interest," then faced the problem of how to ascertain "what is *really* valuable" and how to account for "the case in which an interest or its object is morally condemned." His reply was that the general definition of value must be distinguished from the solution of special problems of value and that in the special case of moral judgments, these "are concerned with organizations of values . . . with the relation of interests to the comprehensive purposes in which they are incorporated." This reference to "comprehensive purposes" suggests a quantitative approach to morality similar to that of James, with its corresponding difficulties. Subsequently, however, Perry tried to deal with the problem of justice by means of what he called "the postulate of concurrence," according to which "we have found a solution when, and only when, the wills of all are so attuned that each is content with a situation in which provision is made for all." This postulate obviously departs considerably from Perry's general definition of value, for it requires not merely that as many interests as possible be satisfied as fully as possible but that there be unanimous consent, even at the cost of lesser satisfaction of some interests. Nevertheless, this postulate can be accommodated to his naturalistic framework, since all of its concepts can be defined empirically.

General definition versus special problems of value

Dewey's ethics of intelligent action and enjoyments. The complex naturalistic doctrine of John Dewey took a somewhat different approach to these problems. He set himself in explicit opposition to Perry's equation of value with any object of any interest, declaring that this equation omits the normative aspect of value: it ignores "the difference between the enjoyed and the enjoyable, the desired and the desirable, the satisfying and the satisfactory." To account for this difference, an additional factor must be introduced. The enjoyments in which fulfillment of desires or interests directly consists must be subjected to experimental, operational tests in order to ascertain their consequences. Hence, Dewey advocated "defining value by enjoyments which are the consequences of intelligent action. . . . Enjoyment becomes a value," he held, "when we discover the relations upon which its presence depends."

This position, with its reference to enjoyments and consequences, is reminiscent of Utilitarianism. But Dewey did not indicate just how the consequences must be related to the initial enjoyments in order to yield a judgment of value. Perhaps he had in mind the sort of situation in which a person tests whether his enjoyment of smoking, for example, is a genuine value by ascertaining through statistical or other empirical methods what are its likely consequences. Here, however, a standard of value is already assumed: the maintenance of physical health. But suppose, instead, that one is testing the enjoyment derived from inflicting pain on innocent persons. By what standard does he test its consequences? The moralist may subject such enjoyment to "intelligent action" by ascertaining its consequences. But Dewey presents no way of showing that, when the consequences include increased suffering for others, the value of the enjoyment is removed, and it is rendered a moral disvalue. His naturalistic definition of value as "enjoyments which are the consequences of intelligent action" hence fails, not because it does not pass Moore's open-question test but rather because it fails to meet a simpler exten-

Relation of enjoyments to consequences

sional requirement: it would entitle the moralist to regard as genuine values sadistic or other acts that are unquestionably moral disvalues.

A possible solution of this problem in Dewey may be found by considering first his conception of the kind of intelligence that he insisted must be brought to bear on enjoyments before they can be considered to be values. He regarded science as the fullest development of intelligence, and science in its own right is both practical and valuational: "This element of direction by an idea of value applies to science as well as anywhere else. For in every scientific undertaking, there is passed a succession of estimates; such as 'it is worth treating these facts as data or evidence; it is advisable to try this requirement.'" The values here in question, however, are purely scientific or intellectual; how, then, do they bear on moral values? Dewey answered this question by emphasizing that scientific values, like moral ones, are interpersonal and intersubjective: the scientist must put his opinions to the test of scrutiny through free inquiry by other scientists using methods whose application and results are available to all. Dewey suggested that, in this respect, there is a potential alliance between scientific and democratic method:

Alliance of scientific and democratic methods

> It is of the nature of science not so much to tolerate as to welcome diversity of opinion, while it insists that inquiry brings the evidence of observed facts to bear to effect a consensus of conclusions—and even then to hold the conclusion subject to what is ascertained and made public in further new inquiries. I would not claim that any existing democracy has ever made complete or adequate use of scientific method in deciding upon its policies. But freedom of inquiry, toleration of diverse views, freedom of communication, the distribution of what is found out to every individual as the ultimate intellectual consumer, are involved in the democratic as in the scientific method.

The question still remains, however, whether these analogies between science and democracy, in respect of their methods of reaching conclusions, are sufficient to remove the possibility of the immoral (or undemocratic) use of scientific method. It might be held that all such uses are really abuses of the libertarian, egalitarian features of scientific method. But, even if scientific and democratic methods exhibit certain freedoms and equalities in their respective spheres and are to that extent similar to one another, this does not answer the question of whether science ought to be used to extend men's freedom and equality generally. Science and democracy may well be similar to one another in the ways indicated by Dewey; but this does not prove that science may not be used for undemocratic purposes. Hence, Dewey's appeal to the operations of scientific intelligence as the way to convert enjoyments into genuine values does not suffice to prevent his definition of value from being compatible with serious moral disvalues.

NONCOGNITIVISM

Although the main development of noncognitivist theories has occurred since the first third of the 20th century, their lineage can be traced back to various emphases in philosophers such as Hobbes and Hume and even to some points in Aristotle and in the medieval voluntarists.

Denials and avowals of noncognitivism

In contrast with both the intuitionists and the naturalists, the noncognitivists deny that moral terms have descriptive meaning, either at all or primarily. They also deny that moral judgments are propositions that state facts about the world or that reason or any other intellectual faculty can prove or justify any moral principle or, more generally, that there can be any conclusive justification of a moral judgment, as the intuitionists and the naturalists had assumed because of their assimilation of such judgments to knowledge certified either by intuition or by empirical methods.

Complementing these negative theses, the noncognitivists hold that moral words and sentences have various kinds of nondescriptive meaning; they express feelings or emotions; they evince attitudes; they exert an influence; they have a certain magnetic force; they issue commands or prescriptions; they guide choices; they commend, recommend, condemn, or disparage.

Concern with language and verifiability. An initial important characteristic of all of the contemporary noncognitivists, in contrast with the cognitivists, is their great concern with language. This concern is indicated in the titles of some of their chief books: A.J. Ayer's *Language, Truth and Logic* (1936; rev. ed. 1946), Charles Stevenson's *Ethics and Language* (1945), R.M. Hare's *The Language of Morals* (1952).

The reasons for this linguistic emphasis are similar to those that motivated the earlier emphases on language by Hobbes and Locke in the 17th century. In general, an explicit, direct study of language can help to clarify philosophic issues. But, more specifically, the philosophers in question hold that it is only by reference to language that the philosopher can account for many of the so-called entities or objects that, according to earlier Rationalistic or ontologically inclined thinkers, comprise an important part of the existing universe. Just as for Hobbes and John Locke the "intelligible forms" and "real essences" that Scholastic Rationalists had held to exist were mere words, whose purported ontological reference was to be accounted for by the philosophers' misuses of language, so, for the noncognitivists, the nonnatural or natural objects that the cognitivists had held to be the referents of "good" and "ought" are to be explained in terms of these philosophers' misinterpretations of the meaning or use of ethical words. There are no such objects in the world, or, at any rate, philosophers have no warrant for asserting that there are; there is only the use of language in certain nondescriptive ways. It is because the cognitivists have interpreted moral judgments as being descriptive that they have held that goodness or obligation is a real quality of things. Once it is seen that the only or the primary meaning of moral judgment is nondescriptive, this view of the ontological or objective reference of moral words must be relinquished.

Another way in which the noncognitivists' concern with language bears on their ethical theory is through their application of the verifiability principle to ethics. Earlier noncognitivists, such as Ayer and Stevenson, were strongly influenced by the Logical Positivists, philosophers such as Moritz Schlick and Rudolf Carnap, one of whose main tenets was that, for a sentence to be cognitively meaningful, it must be verifiable, at least in principle. They further held that there are only two mutually exclusive means of verification and hence only two kinds of cognitively meaningful sentences: (1) analytic sentences, verifiable merely by the meanings of their terms, and (2) synthetic sentences, verifiable by consideration of empirical data. Ethical sentences, however, are of neither of these kinds. For, if they were, then disputes over moral matters could be settled as readily as can disputes over analytic sentences (settled by logic or mathematics) and over synthetic sentences (settled by empirical observations, including those performed by experimental scientists). The fact is, however, that men may agree on all of the empirical facts and logical rules and still disagree on some moral issue. Thus, ethical sentences are not verifiable; hence, they are not cognitively meaningful. The earlier Logical Positivists tended to say, on the basis of this argument, that ethical sentences are not meaningful at all, that they are mere noises or ejaculations, as when one says, "Hurrah!" or "Ugh!" The later Logical Positivists, however, especially Stevenson, admitted that ethical sentences are indeed meaningful, but, since their meaning could not be cognitive, they invented a new rubric, "emotive meaning."

Cognitive meaning through verifiability

Moral statements as "commending." In addition to this argument from the cognitive undecidability of moral disputes, the noncognitivists supported their rejection of the cognitivist doctrines by several other kinds of argument. One was an appeal to Moore's open-question test. Ayer and Stevenson repeated this test in a form that was quite similar to Moore's; hence their views were vulnerable to objections similar to those given above. R.M. Hare, an Oxford "ordinary-language" philosopher, however, put the argument into a new form directly linked to his own noncognitivist thesis.

Value-terms have a special function in language, that of

commending; and so they plainly cannot be defined in terms of other words which themselves do not perform this function; for if this is done, we are deprived of a means of performing the function.

Analysis of attempts to define values

One of the examples that Hare gives to illustrate this impossibility of a "descriptive" or "naturalistic" definition of value terms is the attempt to define the goodness of a strawberry as meaning that the strawberry has such descriptive characteristics as being sweet, juicy, firm, red, and large. But, as Hare writes, "we sometimes want to say that a strawberry is a good strawberry because it is sweet, etc. This—as we can at once see if we think of ourselves saying it—does not mean that a strawberry is a sweet, etc., strawberry because it is sweet, etc. But, according to the proposed definition, this is what it would mean."

This argument can be formally presented in a way parallel to the presentation of Moore's open-question argument given above. Someone may define the goodness of a strawberry in terms of sweetness and other descriptive characteristics (*DC*):

(1) $G =_{df} DC$

If one now wants to say that a strawberry (*S*) is good (*G*) because it is sweet, etc. (*DC*); *i.e.*,

(2) *S* is *G* because *S* is *DC*,

then, according to the definition in (1), in saying (2) he would be saying:

(3) *S* is *DC* because *S* is *DC*.

There are two points here. One is similar to Moore's point: (3) is a tautology, but (2) is not a tautology, as it would have to be if (1) were a correct definition. Moore's emphasis in this connection had been that (3) does not capture the meaning of (2) because the property held to be synonymous with and definitive of goodness is seen by inspection not to be synonymous with it—so that the point concerns the nonidentity of two properties. Hare's emphasis, however, involves the quite distinct point that "good" does not signify any property at all, because a person in saying, "This strawberry is good," is commending the strawberry, not describing it, and this commending function is quite different from uttering the mere tautology that (3) is.

This argument of Hare's raises several important issues. One concerns what it means to refer to the function of words and sentences (see below *Function of moral judgments*). Another issue bears on the basic contention that the function of "good" and other value terms is exclusively or primarily prescriptive or commendatory, not descriptive. In connection with this issue, two questions must be distinguished: (1) Is the function of describing different from the function of prescribing (or commending)? and (2) May the same word or judgment perform both functions? At one point, Hare gives an affirmative answer to both questions: "Although prescription and description may be combined in the same judgment, description is not and never can be prescription." If, however, the same judgment can be used to perform both functions, some scholars have wondered what becomes of Hare's argument just presented. May not a man then use "This strawberry is sweet, juicy, firm, red, and large" to commend the strawberry, to urge others to taste it (or one like it), to guide persons' choices with regard to strawberries, and so forth? And may he not then define "good" and other value terms by means of descriptive terms such as "sweet," "juicy," etc., because these terms can be used to commend, prescribe, or guide actions?

Description versus prescription

In answer to the argument that to give such a descriptive or naturalistic definition of the goodness of strawberries is to incur the penalty of reducing a commendatory or prescriptive judgment to a tautology, one can again argue as above against Moore. The definition may be held to explicate or analyze the complex meaning of the value words; hence, because there is a difference in clarity between the word to be defined and its definition, the one cannot be directly substituted for the other so as to yield a tautology, for such substitution ignores the explicative function of the definition.

Moral statements as "commanding." Moral statements may be interpreted not only as commending but also as commanding.

Guides to action. More generally, it can well be argued that most descriptive words, which can be given descriptive definitions, may be used for prescriptive or action-guiding purposes. ("To commend," Hare says, "is to guide choices.") The Pragmatic philosophers, from Charles Sanders Peirce to C.I. Lewis, had indeed held that a bearing on action is basic to all descriptive language, in that to define any descriptive word, such as door, or to verify any descriptive statement, such as "This is a door," one, at least in principle, would have to refer to various relevant operations, such as turning a knob or pushing. Or, in order to ascertain what are the descriptive statements that a person believes to be true, one would have to examine what actions he performs or would perform in relevant situations. Hence, it is impossible to separate the meanings of descriptive words from the ways in which they either guide actions or are guided by actions.

Action-guiding function

Even without going the full length of the Pragmatic position, it is clear that descriptive words and statements can serve to guide action. An obvious example is, "There is a fire in the next room." Hare recognizes this point when he says: "Even plain judgments of fact may be action-guiding. For example, if I say 'The train is just about to depart,' this may guide a person who wants to catch the train to take his seat." But it is difficult to reconcile this recognition with statements that Hare makes when he is pressing his noncognitivist or prescriptivist thesis, as when he says: "A statement, however loosely it is bound to the facts, cannot answer a question of the form 'What shall I do'; only a command can do this."

For a statement of fact to guide action, it must in some way be related to the wants, desires, or interests of the person whose action is to be guided. If someone in St. Louis, for example, should be told, "A tree fell yesterday in central Siberia," this statement is quite unlikely to have any impact or bearing on his action, unless he perchance should have some interest in it. This factor of interest also operates on commands. If someone is told, "Go jump in the lake," such a command will not even tend to guide his action if he has no interest in jumping in the lake or in obeying the person who gives the command. To generalize, for any person's action to be guided by words or sentences in any use of language—prescriptive, descriptive, or other—what is said in those words or sentences must have a bearing on that person's wants or interests, and he must recognize and accept that bearing.

Commands versus statements. Hare also tries, however, to distinguish between commands and statements (the grammatical analogues of prescriptive and descriptive judgments, respectively) in terms of what is involved in sincerely assenting to each of them. To assent to a statement is to believe it to be true; but to assent to a second-person command addressed to oneself is "to do or resolve to do what the speaker has told us to do." Here, through the difference between belief and action, there would seem to be the basis for Hare's doctrine that descriptive statements, as such, present only what is to be believed, not what is to be done, so that moral judgments, because of their action-guiding function, cannot be descriptive.

Lack of parallelism

The contrast between statements and commands, as Hare here gives it, seems to rest, however, on an important lack of parallelism. He stipulates that the commands must be second-person ones, but he does not in parallel manner stipulate that true–false statements must also be second-person ones concerning the intentional action of the person addressed, of the form "You are doing *x*" or "You will do *x*" or "You did do *x*." But, in fact, to assent to such second-person statements about one's own intentional actions is either to be doing *x* or to resolve or intend to do *x* or to have done *x*. Even if it be admitted, as it should be, that to assent sincerely to a statement is to believe it to be true, it is impossible to believe in the truth of a second-person statement about what one will in the future do without intending to do it. And, because second-person commands concern the (directly or remotely) future actions of the persons to whom they are

addressed, the proper analogue of these imperatives is the true–false statement that the person addressed will perform some intentional action. But sincerely to assent to such a statement is to resolve to do the action, just as to assent to a second-person imperative is to do the action. In this respect, then, the purported distinction between what is involved in assenting to statements and to imperatives does not hold up. Insofar as statements describe one's own intentional future actions—and to be parallel to second-person commands they must be of this sort—assenting to them commits one to action just as does assenting to commands. Yet, it may still be contended that there is a difference between such statements and commands. It lies in the intention of the person who issues each kind of utterance; *i.e.*, the difference is not on the side of the hearer or addressee but rather on the side of the speaker.

Function of moral judgments. The contention that there is a difference between assenting to statements and complying with commands reopens the basic issue of the function of moral judgments and the relation of that function to meaning.

Austin's locutionary distinctions. Some philosophers have tried to clarify this issue by means of an important, though controversial, set of distinctions presented by the Oxford "ordinary-language" philosopher J.L. Austin. (The summary that follows cannot be completely faithful to Austin's position because of persisting unclarities as to what he meant by what he said.)

Locutionary, illocutionary, and perlocutionary

Austin distinguished three different kinds of speech acts —the various actions that a speaker may perform when he utters a sentence. The locutionary act is that of uttering a sentence that has a certain sense and reference. When a person says, for example, "The gun is on the table," this sentence has a certain sense, derived from the definitions of its constituent words and a certain reference, viz., to the state of affairs of the gun's being on the table. The illocutionary act is what the speaker intends to do in uttering the sentence: he may utter it to inform someone about the gun's location or also to urge someone to pick up the gun or to warn him to be careful about the gun and so forth. Informing, urging, warning are different illocutionary acts; *i.e.*, they are different intentions that a person may have in uttering the same sentence, in performing the same locutionary act. The perlocutionary act comprises the effects that the speaker produces or achieves by uttering the sentence—effects on the feelings, thoughts, or actions of other persons (or perhaps of himself). By saying, "The gun is on the table," for example, he may frighten his hearer or alert him to danger or make him leave the room. Thus, the purview of perlocutionary acts is causal, while that of illocutionary acts is intentional.

A basic question raised for ethics by Austin's distinctions is whether moral judgments can be viewed as locutionary acts the content of which has sense and reference or whether they must be viewed only as performing one or both of the other two kinds of acts. Before attacking this question directly, however, it must be noted that Austin's distinctions serve to differentiate some of the earlier noncognitivists, the emotivists, such as Ayer and Stevenson, from the prescriptivist Hare and the multifunctionalist Patrick H. Nowell-Smith, both of whom are Oxford Analysts. The difference is that the emotivists regard moral judgments primarily as causing or tending to cause certain reactions in other persons (as perlocutionary acts), whereas Hare and Nowell-Smith regard them rather as expressing the speaker's intentions of commending, urging, and so forth (as illocutionary acts). The meaning of a sign is thus, as Stevenson defines it,

> a dispositional property of the sign, where the response, varying with the varying attendant circumstances, consists of psychological processes in a hearer, and where the stimulus is his hearing the sign.

Psychological–causal definition of emotive meaning

When the response or the stimulus is a range of emotions, the meaning in question is emotive meaning. It is this emotive meaning, defined in this psychological–causal way, that Stevenson holds to pertain to moral judgments. Of such judgments, he writes:

> Their major use is not to indicate facts but to create an influence. Instead of merely describing people's interests, they *change* or *intensify* them. . . . The word "good" has a pleasing emotive meaning which fits it especially for the dynamic use of suggesting favourable influence.

Opponents of this perlocutionary, or psychological–causal, conception of the meaning of moral judgments have claimed that there are obvious objections to it. It equates moral judgments with propaganda, they argue, ignoring the many contexts in which they do not have the use or effect of moving, inciting, or propagandizing. The analysis also deprives moral judgments of rationality. When a man makes a moral judgment, he implies that he has reasons for it, and this having or giving of reasons is different from the exertion of causal pressures. (Stevenson wrote that, when one gives reasons, "these reasons which support your ethical judgment are simply a means of facilitating your influence.") Hence, this conception is charged with being false to many facts about moral language.

Hare's illocutionary view. Hare differentiates his own view of both imperatives and moral judgments from the perlocutionary interpretation. He points to the distinction "between telling someone something and getting him to believe or do what one has told him," and he says that commands are similar to statements as well as to moral judgments in that all of them "consist in telling someone something, not in seeking to influence him." But the "telling" in the case of moral and other "value" judgments, according to Hare, consists in a certain kind of illocutionary act. "Value-terms have a special function in language, that of commending. . . ." And it is this commending function that he holds to constitute the primary meaning of "good" and other value terms. But others have argued that, if the perlocutionary force of moral judgments can be separated out as not essentially characteristic of them, then why may not the commending function also be separated out? May not moral judgments function simply to make statements concerning salient facts about the situation, statements that are held to be rationally justified by considerations of happiness or welfare? One may, for example, say that Martin Luther King, Jr., was a good man, not with any thought of commending or praising him but simply with that of indicating his possession of certain qualities importantly related to human welfare. If motivating force is not necessary to a judgment's being moral, then it may be contended that a commending force or intention is also not necessary.

Difficulties posed against Hare's view

It has also been argued that Hare's identification of the meaning of "good," with its illocutionary force (commending), cannot be correct, because "good," as used in affirmative–indicative sentences, has the same meaning as it has when used in negative, interrogative, optative, hypothetical, and other sorts of sentences; yet, in none of these latter is it used to commend. The more general point here, which goes back to Austin, is that meaning taken as locutionary must be distinguished from speech acts that use words having meaning. To this argument Hare has replied that all meaning consists in the illocutionary force of speech acts and that the instances of sentences other than affirmative–indicatives involve combinations of two different speech acts (*e.g.*, that of commending and questioning) such that the original speech act (commending) is transformed by applying to it the other speech act (questioning). The issue is too complex to be dealt with further here. But the important question still remains of why moral words cannot be defined in terms of a certain content, having a certain sense and reference (locutionary), as against being defined only or primarily in terms of certain kinds of speech acts (illocutionary).

Commendatory invariance argument. In addition to the antinaturalistic argument examined above, Hare and other noncognitivists offer another argument for the exclusively illocutionary meaning of "good" in terms of commending. Their argument also purports to show why the meaning of good cannot be exclusively or primarily descriptive or naturalistic. They point out that "good" is applied to many kinds of objects that are descriptively

different—men, cars, apples, breezes, hammers—and yet it has the same meaning in all of these diverse applications. But, if good had primarily a descriptive meaning, consisting of the descriptive features that render these various objects good, then, these noncognitivists argue, its meaning would vary greatly. The descriptive features that make a car good, for example, are its smooth riding, fuel economy, and so forth; those that make an apple good are its being juicy, sweet, firm; those that make a breeze good are its being fresh and invigorating. It seems absurd, however, that good should have a different meaning in all of these contexts. Yet, this is what one would have to say if the meaning of good were exclusively or primarily descriptive.

The only way to avoid this absurd consequence, they hold, is to recognize that the primary meaning of good is not descriptive at all but rather evaluative or prescriptive: it is used to commend. For it is this commending function or illocutionary force that remains invariant amid all of the varying descriptive features that are referred to as good. Hence, according to the argument, a distinction must be drawn between the invariant meaning of good—which is illocutionarily commendatory or evaluative—and the varying descriptive criteria of its application to different objects.

Replies to the argument

To this argument, several replies have been made. One is that, if the argument's premise is that the meaning of a word must be invariant with every standard use of the word, then the commendatory aspect of the use of good varies as well as does the descriptive aspect. For the commending function itself varies with different dispositions of various commenders. The commending engaged in by a bully, for example, differs from that engaged in by a self-effacing person not only as to the objects they respectively commend but also as to the psychological features that directly enter into the commending itself. On the other hand, just as these commendings are generically alike simply insofar as they are commendings, so too the descriptive features to which good is applied are also generically alike simply insofar as they are features of objects. Hence, the argument does not show that the commending function of good is any more invariant than is its describing function.

The commendatory invariance argument has also been answered by pointing to more specific descriptive features that all objects called good have in common. All such objects are regarded by the persons who call them good as satisfying certain wants or answering to certain interests, and their being regarded in this way is just as pervasive a feature on the descriptive side as is the commending function on the evaluative side. It may, of course, be contended that such reference to wants or interests is but another way of indicating the commendatory aspect of good. But wants or interests are, after all, the conative and emotional factors that have figured in the nonepistemic forms of naturalism considered in the preceding section.

A related answer to the commendatory invariance argument has been set forth by another Oxford philosopher, Philippa Foot, who has contended that evaluative adjectives, such as "rude" and "courageous," as well as "good," carry a certain constant descriptive meaning with them. Reference to certain descriptive qualities or objects is essential rather than incidental to the meaning of such words. One cannot apply "rude" to any qualities indifferently but only to certain qualities connoted by it. So too with "good"; at least in its moral use, it always bears some reference to human needs and, in other uses, to specific purposes. This argument, if successful, hence serves to refute the idea that the descriptive meaning of good can vary randomly and that its evaluative meaning is independent of its descriptive meaning. It has been objected, however, that the concept of human needs is itself evaluative as well as descriptive and that the concept's descriptive component has considerable variation.

Cognitive undecidability of moral issues. A noncognitivist argument that has been even more influential than that of commendatory invariance has been the contention, mentioned earlier, that moral disagreements are not decidable by cognitive considerations alone. Noncognitivists recognize that, if two disputants agree on their moral principles, then cognitive considerations alone—especially appeal to empirical facts—will enable them to resolve their dispute. But the crucial question concerns precisely situations in which men disagree on moral principles. Here, cognitive considerations do not avail, so that moral judgments, in this view, are not (or at least not in principle) cognitive. Opponents have contended, however, that this argument overlooks the parallel fact that, when men disagree on matters of basic scientific principle—of fundamental ways of approaching problems—cognitive considerations also do not avail to resolve their dispute, because what they accept as cognitive is then also different. The trouble with this argument—from the cognitive undecidability of disputes over moral principle—has often been that, whereas philosophers refer to moral principles at a quite basic level, they do not refer to scientific principles at a comparably basic level. But a dispute between adherents of Christian Science and of physiological science is no more resolvable by cognitive means alone than is a dispute between adherents of Nazi ethics and of democratic ethics. The difficulty about cognitive decidability applies to all disputes in which there is disagreement on basic principles, not only to moral disputes.

Disagreement in ethics and in science

Logical structure. A further phase of the contention that moral disagreements are not cognitively decidable bears on the logical structure of moral argument. Noncognitivists appeal to the principle that no "ought" conclusion can be derived from exclusively "is" premises; *i.e.*, that purely factual–descriptive premises cannot logically justify any moral conclusion. The reason for this contention is that "nothing can appear in the conclusion of a valid deductive inference which is not, from their very meaning, implicit in the conjunction of the premises." And, because naturalistic and other cognitivist arguments purport to infer "ought" conclusions from premises in which "ought" or some word definitionally related to it does not appear, their inferences are invalid.

One aspect of this argument has already been considered above: that for a judgment to be a moral one, it must serve to guide choices, and that such guidance cannot be given by descriptive statements. Another aspect of the argument, however, bears more narrowly on the contention that an ought statement cannot be derived from non-ought premises. One objection to this contention is based on the logic of material implication, in which "If f then q" is true, except when f is true and q false. Among the theorems of this logic is the proposition that a false statement materially implies any statement. Hence, according to this theorem, an ought statement is logically derivable from any false is statement; *e.g.*, "The moon is made of green cheese" materially implies "One ought to tell the truth." Another theorem of material implication is the law of addition, according to which any statement materially implies (symbolized $\supset$) the disjunction (symbolized $\vee$) of itself and any other statement; *i.e.*, $p \supset (p \vee q)$. According to this rule, the statement "Chicago is in Illinois" materially implies "Chicago is in Illinois or one ought to keep one's promises." These inferences of ought statements from is statements, however, are admittedly vacuous, because any statement could be substituted for the ought statement without changing the validity of the inference.

"Ought" from "is" in material implication

Sets of terms. A further kind of objection to this principle has appealed to the specific meanings of certain crucially relevant terms. One set of such terms is ought and can. It is often held that ought implies can; *i.e.*, that, if one ought to do something, then it must be the case that one is able to do it. But this, in turn, entails that cannot implies "It is not the case that one ought"; *e.g.*, that "A cannot tend B's garden" implies "It is not the case that A ought to tend B's garden." Thus, from a certain recognized principle, it follows that an is statement about someone's inability to do something implies a negative ought statement.

Another set of terms that has been held to refute the no-ought-from-is principle derives from institutional facts; *i.e.*, facts based on institutions such as promising

"Ought" from "is" in such institutions as promising and trading

and buying and selling. It is held that such facts entail corresponding obligations; for example, if *A* says to *B*, "I promise to pay you five dollars"—which is an is statement—then from this there follows the ought statement "*A* ought to pay *B* five dollars." This argument, which has been advanced by John R. Searle, a California philosopher, has aroused great controversy. It may be suggested, however, that, if the entailment validly succeeds in generating an ought conclusion, then its premise is not a purely factual, nonmoral is statement. For only if someone already believes in the moral obligatoriness of keeping one's promises will he agree that *A*'s saying to *B* "I promise to pay you five dollars" entails that *A* ought to pay *B* five dollars. Hence, in this view, the argument does not genuinely go from a nonmoral is to a moral ought.

There may well be, however, other ways in which normative conclusions can be derived logically and nonvacuously from factual premises. There may also be ways in which basic moral principles can be cognitively or rationally established from premises partly conceptual and partly factual. Some of the considerations relevant to these suggestions will be presented below (see *Formal deontological ethics: The principle of universalizability*). The above summaries of recent controversies over noncognitivist meta-ethics, however, indicate that there are strong current tendencies toward a return to naturalism and cognitivism in meta-ethics.

III. Normative ethical theories

Normative ethics is that part of moral philosophy that is concerned with justifying and applying criteria of what is morally right and wrong, good and bad. These activities of normative ethics issue in normative moral rules and judgments, and these, in turn, have direct implications for what human actions, characters, institutions, and ways of life should be like.

PRINCIPAL ISSUES OF NORMATIVE ETHICS

The central question of normative ethics is: What are the criteria for determining which actions are morally right and morally wrong? This question is related in various ways, however, to three other important questions.

Virtues and vices and social ethics

Three traditional questions. The first of these questions is that of moral virtues and vices: What characters of men are morally good and bad? Because men's characters, however, are both developed by and exhibited in the corresponding actions, the question of actions is logically prior. A morally good man is one who does what is morally right, from a firm disposition to do such actions because they are morally right. Thus, as far back as in Aristotle's *Nicomachean Ethics*, the various moral virtues and vices were defined in large part through definitions of the corresponding kinds of action.

Another traditional question of normative ethics concerns social institutions: How should society be organized and what specific policies should its various agencies pursue? This question, which is basic to social ethics, has a twofold relation to the two previous questions of normative ethics. On the one hand, the kinds of actions that men perform and the kinds of character that they acquire are strongly affected by their social environment. On the other hand, what kind of social environment there ought to be is to be determined by considering what kinds of action and character ought to be fostered among men. Consequently, the question of social ethics, like that of moral virtue, logically presupposes the central question of the criteria of morally right actions.

Intrinsic values

A third traditional question of normative ethics concerns intrinsic values: What things or experiences are worthwhile or desirable on their own account? This question constitutes the whole subject matter of axiological (or value) ethics and is basic to teleological ethics. Philosophers have given many different answers to it. The most famous answers are that pleasure and happiness are the intrinsic values; but power, beauty, knowledge, prestige, companionship, love, and many other values have been upheld either as basic to all the rest or as essential components in the whole spectrum of values.

Two points must be made about this question of values. One bears on the relation between values of the kinds just listed and the moral virtues or goods of character previously mentioned. These two kinds of goods or values seem clearly to be independent of one another, for one may have the worthwhile things just listed without having a morally good character, and conversely. Some philosophers, however, have refused to accept this distinction between moral and nonmoral goods. In a tradition going back at least to Socrates, they hold that pleasure, power, and the like are not goods at all unless they are had by morally good men. On this refusal to accept the distinction depends Socrates' famous paradox that it is better to suffer injustice than to do injustice, because any action, no matter how much pleasure it gives the doer, is no good at all if the psyche in which the pleasure occurs is not morally good. In the same tradition belongs Kant's insistence that the only thing that is good without qualification is a morally good will; all other goods become bad in the absence of this good.

These contentions serve to emphasize a second point about the moral import of the whole question of nonmoral values. Persons differ over these values, partly in terms of their native temperaments and other endowments. It seems morally appropriate, therefore, not to try to legislate about what the "true" values are but rather to allow persons to pursue whatever values they want, so long as what they do is not morally wrong. This consideration can be made to fit into both a deontological and a teleological framework—into the deontological on the grounds of justice and equal freedom, into the teleological on the ground of the fullest possible maximization of values. Hence, the question of what actions are morally right or wrong retains its centrality and logical priority in normative ethics, because, so long as the grounds or criteria of moral rightness are fulfilled, the question of intrinsic values may not pose any separate moral issue.

Deontological versus teleological theories

The central question. The answers to the central question of the determining criteria of normative ethics fall into two broad groups, deontological and teleological. The main difference between them is that deontological theories do not and teleological theories do appeal to value considerations in answering the question. The only normative moral concepts used by deontological theories in their answers are themselves deontic ones, whereas teleological theories insist that value concepts must be used centrally in answering the question.

Five further differences between deontological and teleological theories may be indicated. (1) Deontological theories advocate doing certain things on principle or because they are inherently right, whereas teleological theories advocate doing things because of their consequences. (2) Deontological theories set forth absolute obligations, whereas teleological theories set forth only conditional obligations. (3) Deontological theories set forth formal or relational criteria, such as equality or impartiality, whereas teleological theories set forth material or substantive criteria, such as pleasure or happiness. (4) Deontological theories set forth distributive criteria, concerning how goods or evils should be allocated or divided among men, whereas teleological theories set forth aggregative criteria, concerning the maximization of goods and minimization of evils. (5) Deontological theories set forth plural as well as unitary criteria, whereas teleological theories set forth only a unitary criterion, such as doing the greatest amount of good.

Although there are important bases for most of these differentiations, they do not correspond at all points to the full complexities of deontological and teleological theories. Some of the main varieties of these theories will now be examined.

DEONTOLOGICAL THEORIES OF NORMATIVE ETHICS

Deontological theories fall into two main groups, material and formal. Material theories hold that the criterion of the rightness or obligatoriness of actions consists in some feature either of the actions themselves or of the background of the actions, whereas formal theories

hold that it consists rather in some logically necessary relation between the judgments or rules in accordance with which the actions are performed.

Material deontological ethics. Material theories, in turn, vary in several respects. One important respect involves the alternatives of pluralism and monism: whether the features of actions or of their backgrounds that make them duties are of many different kinds or whether they are ultimately of one kind.

Pluralistic theories. A prime example of pluralism is the doctrine of Prichard discussed above in connection with intuitionism. According to that doctrine, actions such as keeping one's promises, repaying benefits, speaking the truth, and showing consideration for the feelings of others are inherently and self-evidently obligatory—*i.e.*, each of these kinds of action is obligatory simply in virtue of being the kind of action that it is, so that no further reason can be given for its being obligatory. Various doctrines of Existentialism and of situation ethics are also of this pluralistic sort.

Classification of prima facie duties in Ross's intuitionism

A closely related example of a pluralistic material deontological theory is that of W.D. Ross, also discussed above, especially his doctrine of prima facie duties. Ross's theory, however, was more complex than Prichard's. He not only expanded Prichard's list of inherently right actions, but he also grouped such actions in ways that came close to accepting, in part, a more unified teleological position. He distinguished six kinds of circumstance having moral significance, each of which was the basis of one or more kinds of prima facie duty. These circumstances include, for example, previous acts of one's own, including one's implicit promise "not to tell lies which seems to be implied in the act of entering into conversation (at any rate by civilized men)"; this is the ground of duties of fidelity. In addition, one's previous wrongful acts are the ground of duties of reparation. Other duty-grounding circumstances include previous acts of other men that are helpful to oneself, on which rests the duty of gratitude; distributions of happiness not in accordance with merit, on which rests the duty of doing justice; and the facts that one can both benefit and harm other men, on which rest the duties, respectively, of beneficence and of nonmaleficence.

Ross went on, however, to "arrange in a more systematic way the main types of duty," and his first step in this direction was to propound as "self-evident that if there are things that are intrinsically good, it is *prima facie* a duty to bring them into existence rather than not to do so, and to bring as much of them into existence as possible." This consideration, however, because it bases the duty to perform certain actions on the intrinsic goodness that they bring about, is close to, if not identical with, a teleological, Utilitarian position. Ross held, in contrast to Prichard, that the duties of beneficence, nonmaleficence, self-improvement, and justice can all be shown to follow from this consideration. He also insisted, however, that this good-maximizing consideration is not the ground of such other duties as that of keeping one's promises, so that his position is not completely teleological, and it remains pluralistic. In this connection, he maintained explicitly, in contrast to the Utilitarians and to monistic deontologists, that "in principle there is no reason to anticipate that every act that is our duty is so for one and the same reason."

Difficulties stemming from pluralism: conflicting duties

In addition to the difficulties that material deontological theories such as those of Prichard and Ross incur through their intuitionist meta-ethics, there are further difficulties that stem from their pluralism. Critics have argued that, if no general reason can be given why certain kinds of acts are obligatory, then the rational justification of moral judgments becomes difficult or impossible; and there is then no rational way, for example, to deal with moral disagreements. Moreover, in the absence of a general reason or principle, there is no way of resolving conflicts between duties—for example, between a man's duty to tell the truth and his duty to refrain from harming other persons.

An attempt to resolve this problem of conflicting duties within a pluralistic deontological framework has been made by A.C. Ewing, a contemporary Cambridge philosopher. He contends that the resolution of such conflicts does not require that all duties be deducible from a single principle, as the Utilitarians insisted; it is sufficient to be able to show that all duties harmonize with one another, in that they form a consistent system and fulfill a kind of coherence test similar to that envisaged in the coherence theory of truth, which defines truth by the cohering relations among propositions. Specifically, this means that "to further one [prima facie duty] on the whole furthers others and to violate one tends to involve sooner or later violating others." Ewing admits that not all conflicts between duties could be resolved in this way; but he holds that, at least in their general tendency, the prima facie duties fit together. This still leaves open, however, the possibility of clashes between duties in particular cases.

Monistic theories. In contrast to the pluralistic varieties of material deontological theory, monistic theories set forth a single, nonformal principle as the criterion of the rightness of actions. Because this principle has to account for the moral quality of a large variety of actions, its content is obviously much less closely related to the description of the specific actions, the rightness of which it is held to ground, than is the content of the pluralistic deontological theories. Examples of monistic material principles to which deontological philosophers have appealed include the will of God, conformity to nature or to human nature, human dignity, the march of history, and consent or contract. Such theories may be more or less naturalistic. Often, however, the principles to which they appeal are not amenable to direct empirical scrutiny, either because they refer to the supernatural, or because of their generality, or because the normative character of the principle is interpreted in such a way as to withstand empirical testing, as when it is denied that any possible empirical test could refute the principle that all men are equal in dignity.

Mixed and pure deontological theories

Monistic deontological theories sometimes include teleological elements. Theories that hold, for example, that what God has commanded is the basis of all right action sometimes also say that obedience to God's commands yields more good than does nonobedience. In such cases, however, the theories may not be properly deontological, if what they mean is that the ultimate reason for obeying God's commands is the good consequences of such obedience. Many monistic material theories, however, remain within a strictly deontological emphasis. They insist, for example, that God's commands must be obeyed not because of any good to be derived therefrom but simply because they are God's commands; or that the imperatives of history must be obeyed regardless of the suffering that this obedience may cause; or that the duty to respect the rights of persons has nothing to do with whether anyone stands to benefit thereby; or that obligations must be grounded on consent even if acting on certain persons without their consent will yield more overall good; or that the dignity of criminals as human persons requires that they receive the legally allotted punishment even if giving them psychotherapy instead will do them or the society more good in the long run.

The principle of consent is an example of a material deontological principle that has had a powerful influence both in general ethics and in political philosophy during the modern era of individualism and liberal democracy. According to this principle, an action is obligatory if and only if the person who has the obligation has consented to undertake it, in the sense that he has voluntarily accepted it either directly or through some indirect, implicit, or tacit agreement. Varieties of this principle range from making the act of promising the paradigm of all grounds of obligation to theories that hold that obligation may derive from a variety of human transactions—but still, the transactions at some point must contain a consensual element. Through the theory of the social contract, political obligation has been held to rest on this basis of consent. Although in its political form the theory has regularly been attacked as fictional, in its more specific interpersonal form, it has often been accepted as

giving a sufficient, even if not a necessary, condition of obligation.

Advantages and problem of monistic theories

The monistic material versions of deontological theory seem to avoid some of the problems of the pluralistic versions. They provide a general reason for men's obligations and hence also a possible way of ordering obligations and resolving conflicts among them. On the other hand, they are faced with the difficult problem of explaining why, apart from any reference to good consequences, the principle they uphold should be regarded as grounding men's obligations. This problem persists even when the principles are fitted into broader philosophical frameworks, such as that provided by proofs of God's existence and creation of the world or by doctrines of the historical development of societies. In one respect, the problem is that of deriving an ought from an is; for, even if the principle is itself held to contain an ought, there still remains the question why the specific descriptive features that define the principle are held to include the ought; *e.g.*, the question of what are the grounds for the connection between God as supreme being or creator of the world and as one who ought to be obeyed.

Formal deontological ethics. Because material deontological theories seem not to be self-validating, for the reasons just suggested, many deontologists have developed formal theories of obligation. Such theories appeal to a logically necessary relation conceived as holding between moral judgments, so that to perform actions that contravene certain judgments is to incur inconsistency or self-contradiction. The appeal to consistency in such theories is hence much more logically rigorous than that contained in Ewing's coherence test, discussed above. The general point of formal deontological theories is that, if to perform certain kinds of action logically commits the agent to self-contradiction, then his performance of these actions is not rationally justifiable; thus, it is both rationally not his duty to perform them and it is rationally his duty to refrain from performing them. Sometimes, this latter point entails also the statement that it is rationally his duty to perform some other, opposed action; for example, if a man is asked a question to which he has to reply and if it is rationally his duty to refrain from giving a false reply, then it is rationally his duty to give a true reply.

The principle of universalizability. Formal deontological theories attain a logically necessary relation between moral judgments by using the principle of universalizability. In its most general logical form, this principle says: if some predicate P belongs to some subject S because S has some quality Q (where the "because" is that of sufficient reason), then P logically must belong to all other subjects S_1, S_2, . . ., S_n that have Q. If someone denies that P belongs to S_1 . . . while admitting that P belongs to S because S has Q, then he contradicts himself, because, in saying that P belongs to S because S has Q, he implies that all that has Q is P; but this is inconsistent with his denial that S_1 . . ., which has Q, is P.

Logical and applied forms of the principle

The principal of universalizability is given a moral application in such a form as: What is right (or wrong) for one person must be right (or wrong) for any similar person in similar circumstances. This is a direct application of the logical form of the principle, in which the P means right (or wrong) and the S signifies some person who performs some action. Thus, if some person S says that it is right for him to do x—for example, to make a lying promise—because he has Q (in which Q means, for example, wanting to avoid mild inconvenience and the "because," as before, is that of sufficient reason or sufficient justifying condition), then he must admit that to make a lying promise is right for any other person who wants to avoid mild inconvenience. If S denies that such action is right for some other person S_1, then S contradicts himself. For, in saying that it is right for him to make a lying promise because he wants to avoid mild inconvenience, he implies that to make a lying promise is right for all persons who want to avoid mild inconvenience; whereas, in denying this judgment in the case of S_1, he is saying that to make a lying promise is not right for some person who wants to avoid mild inconvenience.

Controversial questions on universalizability. Although the logical aspects of the principle of universalizability are clear, its moral application raises at least three controversial questions.

(1) There is the question of commitment to a general rule. This controversial question deals with whether every agent in performing some action logically commits himself to a general rule; *i.e.*, whether every particular moral judgment logically commits one to such a general judgment as: "All persons who want to avoid mild inconvenience have a right to make lying promises." The objection sometimes made is that every situation is unique, so that a person who makes a moral judgment in his own case is not logically required to "legislate for everyone." The chief answer to this objection is that, insofar as a person gives reasons for actions and these reasons are regarded as sufficient justifying conditions of the actions, he is logically required to accept the corresponding general rules that apply the predicate of the original singular judgment to all other persons who fit the description contained in the reason he has given.

Three evasions of universalization

Philosophers have suggested at least three ways of avoiding such generalization beyond one's own particular case. One way is by not giving or having any reasons at all for one's actions. But there are several difficulties with this ploy. It would then become impossible to teach or discuss how to act with other persons, including one's children; for all teaching and discussion involves at least an implicit giving of reasons. A man would also condemn himself to arbitrariness and purposelessness, for it not only is a matter of giving of reasons to others but also of having reasons for one's own actions. The issue, in other words, is not only or mainly the rhetorical one of what to say to other persons but rather the logical one of the agent's own rationale or basis for acting. If he has no reasons, then he cannot guide or control his actions at all; if he has reasons, then he is logically committed to general rules.

A second way of avoiding such generalization beyond one's particular case is by having or giving insufficient reasons for one's actions. The reasons that a man gives would thus be, at most, only necessary conditions, so that, in saying that he had a right to make a lying promise in some particular case because he wanted to avoid mild inconvenience, he may have left out other reasons that are required for the full justification of his action. Hence, if someone is accused of inconsistency because he denies that someone else who wants to avoid mild inconvenience has the right to make a lying promise, he can always answer that the circumstances are different in the two cases. His full reason for making a lying promise may include other factors besides wanting to avoid mild inconvenience, factors that are not present in the second case. In this way he can manage to float along without logically committing himself to accept for others the moral judgments that he accepts for himself. An answer to this criticism, however, is that it is always rationally possible to remove the vagueness of one's reasons for acting by insisting on a fuller account of them, so that he must fill out his previous necessary justifying conditions to make them sufficient conditions. Here again, the rhetorical point must not be confused with the logical one.

A third way of avoiding the generalization of reasons beyond one's particular case is to make the reasons so detailed that they apply only to oneself or to one's particular group and not to any other person or group. Such individualizing or specializing of reasons is prominent in moral and political controversy. "The reason why I have the right to take this money and you do not has to do with the features of my peculiar situation, involving as they do the various debts I have incurred, the causes of my incurring them, and so forth—all of which are quite different from your situation or that of anyone else." "Our group has the right to engage in this violent action, while your group does not have this right, because our group uniquely possesses certain insights about the evils of our society and about their remedies, while no other group, including yours, has these insights." In extreme cases, this individualizing of reasons even adduces the

proper name of the person in question as all or part of his reason for performing his actions.

Arbitrariness of such evasions

Such objections to the universalizability principle based on the possibility of individualizing or specializing one's reasons, however, have been charged with suffering from one or both of two kinds of arbitrariness. One kind consists in the falsity of many of the claims of uniqueness, the other in the fact that the individualized or specialized reason may be irrelevant to the claim being made. This is most obvious when a person's proper name or some other unique property is given as the reason for his having a certain right.

(2) There is also the question of criteria of relevant similarities. The issue of relevance raises a second controversial question for the moral application of the principle of universalizability. The principle refers to similar persons and similar circumstances but does not reveal which similarities are to count as the relevant ones or which reasons that might be adduced for an agent's action are to be regarded as appropriate. Moral philosophers have suggested at least seven different ways of answering the question of the criterion of relevant similarities. They have held that this criterion is to be determined (1) by the specific purposes of acts or institutions, (2) by the consequences for the general welfare, (3) by what will serve to achieve equal rights for all persons, (4) by a person's willingness to allow some quality to be relevant regardless of whether he is on the giving or the receiving side of policies that make that quality decisive, (5) arbitrarily, (6) by history or precedent, (7) or by what is necessarily and hence nonarbitrarily connected with the category of action. Some of these criteria are obviously based on the philosophers' own normative moral principles, whereas others are determined by meta-ethical considerations, including doctrines about the kind of rationality available in moral situations. Further examination of this problem entails more complexities than can be gone into here.

(3) There is the question of its standing as a moral principle. The problem of relevance leads to a controversial question about the moral application of the principle of universalizability: In what sense is it a moral principle?

Necessary or sufficient condition of rightness

It directly requires consistency; but does this requirement give a necessary condition of moral rightness or a sufficient condition or neither? Some philosophers have upheld the last alternative, insisting that the requirement of consistency is in no sense a moral requirement and that, in particular, it must not be confused with the requirement of impartiality; *i.e.*, that one not make exceptions in one's own favour. The contention here is that a person can conform to the principle of universalizability and still be partial toward himself or his group. This, however, involves the difficulties already raised in determining relevant similarities; to the extent to which they can be overcome, the requirement of consistency would entail at least a modicum of impartiality. This point is basic to distributive justice: if a man claims certain goods for himself on the basis of some rule, then he must acknowledge the right of other persons to such goods insofar as they fall under that same rule. Such distributive considerations based on formal criteria are a characteristic feature of the kinds of formal deontological theories encountered in ethics.

The principle of universalizability states only a necessary but not a sufficient condition of moral rightness. The moral rightness in question, moreover, is confined to matters of distributive justice and hence excludes supererogatory actions that, by difinition, go beyond the dictates of duty. Such a limitation entails the conclusion that morally wrong as well as morally right actions and judgments can be universalized, so that conformity to the principle of universalizability is not of itself a guarantee of moral rightness. It also entails the consequence that theories that use this principle cannot rationally be monistic ones, because they must appeal to at least one other, substantive principle in order to supply a sufficient condition for grounding what is morally right.

Universalizability and justice. The contemporary Belgian philosopher Chaïm Perelman has discussed in some detail the bearing of the principle of universalizability on the criterion of justice. He holds that it provides, as such, only for formal justice. To give content to this form requires filling in the criterion of relevant similarities; *i.e.*, what determines the essential category into which persons should be classified for purposes of determining when they are entitled to similar or dissimilar treatment. Examples of such essential categories are the various criteria that have historically competed with one another for acceptance as the basis of just distribution: strict equality, merit, work, need, rank, or legal entitlement. In his earlier work Perelman held that this filling in was a matter of arbitrary preference. In his later work, however, he has emphasized the importance of historical context and precedent in determining the content for the form of justice. In any case, it is importantly true that what is here called the principle of universalizability in its moral application has been basic to the theory of justice from the Greeks until today—but only as a necessary not a sufficient condition of justice.

Need to define content for the form of justice

Kant's use of universalizability. Immanuel Kant was by far the most famous philosopher who has tried to use the principle of universalizability as a sufficient condition of moral rightness and hence as the sole basis of monistic formal deontological theory. The interpretation of Kant's moral philosophy, however, has raised controversies, principally over the question of whether his main doctrine is purely deontological or whether it is also and even mainly teleological. There is also controversy over whether, insofar as it is deontological, it is purely formal or whether it also upholds a material principle.

The chief ground for regarding Kant as a formal deontologist is that he set forth as the criterion of actions that are to be morally right or obligatory the stipulation that their maxims—the rules that the agent intentionally follows in performing the actions—must be universalizable without self-contradiction. This criterion is purely formal in that it stipulates nothing about what makes actions duties except that their maxims must be free from self-contradiction when universalized. This is the first and most famous formulation of the categorical imperative: "Act only on that maxim through which you can at the same time will that it should become a universal law."

Maxims free from self-contradiction: the categorical imperative

It will be helpful to compare the sense in which Kant's criterion is formal with that in which Aristotle's principle of the Golden Mean might be considered to be formal. Aristotle said that, for an action to be morally right, it must fall between the extremes of excess and defect; and this is, in one sense, a formal principle because it is put in terms of certain ratios or proportions. But Aristotle's principle requires careful consideration of the particular contents and circumstances of actions in order to determine just where the Golden Mean lies. Kant's categorical imperative, on the other hand, might seem to abstract from such contents and circumstances. If this were so, however, it would be difficult to see how the categorical imperative could be applied at all; for, in order to determine whether the rule on which one intends to act can be a universal law, he must be as clear as possible about just what the rule is, and this clarification requires consideration of the contents and circumstances of his projected action.

Despite such substantive considerations, Kant's categorical imperative can still be regarded as a formal deontological principle because it makes the universalizability of the rules or maxims of actions the test of the actions' moral rightness. It is important, however, to see just how universalizability operates in such a test. One of Kant's examples was that of the wrongness of making a lying promise in order to advance one's self-interest. His argument was that, if the rule that a person may make a lying promise to advance his self-interest were made a universal law, this would mean that no promises could exist because of the breakdown of the mutual confidence on which promises depend. But this nonexistence of promises would contradict the assumption on which the man acts, viz., that some promise does exist—his own. The contradiction or logical impossibility, then, lies in the relation between the existence of a certain kind of action

Example of a lying promise

when performed by the agent and the nonexistence of that same kind of action given the rule that everyone performs that kind of action. It is the contradiction between "Some promise exists," when made by the agent, and "No promises exist," when there is universal disbelief in anyone's saying that he promises something. Kant's point, then, is that an action is morally wrong if, when its maxim is universalized, it would be both self-annihilating (because such action would become impossible) and self-defeating (because the achievement of the action's purpose would become impossible).

There is an important difference between the ways in which universalizability was used by Kant and by many contemporary moral philosophers. The latter use it not to establish the logical impossibility of the universal performance of certain kinds of actions but rather to establish that a contradiction arises when an agent holds that a rule that he applies in his own case does not apply in other relevantly similar cases. The contradiction or logical impossibility here attaches not to the universal application of certain rules or maxims but to the rightful application of different rules to similar cases. In Kant's use of the principle, the logical impossibility is one of the actual occurrence of the universal application of the rules in question—*e.g.*, it is logically impossible that all promises be lying ones. In the moderns' use of the principle, on the other hand, the logical impossibility is one of the joint rightness of rules or judgments that are mutually inconsistent. When modern moral philosophers prove that to make a lying promise in order to advance one's self-interest is wrong, their argument is not that, if everyone made such promises, there would be no promises but rather that the agent contradicts himself if he holds both that it is right for him to make a lying promise for the reason indicated and that it is not right for some other person to make a lying promise for this same reason. The contradiction is not, as with Kant, between the two statements that a certain kind of action both exists and does not exist but rather between the two judgments that a certain kind of action performed by similar persons in similar circumstances is both right and not right.

Its effectiveness in demarcating wrong from right

The crucial question that Kant's use of universalizability raises for normative ethics concerns the extent to which it can serve to demarcate the morally wrong from the morally right. If actions that are admittedly morally wrong can be universalized in Kant's sense, then his criterion is unsuccessful. The attempt to answer this question is complicated, however, by the fact that Kant introduced various teleological or at least axiological elements into his further discussion of his principle. The most obvious such element is his reformulation of the categorical imperative in terms of ends: "Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end." Another reformulation appeals to autonomy, a point that became central to Kant's political philosophy. In addition, even in his illustrations of the first, formal version of the categorical imperative, Kant distinguished between perfect and imperfect duties on the ground that violations of perfect duties cannot even be conceived as universal laws without self-contradiction, whereas violations of imperfect duties can indeed be conceived in this way but cannot be willed to be universal laws without self-contradiction. This inability seems to introduce a definite teleological consideration. One of Kant's examples of a violation of an imperfect duty is refraining from helping someone who is in need. As Henry Sidgwick, author of a fundamental work, *The Methods of Ethics* (1874), pointed out, however, a man of proud self-reliance might reject being helped by others and could hence will, without self-contradiction, that persons in need not be helped. Kant's formal principle also fails to account even for such a perfect duty as the requirement that deformed infants not be exposed to die.

TELEOLOGICAL THEORIES OF NORMATIVE ETHICS

General considerations. Teleological theories of moral rightness emphasize the vital component that is lacking in deontological theories: goodness or value. An initial way of defining such theories is to say that, in contradistinction from deontological theories, they ground the moral rightness of actions not on nonvaluational features of the actions themselves or of their backgrounds or on formal relational principles but rather on the consequences of the actions evaluated in terms of their nonmoral goodness or badness.

In addition to their concern with good or bad consequences, another important feature common to all teleological criteria is their aggregative emphasis; *i.e.*, their view that the moral rightness of an action depends on its attaining as much good as is possible under the circumstances. This maximizing emphasis stands in contrast to the distributive concern of formal deontological theories.

Unclassified criteria of rightness. Whereas this definition is in many respects appropriate, it has the unfortunate result that two important criteria of the moral rightness of actions are left unclassified, because they fit under neither the deontological nor teleological criteria so far defined.

Intrinsic goodness

One omitted criterion is that which regards actions as morally right not because of their good consequences but rather because of their inherent or intrinsic goodness. Many hedonists (proponents of a pleasure-based ethic), for example, have held that an action is morally right if doing it is in itself pleasant, regardless of the further consequences of doing it. To be sure, there is often a further prudential emphasis on measuring the action's immediate pleasure against its further consequences (which may be painful). Nonetheless, not all nondeontological criteria of moral rightness are consequentialist: the inherent value of experiencing pleasure, exerting power, or "doing what comes naturally" is sometimes held to be the determining factor of the moral rightness of actions. Such criteria may be called axiological: in contrast to deontological criteria, they make goodness or value the determinant of rightness; but, in contrast to teleological criteria, they constitute intrinsic goodness or value. Indeed, it may well be argued that, if there were no intrinsic goods, there could be no consequential goods; for an action is consequentially good only if it produces what is intrinsically good. But such intrinsic goodness may belong to certain actions themselves.

Goodness of act or its consequences

A second criterion omitted from the above classification is that which regards actions as morally right on the basis either of their own moral goodness or of the moral goodness of their consequences. It may be objected that to make moral rightness depend on moral goodness would be circular. There need not be any circularity, however. One reason is that moral rightness pertains to actions, whereas moral goodness pertains to character. A second reason is that, whereas the moral goodness of character may be defined in terms of the moral rightness of actions performed by persons having that character, it may also be defined in other ways; *e.g.*, as the disposition to maintain a certain balance between the well-being of oneself and of others or among various aspects of one's own psyche. These dispositions, like all dispositions, will be exercised in actions. Some philosophers have held, however, that the correct sequence is that the moral rightness of actions is to be determined by their emerging from or by their fostering such dispositions, rather than that the dispositions' moral value is to be determined by their emerging in certain actions. An important, though controversial, example of the use of such a criterion is found in Plato's *Republic*, in which he argues that what makes actions just or morally right is that they either foster or emerge from a harmonious soul in which the rational part rules the other parts for the good of the whole. Moral theories that emphasize the agent's rational autonomy as the criterion of the moral rightness of his actions are also of this sort.

Distinction by the nature of the goods. Teleological theories are differentiated from one another according to what they say is (1) the nature of the goods or values to be maximized and (2) the locus of the goods; *i.e.*, the persons or groups whose good is to be considered or for whose interests the goods are to be maximized. To a

certain extent, the answer to (1) helps to determine the answer to (2), for it is sometimes held that only certain kinds of persons are capable of attaining certain kinds of goods. This is especially true with aristocratic theories such as those of Plato, Aristotle, and Nietzsche, who assert that only the philosopher-kings or the theoretic philosophers or the supermen are capable of having, respectively, the goods of a fully harmonious soul, of theoretic contemplation, and of power. The same point also holds for various nonaristocratic but still exclusivist theories such as that of Augustine and other Christian philosophers who held that only Christians can attain the highest good, which is the vision of God. On the other hand, when the good is identified with such a universally attainable experience as that of pleasure, such limitations of its locus are less inevitable.

In addition to the goods so far mentioned, other goods that have figured prominently in teleological theories include happiness (in some more or less specific sense), knowledge, love or personal affection, aesthetic activity, and various traits of character. Correspondingly, teleological theories, especially utilitarian ones, have been called "hedonistic," "eudaemonistic" (from the Greek word for happiness), "ideal" (in which the good is held to be of various sorts), and so forth.

Distinction by the locus of the goods. The two extremes of teleological theories, so far as concerns the locus of goods, are egoism and utilitarianism, which hold, respectively, that it is the good only of the agent himself or the good of all mankind (or even of all sentient beings) that determines whether actions are morally right. Intermediate between these extremes are theories holding that the goods that ought to be maximized pertain to more persons than the agent himself and yet not to everyone. They include doctrines that say that the goods in question should pertain only or primarily to persons belonging to a certain religion, race, nation, tribe, economic or social class, intellectual or political elite, or some other restricted group. In many obvious respects, such doctrines have been far more practically influential than universalist utilitarian theories, even if not more than egoistic theories.

Locus in the agent or in all mankind

Egoism. Ethical-egoist theories are of two main kinds. The individual egoist holds that the duty of all persons is to maximize the good of only one person, himself. The general egoist holds that the duty of each person is to maximize his own respective good. While individual egoism may be upheld by some persons, it is difficult to think of plausible arguments in its favour. If it appeals to certain characteristics of the one individual whose good is to be maximized, then it must admit that the good of any other persons who have those characteristics should also be maximized. If it appeals only to characteristics that are really unique to the one individual in question, then it is difficult to see how these characteristics could be relevant to the allocation of goods; the difficulty here is the same as that considered above concerning the criterion of relevant similarities in connection with the principle of universalizability. If the individual egoist appeals to no characteristics of any person, then his doctrine is left with no justification at all.

These considerations also apply, with appropriate adaptations, to at least some of the various particularistic theories mentioned above, which single out some favoured religion, race, or nation as the group whose good alone ought to be furthered.

Ethical and psychological egoism

General egoism draws much of its support from psychological egoism, the doctrine that each person in fact pursues or seeks to maximize his own respective good. The support may be merely contingent, or it may be logical. In the latter case, the implicit premise is that each person ought to pursue what he in fact does pursue. This premise, however, may be interpreted as meaning either (1) that men's goals, whatever they are, are always right (which seems hardly plausible) or (2) that it is impossible for persons to pursue anything other than what they in fact pursue, and each person thus pursues his own self-interest because he cannot pursue anything else.

Such a deterministic premise incurs at least three difficulties. First, if men cannot help doing something, there is no point in saying that they ought to do it, for ought implies both can and may not. Hence, psychological egoism, interpreted in this deterministic way, would make pointless both ethical egoism and any other ethical theory of obligation. Second, the deterministic thesis, in general, requires more argument than it has yet received. Third, psychological egoism, interpreted in this deterministic way, attains its counterintuitive result by illicitly moving from the premise that the purpose of any action is the agent's own, to the conclusion that the purpose is always the agent's own good; *i.e.*, the argument confuses the following two assertions: (1) each person in acting always tries to get something he himself wants; and (2) each person in acting always tries to get something he wants for himself or wants for his own interest. But, whereas (1) is true, (2) is false. The doctrine (1) that an agent always has an interest in whatever he does is quite distinct from the doctrine (2) that whatever an agent does is always for his self-interest.

Charge of self-contradiction and rejoinders

It has often been held that the ethical egoist incurs self-contradiction. The argument is that, when there is conflict of interest between two persons A and B, the ethical egoist, holding as he does that each person ought to maximize his self-interest, must hold both that A ought to win at the expense of B and that B ought to win at the expense of A—obviously mutually exclusive outcomes.

Several ways of avoiding this self-contradiction have been suggested. One is that the ethical egoist, in conflicts of interest, will hold not that each person ought to win but only that each person ought to try to win. This solution, however, still leaves a latent inconsistency, and it shifts the emphasis from each agent's achievement of good for himself to each agent's mere attempt at achievement, which weakens the original contention of ethical egoism. Another possible solution is to relativize the ethical egoist's "ought" judgments. What he then says is that A ought-from-the-standpoint-of-A to win, and B ought-from-the-standpoint-of-B to win. Since the hyphenated "oughts" in the two judgments do not have entirely the same meaning, the two judgments are not contradictory to one another. The relativistic interpretation, however, avoids incurring inconsistency only at the price of ceasing to be a theory of what is morally right.

Utilitarianism. Utilitarian theories, in contrast to those of egoism, are universalist; they hold that what a man ought to do is to promote the maximum good for everyone, in other words, the general good.

Good-aggregative versus locus-aggregative views. As it is sometimes formulated, however, utilitarianism raises an ambiguity with the phrase "the general good," in that the emphasis may be either (1) on the greatest good, no matter to whom it accrues, or (2) on distributing the good over as many sentient beings as possible. On the first interpretation the utilitarian might not be concerned at all with whose good is to be promoted: he would be indifferent to its locus, insisting only that good ought to be maximized in the universe, regardless of who gets it. On the second interpretation he would be concerned with its locus, insisting that the good of everyone (or of every sentient being) should be taken into account and maximized so far as possible. Thus, the first type of utilitarian would be concerned only with the total amount of good, whereas the second type would be concerned also or even mainly with how many persons get it. These two types may be called, respectively, "good-aggregative" and "locus-aggregative" utilitarianism. Classical British utilitarians, such as Jeremy Bentham and John Stuart Mill, though not always fully consistent, seem to have been more locus-aggregative than good-aggregative; for the formula that they upheld was "the greatest good of the greatest number."

Possible conflicts and deficiencies

There are two possible conflicts, however, between these two varieties of utilitarianism and, hence, between the two parts of the formula just presented. If one maximizes the locus of goods, spreading them over as many persons as possible ("the greatest number"), then one may decrease the total amount of good. This may be so

for several reasons: by maximizing the locus, one reduces the proportionate share and, hence, possibly the incentives of the more industrious persons, who would work harder if their proportionate share were greater. Some persons also may have a more intense appreciation of certain goods (and hence derive more good) than other persons to whom the goods are distributed. Obviously, many contingent factors, from the economic to the psychological, must enter into such calculations. By the same token, maximizing the total amount of good may require decreasing the amounts of the shares of some persons or even not giving any shares to some persons.

Another point to be noted is that on neither interpretation of utilitarianism, good-aggregative or locus-aggregative, is any specific pattern of distribution required. Even the locus-aggregative type says only that the number of persons or beings receiving the goods must be as great as possible; but this condition is consistent with a markedly unequal distribution of the goods. If, among a total population of five persons, *A*, *B*, *C*, *D*, and *E*, there is a total amount of 15 units of good to be distributed, the good-aggregative utilitarian would be satisfied to give all 15 units to *A*, leaving no units for the others, whereas the locus-aggregative utilitarian would insist that all five persons must have some units of good, but he would be satisfied with such a distribution as 7, 4, 2, 1, 1 or even 11, 1, 1, 1, 1. It is true that Bentham and Mill upheld the dictum, "Everybody to count for one, nobody for more than one." But this formula, as these philosophers interpreted it, provides not for equal distribution of goods among persons but only for impartial or indifferent consideration of persons as possible loci for the distribution of goods. As Mill explicated the formula, it means that "Equal amounts of happiness are equally desirable, whether felt by the same or by different persons." But this is compatible with all three of the unequal distributions indicated above, because, for example, the distribution of 15 units for *A* and none for the others distributes an "equal amount of happiness" and is hence "equally desirable"; with the distribution of 7 units for *A* and 4, 2, 1, 1, respectively, for the other persons. For, as between these two distributions, there are distributed "equal amounts of happiness"; *i.e.* the total amounts of good or happiness in the two cases are equal. In other words, the emphasis falls on aggregative not distributive equality.

Problems of distributive justice. The good-aggregative versus locus-aggregative conflict indicates what is probably the most serious problem confronting utilitarianism: the fact that neither version makes specific provision for distributive justice or for deciding the proportion in which goods are to be distributed among persons, as against the aggregative questions of how either the total amount of goods or the total number of persons receiving the goods are to be maximized. It is true that utilitarians from Hume through Bentham to Henry Sidgwick held, on the ground of the diminishing marginal utility of wealth, that, the more nearly equal is the distribution of wealth in any society, the greater is the total amount of happiness in the society. Even here, however, the determining criterion was the maximizing of good or happiness; equal distribution was valid only as a means to this end and was given no consideration as an independent criterion in its own right. In addition, the utilitarians introduced so many qualifications, based on the possible unfortunate consequences for the total amount of happiness that would follow upon the equalizing of wealth, that they eventually assumed a markedly inegalitarian position.

Ways of formulating distributive justice

Philosophers have suggested various ways of accommodating the claims of distributive justice within utilitarianism. Some hold that an equal distribution of goods is itself a good, so that, of two otherwise equal amounts of good, the amount that has equal distribution is better overall than the amount that is unevenly distributed. This suggestion illustrates the difficulties both of distinguishing between moral and nonmoral goods and of providing noncontroversial criteria of goods. Will persons who regard pleasure, wealth, or power as goods, over the distribution of which they contend, also regard equal distribution as itself a good; and, if so, why should they?

Another suggestion has been made by Nicholas Rescher, a contemporary U.S. philosopher, to the effect that, of two distributions, the one to be preferred is the one that has the higher effective average. The concept of the effective average is defined as the average number of shares had by the population minus one-half the standard deviation from the average. (The standard deviation in any group of numbers is determined by adding the squares of the numbers' deviation from the average, dividing the sum by the number of items, and then taking the square root of the quotient.) Taking the example of 15 units of good to be distributed among five persons, the first distribution (15, 0, 0, 0, 0) would have an effective average of 0, the second (7, 4, 2, 1, 1) would have an effective average of 1.86, and an equal distribution would have an effective average of 3. While Rescher presents some necessary qualifications of his principle, it has the difficulty that the same effective average could characterize distributions that are different in important respects, such as the degree to which they approach equality.

Other distributive problems

"Act" versus "rule" utilitarianism. Utilitarianism is also confronted by other distributive problems closely related to the problem of justice. It justifies the violation of important moral rules, such as those prohibiting the killing of innocent persons or the breaking of promises, whenever the consequences of violating these rules will be better overall than the consequences of obeying them. It ignores the fact that the recipient of one's obligatory actions may be highly particularized: if the same amount of good will be achieved by giving money to *A* or to *B*, then utilitarianism is indifferent as to who receives the money, even though, in fact, because of a previous transaction, the money is owed to *B*. More generally, utilitarianism ignores the fact that actions such as torturing innocent children are wrong on principle, regardless of their consequences. By making all moral obligations matters of calculation of consequences, utilitarianism distorts the nature of at least some basic moral judgments.

To deal with these difficulties, some philosophers have distinguished between "act" utilitarianism and "rule" utilitarianism. According to act utilitarianism, to determine whether an action is right or wrong, one directly considers its overall consequences. According to rule utilitarianism, to determine whether an action is right or wrong, one considers it in the light of some relevant justified moral rule; and, to determine which moral rules are justified, one considers the overall consequences of men's generally acting in accordance with that rule as against its alternatives. Thus, in order to determine whether to keep or break a promise to suit his convenience, an act utilitarian would consider whether keeping or breaking that promise would have better overall consequences; but the rule utilitarian would consider whether the general observance of the rule that men may break their promises to suit their convenience would have better overall consequences than the general observance of the rule that men should keep their promises regardless of whether it suits their convenience.

Philosophers are still in disagreement over whether rule utilitarianism has different implications for action than has act utilitarianism when the rules in question are specified in sufficient detail. Even if the implications are different (so that the act utilitarian may break his promise but not the rule utilitarian), the question still remains: "Why should the rule utilitarian obey a moral rule when in the particular case his obeying it will do more harm than good?" The charge of "rule-worship" has been levelled against the rule utilitarian, for whom, it seems, there implicity enters a supplementary, non-utilitarian consideration of fairness—of not taking advantage of other persons' observance of moral rules, while flouting them oneself, or of not advocating for others rules one does not accept for oneself. Apart from the consideration of the consequences of moral rules, however, rule utilitarianism copes no better than does act utilitarianism with the specific problem of distributive justice.

IV. Good reasons theories

Mixed meta-ethical and normative theories

In addition to the deontological and teleological normative ethical theories that have been considered above, some contemporary philosophers have presented theories of a mixed character. While their theories have obvious similarities to one or another of the normative ethical doctrines, they are presented as being logical or conceptual, as primarily explicating the meanings of such words as moral or ethical or ought, as well as the nature of moral reasons or moral reasoning. Thus, while the theories are presented as mainly or exclusively meta-ethical, their authors derive substantive normative ethical conclusions from them.

Material reasons. These theories focus on the question of logically good reasons in ethics, reasons of two kinds: material and formal. The material kinds of reason have a definite moral content and take one normative ethical position as against another; yet, the philosophers who propound them set them forth as deriving from definitions of what constitutes ethical or moral reasons in general. Thus, Stephen Toulmin, a British-American philosopher of science and of ethics, presents as his main problem to inquire into "the kinds of change in behaviour at which reasoning must be aimed, if we are to call it 'moral' or 'ethical' "; and his solution is that "the 'function' of ethics" is "to correlate our feelings and behaviour in such a way as to make the fulfillment of everyone's aims and desires as far as possible compatible," so that "what makes us call a judgment 'ethical' is the fact that it is used to harmonise people's actions. . . ." It is difficult to avoid the suspicion that a substantive view of moral goodness or rightness as consisting in social harmony is here being set forth as part of the very meaning or definition of ethics or morality. On this view, it becomes logically impossible to refer to ethical egoism as an ethical theory or to the various nationalistic, racialist, elitist, and other particularistic moralities. The view seems to settle by definition what preceding moral philosophers regarded as the central problem for normative ethical argument: how it can be established that one system of normative ethics is rationally superior to its rival systems. Similar considerations apply to the statement of Kurt Baier, a U.S. moral philosopher, that "moral rules must be for the good of everyone alike," especially as this is presented as part of the definition of "the moral point of view" and, hence, also of "moral reasons."

Mixed formal and material reasons. The other kind of reason presented by the recent writers who make "moral reasons" central to their theories is formal, consisting in the principle of universalizability discussed above. Since this principle taken by itself gives no moral results, the "good reasons" writers combine it with various material considerations. Marcus G. Singer, another U.S. moralist, combines it with the negative utilitarian consideration of the bad consequences that would ensue if one were to perform certain kinds of actions. R.M. Hare combines it with the consideration of men's inclinations, consisting mainly in their self-interested desires. The result of this combination in Singer's doctrine yields a line of argument that says that, if the consequences of everyone's doing a certain kind of action would be disastrous, then no one ought to do that kind of action. The corresponding result in Hare's doctrine is a line of argument that may be divided into three steps: First, a person holds that he ought to perform a certain kind of harmful action toward someone else because the latter has some quality. Second, he recognizes that, if he, the prospective agent, were to have that same quality, then it would be right for him to have to undergo that same kind of harmful action himself. Third, since, however, his inclinations are against such undergoing, he comes to judge that he ought not to perform that action on his intended victim.

Bad consequences and human inclinations

Each of these positions makes suggestive points about the logic of moral argument. One of their difficulties, however, is that, in each case, the material considerations that they combine with the formal one are contingent on men's variable beliefs or inclinations. Men may differ in their beliefs about what consequences are disastrous, and they may also differ in the extent to which they are willing to undergo counterinclinational actions. The doctrines of Singer and Hare hence provide no way of arguing against the person who denies that certain consequences, such as colonial war, are disastrous or who persists in certain judgments and actions even when their application to himself would go against his own self-interested inclinations. This means, for example, that the fervent nationalist or racist cannot be proven to be morally wrong by the above kinds of argument.

The normative ethical theories that have been discussed in the last two sections have obvious ramifications for the social problems of our time. The issues of race, poverty, war, crime, ecology, and so forth impinge on some of man's most important goods and bear vitally on the duties of distributive justice and on corresponding rights. The moral philosopher's essential contributions to these problems comprise both clarification and justificatory argument. The contributions have important implications for individual action and social policy.

BIBLIOGRAPHY

Anthologies: R.B. BRANDT (ed.), *Value and Obligation* (1961), readings from both classical and contemporary sources; PHILIPPA FOOT (ed.), *Theories of Ethics* (1967), an excellent collection of recent writings; A.I. MELDEN (ed.), *Ethical Theories: A Book of Readings*, 2nd ed. (1967), mainly classical writings, ending with Prichard; W. SELLARS and J. HOSPERS (eds.), *Readings in Ethical Theory*, (1952), the best collection of writings in meta-ethics from the first half of the 20th century; J.J. THOMSON and G. DWORKIN (eds.), *Ethics* (1968), a copious, excellent collection of recent articles.

Histories: H. SIDGWICK, *Outlines of the History of Ethics for English Readers*, 6th ed. (1931), excellent; A.C. MACINTYRE, *A Short History of Ethics* (1966), provocative, philosophically rewarding, selective in coverage and viewpoint.

Texts: R.B. BRANDT, *Ethical Theory: The Problems of Normative and Critical Ethics* (1959); W.K. FRANKENA, *Ethics* (1963); J. HOSPERS, *Human Conduct: An Introduction to the Problems of Ethics* (1961).

Classical Works (arranged chronologically): PLATO—All of Plato's writings are relevant to ethics, especially *The Republic* (the most basic), *Euthyphro*, *Protagoras*, *Gorgias*, and *Philebus;* ARISTOTLE, *Nicomachean Ethics* and *Politics;* CICERO, *De finibus bonorum et malorum* (*On the Ends of Goods and Evils*), and *De officiis* (*On Duties*); W.J. OATES (ed.), *The Stoic and Epicurean Philosophers* (1957); ST. AUGUSTINE, *De moribus ecclesiae catholicae* (*On the Morals of the Catholic Church*), *De civitate Dei* (*The City of God*), and the *Enchiridion;* PETER ABELARD, *Scito teipsum* (*Ethics, or Know Thyself*); ST. BERNARD OF CLAIRVAUX, *The Steps of Humility;* THOMAS AQUINAS, *Summa theologica* (Eng. trans., 22 vol., 1912–36); DESCARTES, *Les Passions de l'âme* (1649; *The Passions of the Soul*); PASCAL, *Pensées* and *Lettres provinciales* (1656–57; Eng. trans., 1941); THOMAS HOBBES, *Leviathan* (1651); SPINOZA, *Ethics* (1677; Eng. trans., 1927); L.A. SELBY-BIGGE (ed.), *British Moralists*, 2 vol. (1897); and D.D. RAPHAEL (ed.), *British Moralists, 1650–1800*, 2 vol. (1969), two excellent collections; JOSEPH BUTLER, *Fifteen Sermons* (1726); JEAN-JACQUES ROUSSEAU, *Discours sur les sciences et les arts* (1750; Eng. trans., *Discourse on the Sciences and Arts*, 1946); THOMAS REID, *Essays on the Active Powers of Man* (1788); DAVID HUME, *Enquiry Concerning the Principles of Morals* (1751), and *A Treatise of Human Nature*, books 2 and 3 (1739–40); IMMANUEL KANT, *Eine Vorlesung über Ethik* (1924; Eng. trans., *Lectures on Ethics*, 1930), *Grundlegung zur Metaphysik der Sitten* (1785; trans. by L.W. BECK as *Foundations of the Metaphysics of Morals*, 1949), and *Kritik der praktischen Vernunft* (1788; Eng. trans., *Critique of Practical Reason*, 1956); JEREMY BENTHAM, *Introduction to the Principles of Morals and Legislation* (1789); ARTHUR SCHOPENHAUER, *Über der Grundlage der Moral* (1840; Eng. trans., *The Basis of Morality*, 1903); SOREN KIERKEGAARD, *Enter eller* (1843; Eng. trans., *Either/Or*, 2 vol., 1944), and *Frygt og Baeven* (1843; Eng. trans., *Fear and Trembling*, 1939); WILLIAM WHEWELL, *Elements of Morality, Including Polity*, 2 vol. (1845); AUGUSTE COMTE, *Système de politique positive*, 4 vol. (1851–54; Eng. trans., *The System of Positive Polity*, 4 vol., 1875–77); KARL MARX, *Economic and Philosophic Manuscripts* (1932), *Die deutsche Ideologie* (written 1845–46, published 1932; Eng. trans. *The German Ideology*, 1933), and *Die heilige Familie* (1845; Eng. trans., *The Holy Family*, 1956); FRIEDRICH ENGELS, *Herr Eugen Dührings Umwälzung der Wissenschaft* (1878; Eng. trans., *Herr Eugen Dühring's Revolution in Science*, 1934), see ch. 9–11; HERBERT SPENCER, *The Principles of Ethics*, 6 pt. (1879–93); T.H. GREEN, *Prolegomena to Ethics*, ed. by A.C. BRADLEY (1883);

F.H. BRADLEY, *Ethical Studies* (1876); JOHN STUART MILL, *Utilitarianism* (1863); HENRY SIDGWICK, *The Methods of Ethics* (1874); FRIEDRICH NIETZSCHE, *Jenseits von Gut und Böse* (1886; *Beyond Good and Evil*), and *Zur Genealogie der Moral* (1887; *On the Genealogy of Morals*).

Twentieth century: (*Intuitionists*): G.E. MOORE, *Principia Ethica* (1903) and *Ethics* (1912); H.A. PRICHARD, *Moral Obligation* (1949), contains "Does Moral Philosophy Rest on a Mistake?"; W.D. ROSS, *The Right and the Good* (1930) and *Foundations of Ethics* (1939); A.C. EWING, *The Definition of Good* (1947). (*Naturalists*): WILLIAM JAMES, *The Will to Believe and Other Essays in Popular Philosophy* (1897); GEORGE SANTAYANA, *Winds of Doctrine* (1913) and *Reason in Society* (1906); JOHN DEWEY, *The Quest for Certainty* (1930), *Theory of Valuation* (1939), and with J.H. TUFTS, *Ethics* (1908); R.B. PERRY, *General Theory of Value* (1926) and *Realms of Value* (1954); C.I. LEWIS, *An Analysis of Knowledge and Valuation* (1946); P.B. RICE, *On the Knowledge of Good and Evil* (1955); ABRAHAM EDEL, *Ethical Judgment* (1955); H.B. VEATCH, *For an Ontology of Morals* (1971). (*Noncognitivists*): A.J. AYER, *Language, Truth, and Logic*, 2nd ed. rev. (1946); C.L. STEVENSON, *Ethics and Language* (1944); R.M. HARE, *The Language of Morals* (1952); P.H. NOWELL-SMITH, *Ethics* (1954). (*Reason in ethics*): S.E. TOULMIN, *An Examination of the Place of Reason in Ethics* (1950); K. BAIER, *The Moral Point of View*, new ed. (1965); M.G. SINGER, *Generalization in Ethics* (1961); R.M. HARE, *Freedom and Reason* (1963). (*Linguistic analysis relevant to ethics*): C.K. OGDEN and I.A. RICHARDS, *The Meaning of Meaning*, 10th ed. (1956); LUDWIG WITTGENSTEIN, *Philosophical Investigations* (1953); J.L. AUSTIN, *How to Do Things with Words* (1962); J.R. SEARLE, *Speech Acts* (1969). (*Phenomenology and Existentialism*): N. HARTMANN, *Ethik* (1926; Eng. trans., 3 vol., 1932); M. SCHELER, *Wesen und Formen der Sympathie* (1931; Eng. trans., *The Nature of Sympathy*, 1954); M. MANDELBAUM, *The Phenomenology of Moral Experience* (1955); JEAN-PAUL SARTRE, *L'Être et le néant* (1943; Eng. trans., *Being and Nothingness*, 1956), *L'Existentialisme est un humanisme* (1946; Eng. trans., *Existentialism and Humanism*, 1948), and *Saint Genet* (1952; Eng. trans., 1963); ALBERT CAMUS, *L'Homme revolté* (1951; Eng. trans., *The Rebel*, 1954); SIMONE DE BEAUVOIR, *Pour une morale de l'ambiguité* (1947; Eng. trans., *The Ethics of Ambiguity*, 1949). (*Deontology*): In addition to writings by Prichard, Ross, and Ewing cited above under "Intuitionists," see: C.I. LEWIS, *The Ground and Nature of the Right* (1955); C. PERELMAN, *De la justice* (1945; Eng. trans., *The Idea of Justice and the Problem of Argument*, 1963); JOHN RAWLS, *A Theory of Justice* (1971); ALAN GEWIRTH, *Moral Rationality* (1972). (*Utilitarianism*): In addition to writings by G.E. Moore cited above under "Intuitionists," see: H. RASHDALL, *The Theory of Good and Evil*, 2nd ed., 2 vol. (1924); M. BAYLES (ed.), *Contemporary Utilitarianism* (1968), a collection of essays; D. LYONS, *Forms and Limits of Utilitarianism* (1965); N. RESCHER, *Distributive Justice: A Constructive Critique of the Utilitarian Theory of Distribution* (1966).

(Al.Ge.)

Ethiopia

Ethiopia is an independent country in eastern Africa. It includes the former Italian colony of Eritrea, along the Red Sea coast, which was federated with Ethiopia in 1952 and became a province in 1962. The country shares frontiers with The Sudan, to the north and west, with Afars and Issas, to the northeast, with Somalia, to the east, and with Kenya, to the south. It has 628 miles (1,010 kilometres) of coastline on the Red Sea, to the northeast. Its total area is 471,800 square miles (1,221,900 square kilometres). By an estimate of 1975 the population was 27,030,400. The national capital is Addis Ababa ("New Flower").

Ethiopia has had a continuous existence as a nation—though with highly variable frontiers—for no less than 2,000 years. For most of this period its contacts with the rest of the world were limited and intermittent. Ethiopia is one of the developing nations of the world; its backward economy is based almost exclusively on agriculture. Until the military coup d'etat of 1974, its political structure had the trappings of a constitutional monarchy superimposed on a traditional autocracy. Despite this, Ethiopia had taken a leading role in the politics of modern Africa, largely because of its last emperor, Haile Selassie I (1892–1975).

The emergence of the country from its long isolation has brought into sharp focus the themes of paradox and diversity that are inherent in so many aspects of Ethiopian life. Ethiopia, a country that embraced Christianity as early as the second half of the 4th century, was known to Christendom throughout the centuries as a lonely outpost of Christianity in a sea of paganism and Islām. Yet a large minority of its present population is Muslim. Several types of animistic religions and even an ancient form of Judaism are also to be found. While Ethiopia's existence as a nation-state is one of the longest in the world, its population is linguistically, culturally, and to some extent racially so diverse that it has aptly been described as "a museum of peoples" (in *Etiopia e genti d'Etiopia* [1937] by Carlo Conti-Rossini). (See the city article ADDIS ABABA; for associated physical features, see EAST AFRICAN MOUNTAINS and NILE RIVER; for historical aspects, see ETHIOPIA, HISTORY OF; for religion, see EASTERN CHRISTIANITY, INDEPENDENT CHURCHES OF; for language, see HAMITO-SEMITIC LANGUAGES; for cultural aspects see ETHIOPIAN AND SOMALIAN CULTURES.)

A country of paradox

THE LANDSCAPE

The natural environment. *Relief features.* The rock foundation of the Ethiopian region forms part of the Precambrian (from 570,000,000 to 4,600,000,000 years old) crystalline block underlying most of Africa. On top of it, younger limestones and sandstones are capped by successive flows of lava, which have accumulated to a thickness of several thousand feet in some areas. Evidence of volcanic activity is still to be found in the form of a few live volcanoes and numerous hot springs.

An important feature of the relief of the region is the East African Rift System—a mammoth cleavage that runs from north to south and cuts through the central Ethiopian Plateau, dividing it into western and eastern sections. The three major relief regions of the country, which tend to influence climatic conditions, vegetation, human settlement, and drainage, are thus, first, the Western Highlands; second, the Eastern Highlands; and, third, the Rift Valley and Western Lowlands.

The Western Highlands lie at altitudes between 8,000 and 12,000 feet (2,400 to 3,700 metres). The northern end of this region is in northern Eritrea, where erosion has cut down to the basement rock. In the Eritrean Highlands, just to the south, some remnants of basaltic rock can be found. The Tigre Plateau of western Tigre Province for the most part has been eroded down to the layer of sandstones and limestones. In the Semien Mountains and neighbouring massifs to the south, basalt caps the plateaus and *ambas* (small, flat-topped tablelands). In the Semien Mountains, the highest peak in Ethiopia—Ras Dashen, or Dejen—rises to 15,157 feet (4,620 metres). The tablelands are surrounded by deep gorges and canyons. The Abbay (Blue Nile) River gorge runs south and west from Lake Tana and cuts nearly 7,000 feet below the general level of the plateau. The gorge isolates the Gojam Massif, to the north, from the Shewan Plateau, to the south. The Shewan Plateau is the widest area, with a general elevation of more than 7,000 feet in the Western Highlands. Like the areas to the north, it is cut by several deep gorges; the plateau surface itself is formed of basalt rock. Southwest of the Shewan Plateau is a region of lower elevation, which has been subjected to much greater erosion.

The rugged plateau region

The Eastern Highlands, like those in the west, have a largely basaltic cover except in the lower elevations and the river gorges. The highest area is a mountain chain that runs along the Rift Valley escarpment, sloping steeply on its eastern side. A number of high peaks include Mt. Encuolo, which rises to 14,144 feet (4,311 metres). Between the upper courses of the Genale and Shebele rivers, which flow southeastward, is an extensive basaltic plateau—the Bale Massif—which is mostly above 9,000 feet. To the northeast is the Harer Plain, which is less extensive and lies at about 1,000 feet above sea level. The plain continues into the Ogaden region, in the extreme southeast, where it is even more eroded and considerably lower. The southernmost part of the Eastern Highlands is the Sidamo-Borana Plain, which is similar to the Harer Plain in elevation but is more eroded in parts.

The Rift Valley region is dotted by a series of lakes, including Ziway, Abayata, Langano, Shala, Awasa, Abaya, and Chamo. The valley, at an elevation of about 5,000 feet, is bounded by an eastern and a western escarpment that are for the most part from 25 miles to 40 miles apart. To the north the valley widens into the Awash River Basin, where the Awash is joined by a number of tributaries. Volcanic activity is indicated in the Rift Valley by a number of hot springs and by the active volcanoes of the Danakil Plain, to the north. The desertic depression runs parallel to the Red Sea and contains salt lakes and the Kobar Sink, a depression that dips below sea level.

On the western fringe of the Western Highlands, in a thin crescent extending from western Eritrea to the northern shores of Lake Rudolf (on the Kenya border, to the south), lie the Western Lowlands, which are the eroded remains of the rim of the plateau.

Drainage. The Western Highlands slope generally toward The Sudan, and drainage is therefore westward. The eastern edge of these highlands constitutes a major watershed draining into the Rift Valley. Similarly, the western edge of the Eastern Highlands constitutes another watershed, with its eastern slopes draining in a southeasterly direction and its steep western face drained by small rivers into the Rift Valley. Since the Western Highlands are an area of high rainfall, the westward drainage is by far the most important, accounting for nearly half the total volume of water drained annually from the country.

Major rivers

The major rivers in this area are the Blue Nile (Abbay), the Tekeze, and the Baro. In the inland drainage area are the Awash River, which flows northward into Lake Abe, in the Danakil Plain, and the Omo River, which flows southward into Lake Rudolf. The Sagan River flows into Chew Bahir (Lake Stefanie). In the area of southeasterly drainage are the Webi, Genale, and Dawa rivers, which in Somalia join to form the Juba and reach the Indian Ocean, and the Shebele, which flows into Somalia but fails to meet the sea.

Soils. The basalt deposits in much of the highland area are rich in phosphorus and potash. The soils are highly fertile, because they are weathered but not leached. They are, however, deficient in calcium. The texture of these soils, mostly of a loam or silt–loam nature, is judged to be excellent. Phosphorus and potash are also to be found in the Rift Valley soils, but these are highly deficient in nitrogen and are alkaline in some areas. Phosphorus deficiency and acidity are problems in the soils of the northern Eritrean Highlands and the Tigre Plateau. By far the most fertile soils in Ethiopia are the rich alluvial deposits in the river valleys. The extensive valleys of the Blue Nile, Awash, and Omo offer the greatest promise for agriculture, but those of smaller rivers are equally rich.

Climate. Although Ethiopia is completely within tropical latitudes, it enjoys a generally temperate climate in the highland areas. The annual range of temperature is small because of its proximity to the Equator, but the average daily range is large because of its altitude. The average daily range for 15 stations is 24° F (13° C), from a minimum of 60° F (16° C) to a maximum of 84° F (29° C). The annual average temperatures of the highland and lowland regions differ sharply. Addis Ababa, at an altitude of 8,000 feet, has an annual average temperature of 60° F (16° C), while Mitsiwa (Massawa), at sea level, has an annual average temperature of 87° F (31° C).

The coldest month is usually January and the hottest usually May. On some parts of the highlands, however, the rainy season (*kremt*), with its peak in June and July, has a lowering effect on the temperature. In Addis Ababa, the maximum monthly average is 72° F (22° C) and the minimum 50° F (10° C). At Aseb, on the Red Sea coast, however, the maximum monthly average is 95° F (35° C) and the minimum 77° F (25° C).

The rainy season

Rainfall for most of Ethiopia is brought by southwesterly winds. In July the high-pressure system that builds up over the South Atlantic and the Gulf of Guinea causes an air movement toward the low-pressure centres of the Sahara and the Middle East. When the prevailing warm and moist winds reach the Western Highlands they are forced to rise, thereby cooling and losing their capacity to hold moisture, depositing it as rainfall from July to September on the highlands.

In January, northeasterly winds come from a high-pressure area in the dry Middle East, blowing toward a Central African low-pressure area. The northeast trades account for the intensity of the dry season (*bega*), in the generally drier northern parts of the country. Whatever moisture they pick up from the Red Sea is deposited on the slopes of the Eritrean escarpment. There is a short and mild rainy season (*belg*) in March and April. May is a short dry interlude, and June leads into the big rains. The major dry season for most of the country is, therefore, from October to February. In the southeastern lowlands there are two rainy seasons, one in spring and the other in autumn.

Local variations in rainfall are determined by the extremely rugged terrain, which interferes with the movements of major currents and causes more rain to be deposited on the windward slopes than on the leeward slopes. Generally, the amount of annual precipitation varies from almost 100 inches (2,500 millimetres) at Gore, in the southwest, to about 2 inches (50 millimetres) at Aseb, in the northeast. Even at the same location, however, both the amount and the time of rainfall may vary annually to such an extent as to involve serious risks in farming. In addition, the violent nature of the thundershowers is a contributing factor to the serious problem of erosion.

Vegetation and animal life. The two major vegetation regions are the wet and cool highlands and the generally drier and hotter lowlands. The Western Highlands are now mostly grasslands and farmland with occasional clumps of trees. Forests occupy about 7 percent of the total area of the country. The southwestern rain forests, found mainly in Kefa and Ilubabor provinces, are composed of large broad-leaved trees such as *Syzygium guineense* and *Cordia abyssinica* and a thick undergrowth of ferns, creepers, and bushes, including the coffee bush. At higher altitudes the subtropical forests are dominated by the yellowwood, an evergreen tree with a pulpy fruit. Still higher are the coniferous forests, dominated by juniper. The upper limit of trees is about 10,000 feet (3,000 metres); above this altitude are found only coarse grasses, giant lobelias, and other high-altitude vegetation.

In the lowlands the most widespread tree is the acacia, which occurs in all its numerous varieties. The exotic eucalyptus is common in inhabited regions. Where there is a relatively greater amount of rainfall—for example, in the southern part of the Rift Valley—there is rich savanna grassland with numerous trees in addition to the omnipresent acacia. The semi-desert and desert areas have scattered dwarf-acacia thornbushes, some tough grass, and succulent plants.

Indigenous animal species

Among the animal species unique to Ethiopia are the *walia* ibex (a type of mountain goat), found on the higher reaches of the Semien Mountains and numbering not more than 200; the mountain nyala (a kind of antelope) of the Arusi, Bale, and Chercher highlands, which number about 2,000; the *simien* fox; and the gelada monkey. The most widespread animals are the spotted and the striped hyena, various species of jackal, the wild dog, and the bat-eared fox. Antelope include the greater kudu, the oryx, the bushbuck, the bush duiker, the klipspringer, and the common waterbuck. Various species of wild pig (bushpig, warthog, giant forest hog) occur in large numbers. The vervet (a black-faced guenon monkey known for its various cries) and the anubis baboon (a gray or brown baboon, the male of which has a heavy mane) are plentiful and are a serious pest in farming areas. Many smaller animals, from the aardvark (a burrowing mammal that feeds on ants) to the white-tailed mongoose, are widespread.

The richest wildlife areas are the lower Omo Valley, the Gambela region of Ilubabor, and the southwest. Animals include lions, elephants, leopards (now quite rare), buffalo, zebras, giraffes, and black rhinoceroses (also

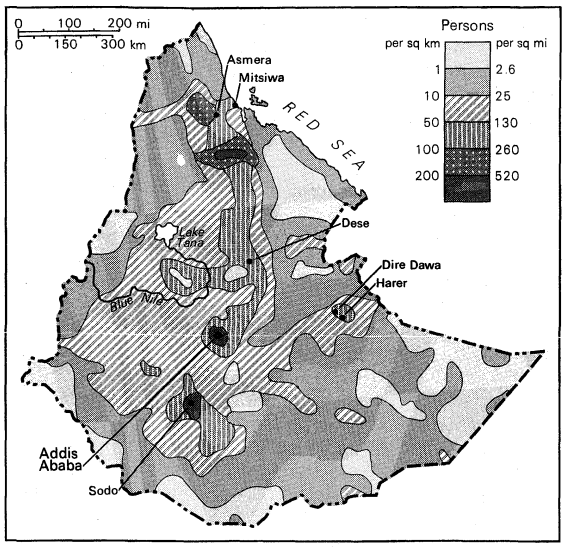
rare). Smaller animals hunted for their skins include the beautiful black-and-white colobus monkey, the serval, and the rare cheetah. Crocodiles inhabit most of the lakes and rivers at lower altitudes.

Rural and urban settlement. Most of the agricultural population lives in walled or fenced family compounds, which are scattered throughout the countryside. Buildings are round with conical thatched roofs in the south and west and rectangular with flat thatched roofs or peaked corrugated-iron roofs in the north. Fields are unfenced, and individual rights to their use are traditional; arguments over land use are a common feature of traditional judicial court procedure. Rural towns are market centres for their surrounding agricultural areas. The cattle breeders of the lowlands are nomadic tent dwellers.

Addis Ababa

The two major cities of Ethiopia are Addis Ababa (*q.v.*), the national capital, and Asmera. Founded in the 19th century, Addis (as it is called locally) occupies the centre of the Shewan Plateau. With a population of about 1,174,000, it is the larger of the two cities and has grown in haphazard fashion. Its modern buildings are surrounded by the traditional housing of the majority of its inhabitants.

Asmera, once a small rural village, developed into the capital of Eritrea, which was under Italian administration from 1890 to 1941. It has a population of about 307,000. In contrast to Addis, the city boasts well-planned streets and avenues and is heavily marked by European influences.



Population density of Ethiopia.

PEOPLE AND POPULATION

Ethnic origins. As there has been so much racial intermixing, racial identifications are of limited use in Ethiopia. Periodic migrations of Caucasoid peoples from across the Red Sea and of Negroid peoples from the West have produced a population that preponderantly manifests a blend of the two physical types. The primary sources of identification for the peoples of Ethiopia are language, religion, and other aspects of culture. The terms Amhara, Galla, and Tigre refer to linguistic and cultural groups rather than to races or tribes.

Languages. There are nearly 100 distinct languages, some of which have several dialects or less well-defined regional variations. The languages fall into the Semitic, Cushitic, and Nilotic language groups; a language group usually classified as Cushitic may be distinct enough to form an independent and coordinate family called Omotic.

Semitic languages

In terms of the historical, political, and cultural role of its speakers, the most important of the language groups is the Semitic. Semitic languages in Ethiopia fall into a northern group—Ge'ez, Tigre, Tigrinya (Tigriña)—and a southern group—Amharic (Amarinya), Gurage, Hareri, Gafat, and Argobba. Ge'ez, the language of the ancient Aksumite Empire, survives as a liturgical and literary language. Tigre is spoken by about 210,000 people in the western and northern lowlands of Eritrea. Tigrinya, with about 3,000,000 speakers, is the language of Tigre province and the highlands of Eritrea.

Amharic, the national language, first emerged into recorded history as the court language in the 13th century and has been the language of the centre of political power ever since. It is spoken as a native language by about 6,000,000 persons in the provinces of Begemdir and Simen, Gojam, Welo, and Shewa but is spreading rapidly and is spoken as a second language by an estimated 3,000,000 persons.

MAP INDEX

Political subdivisions
Arusi 8·00n 39·40e
Bale 6·20n 41·30e
Begemdir and Simen 13·10n 37·00e
Eritrea 15·20n 39·00e
Gemu Gofa 4·45n 37·00e
Gojam 11·00n 37·00e
Harer 8·00n 43·00e
Ilubabor 6·50n 35·20e
Kefa 6·50n 36·00e
Shewa 9·00n 39·00e
Sidamo 5·00n 39·00e
Tigre 13·40n 40·00e
Welega 9·40n 35·50e
Welo 11·50n 40·20e

Cities and towns
Abelti 8·10n 37·37e
Adaba 7·07n 39·20e
Adami Tulu 7·52n 38·40e
Addis Ababa 9·00n 38·50e
Adigrat 14·18n 39·31e
Adi Keyih 14·49n 39·23e
Adi Ugri 14·55n 38·53e
Adwa 14·10n 38·55e
Agere Hiywet 8·56n 37·56e
Akordat 15·35n 37·54e
Aksum 14·08n 38·48e
Alemaya 9·10n 42·30e
Ankober 9·30n 39·44e
Aseb 13·00n 42·45e
Asela 7·59n 39·08e
Asmera 15·20n 38·53e
Asosa 10·03n 34·32e
Awasa 7·02n 38·30e
Awash 8·59n 40·10e
Aysha 10·46n 42·37e
Bahir Dar 11·35n 37·28e
Bako 5·50n 36·40e
Bambesi 9·45n 34·38e
Barentu 15·04n 37·37e
Bedele 8·33n 36·23e
Begi 9·16n 34·33e
Bonga 7·17n 36·15e
Bure 8·15n 35·09e
Dangla 11·18n 36·54e
Debre Markos 10·20n 37·45e
Debre Sina 9·51n 39·50e
Debre Tabor 11·50n 38·05e
Debre Zeyt 8·48n 38·56e
Deder 9·22n 41·26e
Degeh-Bur 8·14n 43·35e
Dembi 8·05n 36·27e
Dembi Dolo 8·30n 34·48e
Deneba 9·50n 39·09e
Dese 11·05n 39·41e
Dire Dawa 9·37n 41·52e
Dodola 7·02n 39·07e
Domo 7·54n 46·52e
Gambela 8·18n 34·37e
Gedo 8·58n 37·27e
Gewani 10·16n 40·44e
Ginir 7·07n 40·46e
Giyon 8·30n 38·00e
Goba 7·02n 40·00e
Gonder 12·40n 37·30e
Gore 8·08n 35·33e
Harer 9·18n 42·08e
Hididola 5·52n 43·42e
Hosaina 7·38n 37·52e
Huwun 4·23n 40·08e
Imi 6·28n 42·18e
Jima 7·36n 36·50e
Karora 17·42n 38·22e
Kebri Dehar 6·47n 44·17e
Kelafo 5·40n 44·20e
Keren 15·46n 38·28e
Kibre Mengist 5·52n 39·00e
Kobo 12·11n 39·33e
Lalibela 12·02n 39·02e
Maji 6·11n 35·38e
Massawa, see Mitsiwa
Mega 4·07n 38·16e
Mekele 13·33n 39·30e
Mendi 9·50n 35·06e
Mersa Fatma 14·55n 40·20e
Metu 8·20n 35·36e
Mieso 9·15n 40·48e
Mitsiwa 15·38n 39·28e
Mizan Teferi 6·53n 35·28e
Mojo 8·38n 39·07e
Nazeret 8·32n 39·22e
Negele 5·20n 39·36e
Nekemte 9·02n 36·31e
Shashemene 7·15n 38·35e
Sodo 6·52n 37·47e
Tendaho 11·48n 40·52e
Teseney 15·07n 36·41e
Waka 7·07n 37·26e
Wonji 8·50n 39·00e
Yabelo 4·54n 38·05e
Yirga Alem 6·52n 38·22e

Physical features and points of interest
Abaya, Lake 6·20n 37·55e
Abbay, see Blue Nile
Abe, Lake 11·06n 41·50e
Abiyata, Lake 7·37n 38·35e
Akobo, *river* 7·48n 33·03e
Atbara, *river* 13·02n 36·08e
Awasa, *lake* 7·05n 38·25e
Awash, *river* 11·45n 41·05e
Barka, *river* 17·20n 37·30e
Baro, *river* 8·26n 33·13e
Batu, *mountain* 6·55n 39·46e
Blue Nile, *river* 11·12n 34·58e
Chamo, Lake 5·49n 37·35e
Chew Bahir, see Stefanie, Lake
Danakil Plain 12·25n 40·30e
Dawa, *river* 4·11n 42·06e
Encuolo, *mountain* 7·22n 39·20e
Eritrea, *historic region* 15·20n 39·00e
Fafen, *river* 6·07n 44·20e
Gash, *river* 15·10n 36·15e
Genale, *river* 5·43n 40·53e
Gilo, *river* 8·10n 33·15e
Gojam Massif, *mountains* 11·45n 38·00e
Guge, *mountain* 6·10n 37·26e
Harer Plain 8·00n 42·30e
Kobar Sink, *depression* 14·00n 40·30e
Langano, Lake 7·35n 38·48e
Ogaden, *physical region* 5·15n 43·30e
Omo, *river* 8·48n 37·14e
Ramlo, *mountain* 13·20n 41·45e
Ras Dashen, *mountain* 13·10n 38·26e
Red Sea 17·30n 40·00e
Rift Valley 8·00n 39·00e
Rudolf, Lake 3·30n 36·05e
Sagan, *river* 5·15n 37·15e
Shala, Lake 7·25n 38·30e
Shebele, *river* 5·00n 45·04e
Shewan Plateau 11·00n 37·15e
Sidamo-Borana Plain 4·45n 39·00e
Simen Mountains 13·00n 38·00e
Stefanie, Lake 4·40n 36·50e
Talo, *mountain* 10·44n 37·55e
Tana, Lake 12·00n 37·20e
Tekeze, *river* 14·15n 36·30e
Tigre Plateau 14·00n 38·30e
Webi, *river* 4·11n 42·09e
Weyib, see Webi
Ziway, Lake 8·00n 38·50e

Hareri (or Adere or Adare) is spoken by about 35,000 people within the walled city of Harer. Gurage, the southernmost extension of the Semitic languages, is in fact a cluster of about 12 distinct dialects rather than one homogeneous language. It is spoken to the south of Addis Ababa by nearly 1,000,000 persons.

Cushitic languages

Among the Cushitic languages, by far the most important is Gallinya. The number of native speakers, known as Gallas, is estimated at about 8,000,000, but the language is spoken as a second language by many smaller linguistic groups in the east, south, and west. Not surprisingly, this widely distributed language has considerable dialect variations.

The other important Cushitic language is Somali, which has about 1,500,000 speakers. Of the remaining 20 Cushitic languages, only three—Afar-Saho, Sidamo, and possibly Hadiyya—have more than 500,000 speakers.

The Omotic branch of the Cushitic languages (called the Western branch by some), chief among which is Walamo (600,000 speakers), is spoken in the southwest, by about 2,000,000 speakers altogether. The Nilotic languages are spoken in the western fringes of the country and have fewer than 40,000 speakers each, with the exception of Kunama, which is spoken by about 65,000 persons. English is taught as a second language in the school system, and Italian and Arabic are widely understood in the north.

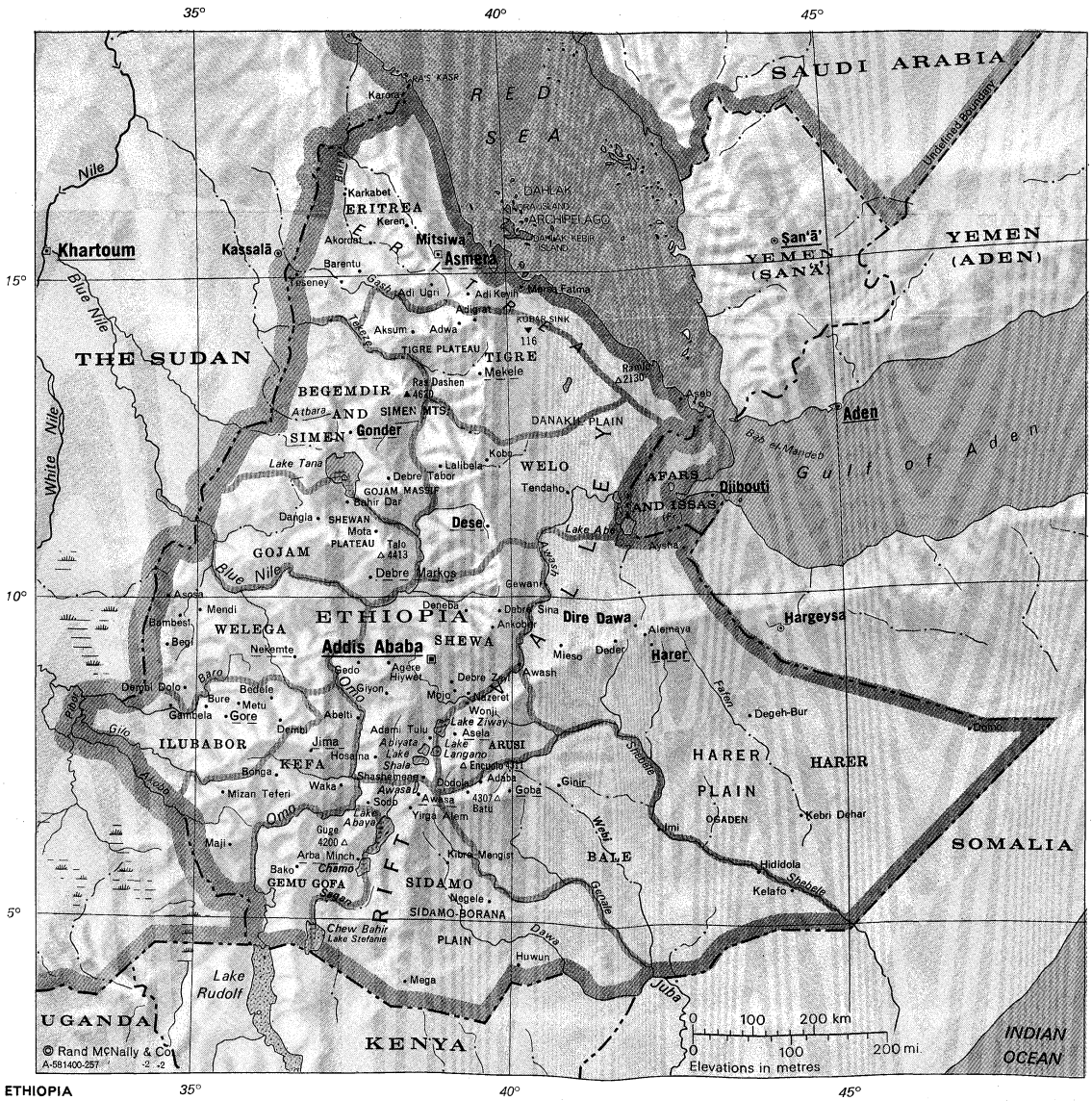
Religions. Christians are generally believed to constitute about half the population. An estimated 35 percent are Muslims, and the remainder are mostly animists; there is also a small number of Jews. The stronghold of Christianity is the northwestern highland region from central Eritrea to Shewa, while that of Islām is the lowland area and much of Kefa province. Animism prevails in the far south and in the southwestern foothills but is losing ground to both Christianity and Islām. The Falashas, pre-Talmudic Jews who were probably converted by Yemenite Jews, are found in small groups in the province of Begemdir and Simen.

The Ethiopian Orthodox Church

Christianity was introduced into Ethiopia in the 4th century, Islām was introduced in the 7th. Historically and culturally, Christianity has been and still is the dominant religion. Ethiopian Christianity had strong formal ties with the Coptic Church of Alexandria, which until 1959 provided the archbishop, or *abuna*, for the Ethiopian Orthodox Church. The Ethiopian Orthodox Church is the established church, and, until the coup of 1974, the emperor and his court were by law of this faith, as were most high government officials, although members of other faiths were not legally proscribed. The church has been viewed in Ethiopian history as the prime source and custodian of the cultural and literary heritage of the nation.

Islām, which includes widely distributed ethnic and linguistic groups, is less homogeneous than Christianity. It has two cultural focuses in Harer and Kefa provinces.

Demography. Demographic data on Ethiopia are limited and unreliable. The average population density is 57



ETHIOPIA

Ethiopia, Area and Population

	area		population*
	sq mi	sq km	1972 estimate
Provinces (*teklai gezatoch*)			
Arusi	9,100	23,500	853,000
Bale	48,100	124,600	708,000
Begemdir and Simen	28,600	74,200	1,356,000
Eritrea	45,400	117,600	1,948,000
Gemu Gofa	15,300	39,500	699,000
Gojam	23,800	61,600	1,750,000
Harer	100,300	259,700	3,359,000
Ilubabor	18,300	47,400	689,000
Kefa	21,100	54,600	1,693,000
Shewa	33,000	85,400	5,370,000
Sidamo	45,300	117,300	2,480,000
Tigre	25,400	65,900	1,829,000
Welega	27,500	71,200	1,269,000
Welo	30,700	79,400	2,460,000
Total Ethiopia	471,800†	1,211,900	26,461,000†

*No census has ever been taken. †Detail does not add to total because of rounding.
Source: Official government figures.

persons per square mile. Densities are about twice the average in the populous provinces of Arusi, Shewa, and Welo. The lowest density, 14 persons per square mile, occurs in Bale. The lowland areas are generally more sparsely populated than the highlands. About 90 percent of the population is rural, the most urbanized provinces being Shewa and Eritrea, which have urban populations of 22.3 percent and 21.5 percent, respectively. Birth and mortality figures are highly unreliable, and immigration and emigration are insignificant. The best estimate for the annual rate of population growth is approximately 2.5 percent.

THE NATIONAL ECONOMY

Ethiopia is among the less developed countries of the world. Many of the characteristics of underdevelopment, such as the predominance of the subsistence sector, the limited role of industry, the underdeveloped economic infrastructure (roads, water supplies, schools, and other facilities), the shortage of domestic capital, and the undiversified export structure, while shared with much of the rest of Africa, are nevertheless more pronounced in Ethiopia. The growth rate of the economy has remained well below the average annual increase in the gross domestic product (GDP) for Africa in general.

Natural resources. Mineral resources had not been accurately surveyed by the 1970s. Preliminary information, however, indicates that mineral deposits are probably considerable. The basement rocks contain valuable metals, the more important of which are gold, platinum, lead, tungsten, and copper. The sedimentary rocks may contain petroleum, but none has yet been discovered in quantities sufficient to make exploitation worthwhile. Limestone deposits are important, and the volcanic deposits of the Danakil Plain contain sulfur, sodium and potassium salts, gypsum, rock salt, and potash.

Mineral deposits

Biological resources include about 22,000,000 acres (9,000,000 hectares) of forest land, covering about 7 percent of the total area of the country. Trees are of varying quality for use as timber; eucalyptus is widely used for construction. Wildlife provides bases for the expanding tourist industry.

Many Ethiopian rivers have great potential for hydroelectric power, while the volcanic areas of the Rift Valley offer possibilities for the development of geothermal power. Accurate surveying of the water resources to enable the selection of suitable dam sites remains to be done. The only river basins surveyed by the 1970s were those of the Abbay (Blue Nile) and the Awash.

Sources of national income. *Agriculture, forestry, and fishing.* Agriculture is the mainstay of the Ethiopian economy. It provides a livelihood for 87 percent of the population, constitutes more than half of the GDP, and accounts for nearly all of the country's exports. Except on the few large-scale commercial farms, land is tilled by the ox-drawn plow or the hoe, and seed is sown by hand. Use of fertilizers is limited; natural manure is mostly used as fuel or as an ingredient in building material rather than as fertilizer. The yield is primarily used by the farmer himself, and only a small proportion of the produce reaches the markets.

More than half of the country's total acreage is devoted to pastures, although most of it is suitable for cultivation. The total of cultivable land and pastureland is about 207,814,000 acres (84,102,000 hectares), of which 25,699,000 acres (10,400,000 hectares) are cultivated. This provides an excellent man-to-land ratio. The fertility of the soil is generally considered to be high, and the climate is suitable for varied agricultural production.

Crops vary greatly, depending upon altitude and climatic conditions. The staple crop in the northern half of the Western Highlands is teff (*Eragrostis abyssinica*), a grain unique to Ethiopia. It grows best on the lower parts of the plateau, where maize (corn) and millet are also raised. Wheat and barley are widely grown on the highlands proper, and oats are raised on the highest parts of the plateau. Pulses (lentils and varieties of peas and beans) are also widely grown. In the southern half of the Western Highlands the staple food is a versatile plant known as ensete (or "false banana"; *Musa Ensete*); an increasing amount of maize and sorghum is also grown. In the west the gathering of wild coffee is increasingly supplemented by coffee plantations, and mixed farming with a wide variety of crops is practiced. In the lowlands and foothill areas, pastoralism is supplemented by seasonal cultivation of crops such as maize and sorghum. In the Danakil Plain, nomadic pastoralism prevails.

Coffee is the most important agricultural export. The coffee plant is native to Ethiopia and grows wild in the provinces of Kefa, Welega, Ilubabor, and Gemu Gofa. There is increasing cultivation of coffee in these provinces and in similar highland areas of Sidamo and Harer. Other important agricultural products are sugarcane, various oil seeds, vegetables, and fruits. There are possibilities of large-scale production of cotton on the lowlands. as exemplified by the plantation at Tendaho.

The native coffee plant

Next to coffee, perhaps the most important agricultural resource is livestock. Ethiopia has a livestock population of about 100,000,000 animals, more than one-fourth of which are cattle; they account for about 20 percent of the total cattle population in Africa. The rest are sheep, goats, poultry, and other domestic animals. Prospects for the expansion of the livestock industry are good because of the abundant grassland and adequate rainfall. The meat yield of Ethiopian livestock, however, is rather poor because of poor breeds and feeding methods.

Despite the considerable forestry potential, the contribution of this resource to the economy is small. This is largely a result of the lack of road or rail access to much of the forested area. Most of the forests are owned by the government. The state forests have remained untouched, domestic timber is expensive, and forest products are increasingly imported.

The role of fishing in the traditional economy is negligible. Commercial fishing is entirely restricted to the Red Sea coastal waters and involves little deep-sea fishing. Production fluctuates from year to year because of poor techniques and equipment and the near total dependence on foreign markets.

Fishing

Mining and quarrying. The contribution of mining and quarrying to the economy is also small. Mining activity includes the exploitation of the alluvial gold deposits at Kibre Mengist and the mining of manganese ore near Mitsiwa, salt in the Danakil region and southern Borana, and platinum in Welega. The exploitation of potash was given the highest priority in the 1970s.

Manufacturing. Manufacturing is little developed. It employs less than 1 percent of the population and contributes about 10 percent to the GDP. It is based almost entirely on the processing of agricultural raw materials and the production of light consumer goods. Most industry is located in the few major urban centres; about half is in Addis Ababa and most of the rest in and around Asmera, Nazeret, and Dire Dawa. The main products

are textiles, processed food, beverages, and shoes. The sugar factory at Wonji, in Shewa province, supplies domestic needs.

Handicrafts and small-scale industries produce about 5 percent of the gross domestic product. Activities include hand spinning and weaving, tailoring, carpentry, and the small-scale production of carpets, rugs, rope, twine, handmade shoes, other leather products, and jewelry. These activities are usually located in households and depend on family labour, but also included are such undertakings as cleaning coffee and grain and milling flour.

Energy. The total installed generating capacity in the mid-1970s was 185,000 kilowatts, of which two-thirds was hydroelectric and the rest thermal. At Fincha, on a tributary of the Abbay (Blue Nile) River, a major dam of nearly 100,000-kilowatt capacity was inaugurated in November 1973, increasing the installed capacity of the country by nearly 50 percent. But this and other exploitation, largely concentrated at three large dams on the Awash River, constitutes only a minute fraction of the hydroelectric potential, which is estimated to be 4,300,000 kilowatts.

Tourism. The animal life in the southwest, the majestic highland scenery, recreational fishing in the Rift Valley lakes, year-round bathing on the Red Sea coast, and places of historical interest such as Aksum, Lalibela, and Gonder could draw tourists in significant numbers. In order to encourage tourism, several national parks, game reserves, and bird sanctuaries have been established, historic cities restored, and hotels built.

Commerce and financial services. Wholesale and retail trade is, at least in urban areas, dominated by the local Italian, Arab, Greek, and Armenian communities. Christian highlanders have shown a traditional disinclination to trade, but other Ethiopians are now gradually displacing aliens in commerce. Ownership of dwellings, on the other hand, is almost completely in Ethiopian hands, because aliens are legally barred from owning land except by special dispensation.

Banking

The main banks are the Commercial Bank of Ethiopia (a government-controlled corporation), the Agricultural and Industrial Development Bank (which provides long-term and medium-term credit for agricultural and industrial development), and the National Bank of Ethiopia (which performs the services of a regulatory central bank). Until the coup of 1974, the private Addis Ababa Bank was also active in commercial banking, and branches of two Italian banks had been granted licenses to conduct business in Addis Ababa and other cities. These three private banks and all insurance companies were nationalized in January 1975.

Foreign trade. Ethiopia's exports consist almost entirely of agricultural products, while its imports are dominated by manufactured goods. Efforts to diversify exports had by the mid-1970s reduced the share of coffee from 60 to less than 30 percent. Next in importance are pulses (peas, beans, and lentils), oilseeds, and hides and skins. The United States buys about 70 percent of all Ethiopian coffee exported. Other major customers for Ethiopia's exports are West Germany, Afars and Issas, Saudi Arabia, Italy, and Japan.

Ethiopia's imports include manufactured consumer goods and capital goods necessary for industrialization and the mechanization of agriculture. The largest categories are machinery and transport equipment; about one-third of the value of imports is accounted for by road motor vehicles and manufactured goods. Imports of crude and processed petroleum, though they declined during 1968–74, will continue to be an important import unless substantial deposits are found in the country. Major import sources are the United States, Italy, West Germany, Japan, and the United Kingdom.

Between 1958 and 1972, imports consistently exceeded exports, creating an increasing negative balance of trade. The deficit was met chiefly by foreign capital in the form of loans and grants. During the mid-1970s, however, this imbalance was ameliorated. The operation of the national airline to foreign countries has had a positive effect on the balance of payments.

Management of the economy. *The public and private sectors.* Large-scale economic activity is confined to the public and expatriate private sectors. The private sector is limited to wholesale and retail trade, in which aliens play an important role, and to small manufacturing establishments and agro-industrial concerns, especially in Eritrea, where Italians are prominent in trade. Private investment also has a significant role in newly developed commercial agricultural enterprises, such as those in the Setit–Humera region and in the Awash Valley.

Public utilities, rail, air, and water transport, telecommunications, banking, and much of industry (oil refinery, tobacco, sugar, textiles, beverages, mining) are generally in the hands of government-owned or government-controlled corporations.

Public finance. The chief sources of government revenue are indirect taxes and taxes on foreign trade. The large subsistence sector and the administrative problems of tax collection make the contribution from direct taxes much smaller than it could be. External aid and loans are an important source, usually amounting to not less than one-quarter of total revenue. These are generally used for government investment expenditure, particularly for infrastructural development, such as the building of roads and dams.

The largest items of government expenditure are national defense and internal order and justice (25 percent or more). Expenditures on education (15 to 20 percent) and health (6 percent) are significant but less than might be expected in a developing nation.

Trade unions and employer associations. The Confederation of Ethiopian Labour Unions (CELU) includes about 150 trade unions with a combined membership of about 85,000 in the mid-1970s. Employers are organized under the Federation of Employers of Ethiopia (FEE).

Economic policies. The government's third five-year plan, for 1968 through 1973, gave high priority to agriculture and to the expansion of commercial farming. Other objectives were promotion of efficient foreign and domestic marketing practices, the provision of agricultural credit, diversification in coffee-growing areas, and improvement of the quality of coffee.

Land reform

The need for land reform has been generally recognized. In March 1975 the provisional military government announced a new land policy in which private land ownership was abolished. The policy provided for 25 acres (10 hectares) of cultivable land to be allotted to each farming family and for the formation of agricultural associations. Implementation of the reforms was expected to be slow, hampered by traditional values and a lack of administrative personnel, seed, machinery, tools, and fertilizer.

Economic policy aims at doubling the output of manufacturing, systematically surveying mineral resources, and giving priority to a dense network of feeder roads in key areas rather than building primary highways in economically unproductive areas.

Transportation. At the beginning of the 20th century, modern transportation of any kind was virtually absent. In the 1970s there were about 4,500 miles (7,200 kilometres) of all-weather roads, 744 miles (1,197 kilometres) of railway line, and an airline with well-developed domestic and international services.

The major roads radiate from Addis Ababa, which is near the geographic centre of the country, connecting it with the provincial capitals. Most of these roads do not extend to the borders, and an interconnecting secondary road system is still being developed. Nearly half the population lives more than nine miles from an all-weather road. Rapid improvements are not to be expected because of the difficult terrain.

Rail transport is also handicapped by terrain, and there are only two railway lines in the country. One line runs from Mitsiwa via Asmera to Akordat, in western Eritrea. The other connects Addis Ababa to the port of Djibouti, in Afars and Issas, on the Gulf of Aden.

Where land transport faces so many handicaps, air transport acquires special importance. Ethiopian Airlines began operating in 1946 and has provided quick access

to areas that would have taken decades to open up by road or rail. There are about 40 airports, most of which can accommodate only small aircraft. Ethiopian Airlines operates an international service from the jet ports of Asmera and Addis Ababa, which are also used by a number of foreign airlines.

Water transport is undeveloped. Most of Ethiopia's rivers are not navigable, and, of the lakes, only Tana has transport of any economic significance. Ethiopia has only two ports, Mitsiwa and Aseb on the Red Sea. Djibouti remains the main outlet to the sea because of its rail link to the capital.

Problems and prospects. The development of an effective transport network is only one of the several problems that loom large in the Ethiopian economy. Another is the necessity of bringing the huge subsistence sector into the money economy, mainly through the commercialization of agriculture. The spread of an educational system oriented in its content and methods toward development is a prerequisite for effective utilization of the country's human resources. To continue to lessen the dependence of foreign trade on coffee, impetus is needed for the raising of other agricultural products and livestock, for which there appears to be a market in the Middle East.

Diversification would be aided if prospecting for mineral resources, especially oil, were accelerated. The possibility of developing a regional common market such as the East African Union for manufactured goods offers some promise. External capital could participate to a greater extent if the necessary steps were taken. All this requires effective administrative machinery and an equitable distribution of income.

ADMINISTRATION AND SOCIAL CONDITIONS

Until the military coup of 1974, the Ethiopian state was formally a constitutional monarchy with a bicameral Parliament. In practice, however, the Emperor exercised effective political power.

The structure of government. *Constitutional framework.* The first formal constitution was granted by Emperor Haile Selassie I in 1931; it was superseded by the revised constitution of 1955, which affirmed the descent of the Emperor from King Solomon of Israel and the Queen of Sheba of the Aksumite kingdom. It established succession to the throne by male primogeniture and vested sovereignty in the Emperor. The wide powers the constitution granted to the Emperor included control of the armed forces, foreign relations, and all government departments. Parliament, however, had the power to assent to a declaration of war and to certain types of treaties. The Emperor also had the power to dissolve Parliament. When Parliament was in recess, the Emperor was allowed to legislate by decree to meet emergencies, and such legislation continued to have legal force unless expressly repealed by Parliament when it reassembled.

The government structure was profoundly altered by the "creeping" military coup of February–September 1974. On September 12 the Emperor was deposed and Parliament and the constitution of 1955 were suspended. The throne was at first offered to the Crown Prince, but on March 21, 1975, the crown was abolished in preparation for the establishment of a socialist republic.

Revolutionary government

The revolutionary government is composed of a 120-member Provisional Military Administrative Council, known locally as the Dirgue, and a civilian Technical Advisory Council of 50 members.

Regional government. Ethiopia is divided into 14 provinces, or governorates (*teklai gezatoch*), each of which is further divided into districts and subdistricts. Lower regional units are not uniform and do not occur in all provinces. Historically, the basis for the provincial units has been the linguistic and cultural affinity of the inhabitants, and this remains so to a large degree.

Under the imperial government, the provinces were much more under the control of the central authority in Addis Ababa than was the case in the traditional feudal structure. Whatever measure of local autonomy existed resulted more from difficulties of communication than from any conscious policy of decentralization. The revolutionary government promised a lessening of centralized power and a return to a more traditional system of local power centres.

The political process. The traditional checks on the power of the Emperor—exercised by the semi-autonomous feudal lords, other nobility, and the institution of the church—were either removed or greatly weakened by modernization and centralization. The Emperor appointed government officials down to the level of director, which was about five ranks below minister. Many of these ranks were held by several persons in each ministry, and the number of appointed officials was therefore quite large. The periodic announcements of government reshuffles were eagerly awaited; they were minutely studied as an accurate barometer of shifts in the Emperor's esteem and therefore in political power.

No decisions of any importance were made without the knowledge and approval of the Emperor. The prime minister was in reality a "first among equals" among the ministers and had little political power of his own. He had the derivative power of being close to the Emperor and being able to influence the latter's decisions on some issues; a measure of such power also was enjoyed by some of the Emperor's personal attendants. The prime minister was the chairman of the Council of Ministers; occasionally, this function was taken over by the Emperor himself.

In its policy of "Ethiopia First," the military government affirmed the right of self-administration at village, district, and regional levels. It reserved to the central government jurisdiction in national and other fundamental matters of state and the obligation to assist and support other levels of government.

The military government also announced its intention to form a national socialist political party. Limited political participation had existed in the traditional structure in the form of informal councils of elders. In a few sections of the country, notably among the Galla of Borana, a democratic indigenous sociopolitical structure known as the *gada* system had a marginal existence alongside the imperial structure.

Eritrean separatism

Separatist movements have been strong in the province of Eritrea, formerly an Italian colony and chiefly Tigrinya speaking. After the coup of 1974 the drive for separatism gained strength, presenting the new government with one of its most serious political problems. Major uprisings have also occurred in the provinces of Tigre, Bale, Harer, and Gojam.

Administration of justice. A variegated system of customary laws has operated among the various ethnic and religious groups. In the Christian highlands, a written code—the *Fetha Negast* ("Justice of the Kings")—was adapted from an Arabic original, which itself took inspiration from the Justinian code. Since 1960, civil and commercial codes, drafted by international legal experts, have been in force and presumably supersede all other laws. In practice, customary law still prevails, especially in lower courts and outside the main urban centres.

The judicial system includes the High Court (which has branches in the provinces), the Awraja ("district") courts, the Wereda ("subdistrict") courts, and the Atbia ("sub-subdistrict") courts. Cases requiring the application of Muslim law of the family, succession, and the like are heard by special Sharīʿah courts (based on the body of established Islāmic religious law). Under the revolutionary government, military tribunals were formed to try former government officials.

The armed forces. The armed forces have always been held in the highest esteem in Ethiopian tradition. Every able-bodied man was a potential soldier ready to bear arms for his feudal lord or emperor. The formation of a regular standing army began in the 19th century, and the armed forces are now among the most effective and best trained forces in black Africa.

The army, with about 45,000 men, has detachments strategically posted throughout the country; they are supported where necessary by airborne command units. The Territorial Army is a sizable force of reservists who

are trained for about two months every year. The Ethiopian Air Force is a small, well-trained, fully Ethiopian force whose airplanes include supersonic tactical fighters. The Ethiopian Navy is in the process of consolidation. Training facilities for the armed forces include the old military school at Genet, the Military Academy at Harer, the Air Force Academy at Debre Zeyt, and the Naval College at Mitsiwa.

Military training

The social milieu. *Education.* Traditional education is almost solely the concern of religious institutions and is exclusively religious in aim and content. Church and Qur'ānic schools are widespread, particularly in areas in which secular education is not available. The modern school system—introduced by Emperor Menelik II (reigned 1889–1913)—was greatly developed and expanded under Haile Selassie I. It is still inadequate, however, and an estimated 90 percent of the population was illiterate in the 1970s. Ethiopia has had an indigenous writing system for nearly 2,000 years, but its use has been largely restricted to religious purposes.

Ethiopia has a four-tiered educational system with elementary schools (grades one to six), junior secondary schools (seven and eight), senior secondary schools (nine to 12), and institutions of higher education. In addition, there are several vocational schools and in-service training programs. About 60 percent of all schools are under the direct control of the Ministry of Education and are fully government supported. The rest are mission, private, or Ethiopian Orthodox Church schools. The approximately 950,000 students constitute only about 10 percent of all school-age children.

The main institution of higher education is the National (formerly Haile Selassie I) University, in Addis Ababa. It has faculties of arts, science, education, technology, business administration, social work, law, and medicine. Public-health courses are offered in Gonder and agriculture courses in Alemaya, near Harer.

Medical services. The bulk of the population must rely on traditional methods of medical treatment, such as herbal medicines and holy waters. Modern medical care reaches only about 15 percent of the population, at most. Medical facilities are poorly distributed, being concentrated in Addis Ababa, Asmera, and a few other major towns. While the doctor-to-population ratio for the country as a whole is about 1 to 70,000, the ratios for Asmera and Addis Ababa are 1 to 2,500 and 1 to 3,000, respectively. In the provinces, the ratio ranges between 1 to 100,000 and 1 to 250,000. The distribution of other facilities is similarly lopsided; the establishment of rural health centres and the training of health officers and community nurses may partially correct the imbalance. Many of the hospitals, clinics, and health centres are understaffed, inadequately supervised, and lacking in essential supplies and equipment. Medical-training facilities include the Faculty of Medicine and the Public Health College (at Gonder) of the National University, and several smaller institutions for the training of other medical personnel.

Living conditions. The average annual income, Eth$190 (U.S. $90), is one of the lowest in the world. Most of the population is outside the money economy and lives on a subsistence level.

Because the population is concentrated in the highlands, most people are free from tropical diseases, but communicable diseases such as malaria, tuberculosis, and smallpox are prevalent. Leprosy is being fought with international help. Schistosomiasis (a parasitic infestation that is spread through snails that live in shallow water) is endemic; all bodies of water except Lake Langano are said to be infested, and irrigation schemes are spreading the disease. Preventive medicine is hampered by popular ignorance of hygiene and sanitation, lack of safe drinking water, and low nutritional standards.

Diseases

Social and economic divisions. On the whole, the cattle herders of the lowlands, and particularly the Somali and the Galla of Borana, have a highly egalitarian social structure. Before the land reforms of 1975, the farmers in the south were generally tenants for northern, and especially Amhara, absentee landlords. Among the Amhara and the Tigre, farmers usually owned the land they tilled, and the privileged group before the military coup consisted of the members of the government bureaucracy and the declining nobility. Special status was once enjoyed by members of the clergy, whose importance has waned. Most groups in Ethiopia include a depressed class, which usually consists of landless artisans or musicians.

Cultural life. Religion and language have probably been the two most important focal points of the cultural life of Ethiopia, both historically and in the 20th century. Of the two major religions in the country, Christianity has been traditionally predominant culturally. National holidays, ceremonies, shrines, symbols, myths, and sources of identification are nearly all Christian. The national literature, both past and present, is Christian in either content or orientation. What are regarded as the national dress and national traits and customs are generally those belonging to the Christian peoples. Only in certain localities such as Harer and Jima does Islām dominate. Painting historically centres on the church, and modern artists often choose religious subjects.

The culturally dominant languages have been the Semitic languages in general and Ge'ez and Amharic in particular. Ge'ez has always been held in special veneration, and Amharic, as the traditional court language for many centuries and as the constitutionally established national language, earned cultural prestige for its speakers and their way of life. The military government, however, pledged itself to encourage the development of all Ethiopian languages and cultures in order to promote national unity.

Libraries include the National Library and the University Library in Addis Ababa and the Asmera Public Library in Asmera. Also in Addis Ababa are the National Museum, which contains archaeological exhibits of note, and the Museum of the Institute of Ethiopian Studies, which contains ethnological exhibits and collections of paintings, coins, and stamps.

The press and broadcasting

The few newspapers of national circulation are all published in Addis Ababa. The Amharic daily *Addis Zemen* ("New Era") and the English daily *The Ethiopian Herald* have circulations of 35,000 and 30,000, respectively, while the circulation of the French daily *Addis Soir* is limited. Of the weeklies, the most important are *Ye-zareitu Ityopia* ("Ethiopia of Today") and *Ye-sport fana* ("The Light of Sports"). There is an Amharic monthly, *Menen,* and an illustrated English quarterly, *The Ethiopian Mirror.* The most popular newspaper, the Amharic fortnightly *Polisenna Ermejjaw* ("Police and Progress"), is published by the Ethiopian Police Force; it relies heavily on crime reporting and has a circulation of about 80,000. Other papers published in Asmera in Amharic, Tigrinya, Arabic, and Italian are of primarily local circulation.

The communications medium that reaches the masses in Ethiopia is radio. Radio Ethiopia, the government-owned station, broadcasts from Addis Ababa, Harer, and Asmera. News and information are broadcast primarily in English and Amharic, but other local and foreign languages are also used. A radio station owned by the Lutheran World Federation broadcasts religious and secular programs in several languages from Addis Ababa to African and Asian countries.

The Ethiopian Television Service, which began operations in 1964, has daily programming for several hours in Amharic and English. Though there is increasing local programming, there is still heavy reliance on imported television programs from the United Kingdom and the United States. There have been experiments with educational broadcasts for classroom use. Outside the capital, broadcasts are received only in a few towns stretching southward from Addis Ababa to Asella.

Prospects for the future. History weighs heavily upon Ethiopia, and most of the country remains as it has been for centuries. Yet time caught up with the long-established order, and the forces of change made themselves felt in the coup of 1974.

Economically, many of the problems of Ethiopia are no

different from those of other developing nations, although they may be more acute. The integration of the self-sufficient traditional sector into the modern money economy is a difficult but essential part of economic development if Ethiopia is to become an integral part of the modern world. The military government announced a policy of public guidance, control, or ownership of economic resources, as a means toward abolishing poverty and exploitation. Land reform is a vital aspect of the new economic program, and the development of the country's rich mineral resources is given a high priority. If economic development is to benefit the people, it must be rapid enough to do more than simply feed the extra mouths resulting from normal population growth and must be accompanied by a system that will make possible an equitable distribution of income.

After millennia of existence as a nation-state, Ethiopia is not free from the problems of national integration that bedevil so many nations in an age of popular political awareness. The great ethnic and religious diversity, which caused little discomfort in the highly autonomous quasi-feudal structure of the past, is today a potentially disrupting force. Such threats to unity and stability are made even more ominous by external encouragement. The nation has so far managed to weather periodic threats to its stability, and the military government has made it clear that the final solution to this problem must include the due participation of all groups in the cultural and political life of the country.

BIBLIOGRAPHY. Of earlier works in English JAMES BRUCE, *Travels to Discover the Source of the Nile*, 6 vol. (1790), is a classic—a keenly observed eyewitness account of medieval Ethiopia. Of modern works, EDWARD ULLENDORFF, *The Ethiopians*, 2nd ed. (1960), is especially strong on northern Semitized Ethiopia. G.A. LIPSKY, *Ethiopia: Its People, Its Society, Its Culture* (1962), is more comprehensive, but less dependable in matters of detail. D.N. LEVINE, *Wax and Gold: Tradition and Innovation in Ethiopian Culture* (1965), is a much discussed study of the culture and social psychology of northern Christian Ethiopia. The history and doctrine of Ethiopian Christianity is treated in A. WONDMAGEGNEHU and J. MOTOVU (eds.), *The Ethiopian Orthodox Church* (1970). J.S. TRIMINGHAM, *Islam in Ethiopia* (1952), the most extensive study of Ethiopian Islām, contains the best general account of Ethiopian history. The best source for all manner of statistical information is the latest issue of the annual *Statistical Abstract*, issued by the Central Statistical Office. DAME MARGERY PERHAM, *The Government of Ethiopia*, 2nd ed. (1968); and CHRISTOPHER CLAPHAM, *Haile Selassie's Government* (1969), are similar in approach and together constitute the best account on the Ethiopian polity before 1974. RICHARD GREENFIELD, *Ethiopia: A New Political History* (1965), has information on the coup d'etat of 1960 that is unavailable elsewhere.

There is a scarcity of general accounts of the geography of Ethiopia. G.C. LAST, "Introductory Notes on the Geography of Ethiopia," *Ethiopia Observer*, vol. 6, no. 2 (1962), is a good introduction. MESFIN WOLDE MARIAM, *An Atlas of Ethiopia* (1970), is comprehensive and based on the latest available statistical data. F.J. SIMOONS, *Northwest Ethiopia: Peoples and Economy* (1960); and Y.A. HAGGAG, *A Contribution to the Physiography of Northern Ethiopia* (1961), are among later firsthand geographical studies. J. BLOWER and J.F. VON WOLFF, *Shell Guide to the Wildlife of Ethiopia* (1969), is practical and well illustrated. The basic work on geology is still GIOTTO DAINELLI, *Geologia dell' Africa Orientale*, 4 vol. (1943). D.A. JELEC, *Mineral Occurrences of Ethiopia* (1966), is a good compilation.

A useful general introduction to the Ethiopian economy is ASEFFA BEQUELES and ESHETU CHOLE, *A Profile of the Ethiopian Economy* (1969). The ETHIOPIAN COMMERCIAL BANK, *The Ethiopian Economy* (1970), treats selected topics, notably commercial banking and the share market. Economic history is treated in RICHARD PANKHURST, *Introduction to the Economic History of Ethiopia*, 2 vol. (1961–68).

Competent studies of specific peoples of Ethiopia are WILLIAM SHACK, *The Gurage: A People of the Ensete Culture* (1966); ERNESTA CERULLI, *Peoples of Southwest Ethiopia and Its Borderland* (1956); G.W.B. HUNTINGFORD, *The Galla of Ethiopia: The Kingdom of Kafa and Janjero* (1955); A.E. JENSEN (ed.), *Altvölker Süd-Äthiopiens* (1959); E. HABERLAND, *Galla Süd-Äthiopiens* (1963); and H.S. LEWIS, *A Galla Monarchy: Jimma Abba Jifar, Ethiopia, 1830–1932* (1965).

WOLF LESLAU, *The Scientific Investigation of the Ethiopic Languages* (1956), gives a general overview of the Semitic languages. Further literature on the subject is fully covered in Leslau's indispensable *Annotated Bibliography of the Semitic Languages of Ethiopia* (1965). For other languages of Ethiopia, the best guide is still A.N. TUCKER and M.A. BRYAN, *The Non-Bantu Languages of North-Eastern Africa* (1956).

(A.Dem./Ed.)

Ethiopia, History of

Ethiopia is a unique phenomenon in Africa. Ethiopia and Egypt are the only African nations that can trace their history into antiquity, but Egypt became more and more alienated from its ancient culture by successive conquests, while Ethiopia, except for a few years of Italian occupation in the 1930s and 1940s, retained its original character through the centuries. Ethiopia's historic continuity is largely the result of a favourable geographic location; its high plateaus, which, both fertile and temperate, border on heavily travelled seas on the east and elsewhere, can be defended like fortresses from the hostile deserts that surround it. Thus, Ethiopia has been able to preserve and perfect a unique culture born of the ancient encounter and the slow fusion of two equally gifted peoples: the Cushites, a white race with dark skin, who were probably indigenous, and Semitic tribes that may have emigrated from Arabia.

Historic continuity

EARLY ETHIOPIA (TO THE 11TH CENTURY)

Remotest antiquity: the land of Punt. The Egypt of the pharaohs believed that it had received certain of its divinities from Ethiopia, which it called Punt, or the Land of God. Later, the Greeks thought of it as the original home of wheat and of the olive tree. The legends are partly true. The most recent discoveries confirm the ancientness of man's presence in Ethiopia; even the remains of an archaic hominid (*Paraustralopithecus aethiopicus*) dating back some 1,500,000 years have been found in the Omo Valley. According to Egyptian hieroglyphic texts, an authentic civilization began as early as the 2nd millennium BC. The pharaohs sent fleets there to buy such precious goods as spices, incense, and myrrh. Bas-reliefs in Dayr al-Baḥrī at Thebes, dating from around 1500 BC, depict certain Ethiopians cultivating myrrh and incense. The text commenting on these drawings calls the Ethiopians Habashat, from which later came the name Abyssinia.

The Sabaean period. An Egyptian dynasty—the 25th (*c.* 750–656 BC)—classified as "Ethiopian," in the larger sense of the word, by classical Greek authors, has been erroneously included in Ethiopian history itself. Its princes were, in fact, Sudanese, or Nubian. Direct illustrations of Ethiopia's past, such as monuments and inscriptions, are found on Ethiopian soil beginning only with the 6th century BC. These testimonies indicate the growth on the Ethiopian plateau of a civilization closely related to the Minaeans and to the Sabaeans from South Arabia. This "Sabaean" stage of its past led the Ethiopian nation to include in its ancient traditions the biblical episode of the Queen of Sheba's visit to Solomon, as told in the First Book of the Kings (10:1–10).

Between the 6th century BC and the 1st century AD, approximately, Sabaean influences reached the Ethiopian coasts and plateaus. Among these were astral religion, sacred royalty and social institutions, intensive agriculture, architecture, and art. Sabaean writing was adopted in Ethiopia in order to transcribe a language almost identical to that of the Arabian Sabaeans; that language and its writing have been preserved by the Ethiopian church in the liturgical language Ge'ez.

During the last centuries BC Ethiopia also felt the influence of Hellenistic Egypt. The Ptolemies' fleets founded trading posts on the Ethiopian coasts, through which entered "modern" merchandise, the beginnings of commerce, and a few traces of Hellenistic art and the Greek language. One of these trading posts became the great ancient port of Adulis, the modern hamlet of Zula.

The Aksumite Empire. From the 2nd to the 9th centuries AD the fusion of new immigrants (primarily the Aguezat, who left their name to the Ge'ez language) with people who had been on African soil for some time created an independent nation centred at Aksum on the

Tigre Plateau, which still remains Ethiopia's religious centre. Its busy ports helped the nation develop its military and maritime power and thus to dominate the neighbouring regions in the 3rd–6th centuries; even the South Arabian kingdom of the Ḥimyarites, a fusion of the Sabaean and Minaean kingdoms, came under the suzerainty of Aksum. This expansion of the Ethiopian peoples, all of whom were related, along the two banks of the Red Sea, the African banks of the Gulf of Aden, and even toward the coast of Benadir (then known as Azania) brought economic and financial prosperity and religious, artistic, and social vigour to that part of the ancient world. This was facilitated by intense maritime traffic in the southern Red Sea, through which the fleets had to cross in order to pass from the east to the west and which brought ships of all origins to Ethiopian ports, Adulis principally. Thus, Ethiopians imported commodities from everywhere in exchange for African products, particularly costly gums (incense, myrrh, etc.), which other peoples were then buying in considerable quantities for the needs of their pagan cults and funeral embalmings. Commerce rested on sound financial methods, attested to by the minting of splendid coins bearing the effigies of Ethiopian emperors and by the fact that the Christian Ḥimyarite city of Najrān, a vassal of Aksum on the Yemen plateau, was an early centre of authentic banking activity.

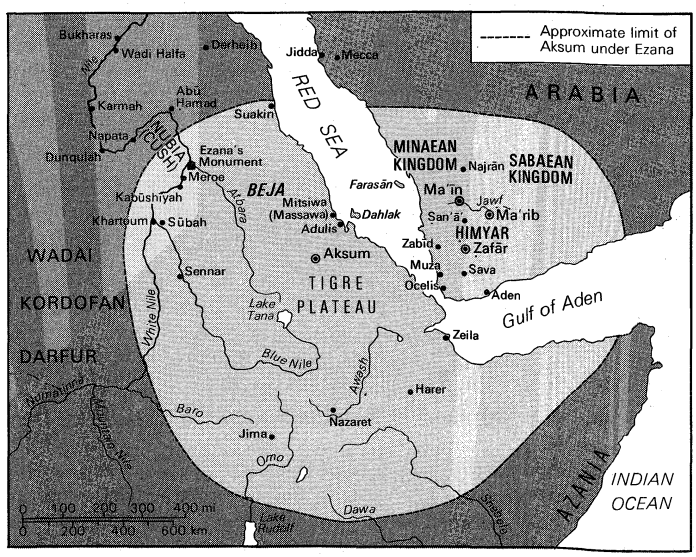
Franco-Ethiopian archaeological investigations, starting in 1954, revealed Aksumite monuments and inscriptions, which, along with data from a few classical texts, have led historians to recognize Ethiopia's role in the development of the southern Red Sea, the Upper Nile, and the eastern Horn of Africa. Some Aksumite emperors were already quite well known, however. Around the 1st century AD an anonymous Greek travel book, the *Periplus Maris Erythraei* ("Circumnavigation of the Erythrean [*i.e.*, Red] Sea"), called King Zoscales a miser but "an expert in Greek language," whose control extended from Aksum to the eastern extremity of Africa. In the 3rd century the Iranian religious leader Mani mentioned the Aksumite Empire in one of his epistles, calling it one of the four greatest existing kingdoms of the world. Sabaean inscriptions in Arabia relate the feats of the Ethiopian emperor Gadarat, who owned the cities of Ẓafār and Sava (now Ta'izz) and the ports of Ocelis and Muza on the Red Sea. Later Arabian inscriptions mention the Ethiopian emperor Wazeb I, to whom certain overseas Habashats made an appeal. Above all, an unnamed Aksumite sovereign toward the end of the 3rd century had engraved in Greek on a monument in Adulis (now lost but copied by the Byzantine traveller Cosmas in the 6th century AD) a recital of campaigns that show that his power extended from southernmost Egypt to the gold-producing regions in the present Ethiopian southwest, including the spice coasts on the Gulf of Aden and even the Sabaean and Minaean kingdoms across the sea. These Ethiopians were powerful enough to defeat Roman ships on the Red Sea.

The Aksumite emperors

Ezana. Ezana came to the throne between 320 and 325. To his title of king of Aksum, he added those of sovereign of Ḥimyar and of Sheba (*i.e.*, Saba), of Cush, or Meroe, and of the Beja, a desert people living between Ethiopia and Egypt. Ezana destroyed the remains of the kingdom of Meroe in the Sudan, erecting a votive monument at the confluence of the Nile and the Atbara. Ezana also caused his nation to adopt Christianity, although it was probably already well implanted by numerous foreign merchants and, more particularly, by the preaching of two young Syrians—Aedesius and Frumentius. Frumentius, still known to Ethiopians as Abba Salama, was the first bishop of Ethiopia, having been consecrated sometime in the 340s by St. Athanasius in Alexandria. This consecration initiated the subordination of the church of Ethiopia to the Egyptian church, which later became the Coptic Church. Ezana also brought about the definitive adoption for Ethiopian writing of an improvement of the Sabaean alphabet.

Ella-Asbeha. Another great emperor was Ella-Asbeha (Caleb), famous for a great expedition to avenge the Christians of Najrān, who had been persecuted around 523 by a local prince who had converted to Judaism. Undoubtedly, also in Ella-Asbeha's time, Christian monasticism was introduced into Ethiopia by a small number of monks who fled Egypt and possibly also Syria because of dogmatic controversies that were then disturbing those countries. Ella-Asbeha had established an Ethiopian viceroy, Abraha, over Ḥimyar. Abraha, who wanted to give his area a special prestige, tried to make the city of Ṣan'ā' and the cathedral he built there a magnet for Christian pilgrims and merchants from all over Arabia, Mesopotamia, and even Iran. Abraha's religious and commercial policies were continued successfully by a nascent Islām. In the Qur'ān there is a powerful, though inexact, reference to an unsuccessful military expedition with which, around 540, he tried to destroy Mecca, already becoming a competitor of Ṣan'ā'. But that period was, above all, marked by the repercussions of the great conflict between Byzantium and the Persians. The Byzantine emperor tried to persuade the Ethiopians to buy as much as possible of the silk produced in China through India and thus diminish the revenues of his Iranian enemies. Meanwhile, the anger of the Persians, stirred up by Abraha's ambition and intrigues, was turned against the regions occupied by the Ethiopians. In 572 the Persians disembarked at Aden, invaded the Ḥimyarite kingdom, and pushed some Ethiopian colonies as far back as the African coastline. Their conquest was followed 30 years later by Arab invasions. Ethiopian power was crumbling in Arabia, while Aksum's decline was

Persian invasion in 572



The kingdom of Aksum under Ezana in the 4th century AD.

beginning on the African coasts and plateaus. Thus disappeared all that still remained in Arabia of the ancient Sabaean and Minaean kingdoms, and the Ethiopians were the only remaining true heirs.

The decline of Aksum. Aksumite power began to decrease in the 7th century, and soon, as a sign of its impoverishment, the empire ceased to mint coins. Already Christianity's expansion in the civilized world had gravely affected the commercial revenues of these regions by devaluing incense and myrrh, resins important in pagan religious rites. The appearance of the Persians and the Arabs and the loss of Ethiopian possessions in Arabia cut Aksum off even further from the West. But not until later did a religious break separate the Aksumites from their Muslim neighbours. During the early years of Islām, relations between the Ethiopians and their Islāmic neighbours across the Red Sea were conducted with a great deal of mutual sympathy. This is easily explained: the Ethiopian Church retained an archaic Christianity that was very Semitic, still imprinted with the Judaism that had nurtured it, and very close to the religious syncretism from which Islām itself developed. Also, when companions of the Prophet Muḥammad fled from Arabia to Aksum in 615, they were well received and returned to their homes full of admiration for the Ethiopian emperor and his faith. Muḥammad expressed great emotion when the Aksumite emperor Armah died and advised his followers never to come into conflict with the Ethiopians.

Ethiopian fleets were active on the Red Sea as late as the 12th century, and in 702 Aksumite ships even harried the Arabian port of Jidda. Even later, the Ethiopians still kept a few possessions on the Arabian coast, such as Zabīd, which was governed by Ethiopian princes. They were no less respected on land: in the Middle Ages the Ethiopian emperors were again considered the masters of the waters of the Nile, which, according to popular belief, they could divert to the point of starving Egypt. But the city of Aksum itself declined to a point at which the seat of power was transferred farther southward, to Nazaret (now Adama), in the mountains of Tigre, while monastic establishments moved even farther to the south—for example, a great monastery was founded near Lake Haik in the 9th century. Most importantly, the land frontiers of the empire came in conflict with new peoples. To the north, along the coast extending from Egypt to Ethiopia, there appeared the short-lived kingdoms of the Naqis (on the southern borders of Egypt), then Baqlin, Bazin, Jarin, and Qata; these last two, vassals of the Ethiopian Empire, occupied the shores and part of the plateaus of the present province of Eritrea. Farther inland, Nubia was entirely Christian, adhering (according to region) either to the Byzantine or to the Coptic Church. From north to south, the kingdoms of Nobatia, Maqurrah, and ʿAlwah (Alodia) resisted the Muslim conquest energetically; not until the 16th century did ʿAlwah finally succumb, the last to do so.

Conflict with new peoples

Toward the south, the coasts of the Indian Ocean were held far back by the Bantu kingdom of the Zanj, the heir to ancient Azania, while in the ports Ḥimyarite merchants were progressively supplanted by Persian or Arabian immigrants. The Zanj kingdom did not decline until it was attacked by the Galla, perhaps from the 11th century, from the interior, and then, after the 14th century came under the attack of an entirely new people —the Somalis, from the mountains near Cape Guardafui. In the south, between the fault of the lakes and the loop of the Blue Nile, was the great Damot (Kingdom), possibly identical to the Demdem people, themselves known as being near Lake Abaya, the homeland of the Galla. A part of these turbulent populations, who were more or less converted to Judaism, launched frequent attacks against the Christian provinces of Ethiopia until the beginning of the 13th century. The most serious of these, led by a queen named Esato (Judith) around 976, laid waste the Christian empire as far as Tigre.

THE LATE MIDDLE AGES (11TH–15TH CENTURIES)

The Zague (Zagwe) dynasty. The Ethiopians saw Esato's invasion as a heavenly punishment for their having failed in obedience to the Coptic patriarchate; once delivered of it, they tightened their bonds with the Egyptian church. Then, under very obscure circumstances—perhaps at the end of the 10th century or perhaps 100 years later—the usurping dynasty of the Zague appeared in Ethiopia. Later ecclesiastical texts accused this dynasty of having not been of pure "Solomonid" stock (*i.e.*, not descended from the union of Solomon and the Queen of Sheba). The Zague, princes from the Lasta region of central Ethiopia, took advantage of the eclipse of the Aksumite sovereigns to transfer the seat of the empire to their own region. Some of them became powerful enough to war effectively against the rebels of southern Ethiopia, to counter Cairo's anti-Christian manoeuvring by threatening the destruction of Mecca, and later to send to Egypt a few elaborate embassies, which are recorded with awe in the Coptic chronicles. But the Zague distinguished themselves above all on the religious plane.

Appearance of the Zague

The Ethiopia of this period was garnering the declining traditions of the last great monasteries of Middle Egypt, the art, rock-hewn architecture, and distinctive literature from which it copied. It was also influenced by the artistic creations of foreign artists, principally Copts escaping persecutions in Egypt in the 11th and 12th centuries. These influences explain the creation by the emperor Lalibela of an admirable holy city that took its name from him. There, a dozen monolithic churches chiselled out of solid rock represent the last and most beautiful artistic expression of the Aksumite stoneworkers.

The Solomonid restoration. Late in the 13th century, partisans of the ancient Aksumite tradition suddenly swept away the Zague, who were replaced by Amhara princes from the Lake Haik region, from Geshen, and from Ambasel. The new dynasty, starting with Yekuno Amlak (ruled 1270–85), transferred the seat of the empire even farther south to Shewa (Shoa), where it remained until the 16th century. The dynasty's rise had been supported by a powerful sector of the monastic church, which proclaimed the new sovereigns to be the legitimate heirs of the Aksumite line as defined by the traditions of the *Kebra Negast* ("Glory of the Kings"). This monastic faction, most closely linked to the Egyptian monasteries of the Wādī an-Naṭrūn and to the Coptic patriarchate, was based in Ethiopia on such great monasteries as St. Stephen on Lake Haik and Debre Libanos in Shewa. In return for its support, the sovereigns gave that faction perpetual title to considerable landed benefices.

In religion and culture the authority of the Coptic Church, represented in Ethiopia by a metropolitan and his following of monks, progressively obliterated the previous Syrian and Byzantine influences. The religious inspiration of the sovereign's policies was unceasingly affirmed in the essays or commentaries of national prophecies. As heads of the church, sovereigns participated actively in the development of religious culture and discipline, as much through their bitter but ineffective battle against dissident sects, such as the Stephanites and Mikaelites (who preserved archaic Christian usages that themselves came from certain Coptic monasteries), as by their repression of surviving pagan superstitions. In addition, such sovereigns as Zara Yakob and Naod made themselves illustrious by beautifying churches and their ceremonies and by composing religious hymns.

This spiritual labour was facilitated by a renewed economic prosperity. Politically, the period was marked by rare and difficult relations with the Christian West—a situation that continued until the Portuguese appeared in neighbouring seas early in the 16th century. The most active diplomatic relations were with the sultans of Egypt. Starting in the 14th century, the empire had to face continual revolts in its southeastern provinces, where Christianity was being pushed back by the force of Islām. These revolts, first caused by economic problems (*i.e.*, the impoverishment of these regions), later took on religious tones. On the Ethiopian frontiers these movements dismembered first the pagan black kingdom of the Zanj and, more slowly, the great Christian kingdoms of Nubia to the north.

Foreign relations

Most sovereigns of this period have justly remained famous. Amda Tseyon (Amda Seyon; ruled 1314–44), who favoured the expansion of Christianity into new areas, defeated four coalitions of rebel Muslim vassals and energetically took up the defense, from a distance, of Egyptian Copts against the Sultan; his fame not only extended to the Muslims but also reached the West. The reign of Saifa Ared (Newaya Krestos; ruled 1344–72) saw the arrival of a Coptic prelate, Abba Salama, who was the author or instigator of numerous translations of Arabic Coptic works into Ge'ez. Dawit (David) I (ruled 1382–1411) obtained in Ethiopia certain highly venerated relics, including an alleged piece of the true cross and an icon representing Christ crowned with thorns, legendarily attributed to St. Luke. Also in Dawit's reign the collection of *Miracles de la Sainte Vierge*, written in France in the 12th century, was translated and introduced into Ethiopia. Yeshaq I (ruled 1414–29) recaptured the port of Zeila from rebel Muslims in 1415 and had the churches there rebuilt. He tried to pacify the northeastern regions, where the judaicized Ethiopians (the Falashas) were in revolt. Cairo suspected him of plotting with the Latins, and, in fact, he had relations with the King of Aragon and with Jean, duc de Berry, in France. These relations were continued, in spite of extreme difficulties, in the time of Zara Yakob (ruled 1434–68), who sent Ethiopian ecclesiastics as delegates to the Council of Florence in 1439. Baeda-Maryam (ruled 1468–78) and Eskender (ruled 1478–94) were distinguished by their struggle against rebellious Muslims. Under Naod (ruled 1494–1508), attempts to initiate some contact with the West became more effective. These efforts even led to the arrival, in 1520, of a Portuguese ambassador, who was received favourably by Lebna Dengal (ruled 1508–40). This meeting was finally possible because the Portuguese had just reached the Red Sea and had eliminated the monopoly held by the Indian and Egyptian fleets.

Arrival of the Portuguese

ETHIOPIA IN THE MODERN PERIOD (SINCE 1500)

Contacts with the Turks and Portuguese (1520–1632). Catholic Europe had been attracted to Ethiopia by the legend, fabricated in the West at the time of the Crusades, that far in the East there was a powerful Christian sovereign, Prester John, whose alliance would have been helpful against Islām. He had at first been vainly sought in Central Asia, where there were Nestorian Christian princes; then he was thought to have been an Ethiopian emperor, about whom there were analogous legends. Also, the discovery of Ethiopia by the Portuguese, starting in 1520, had received as much attention in European literature as had the discovery of America. But this encounter was not to Ethiopia's advantage; the Egypto-Indian fleets that the Portuguese chased from neighbouring seas were quickly replaced by those of the Turks, who had conquered Egypt in 1517 and who had installed themselves in Arabia. The Turks came into conflict with the Portuguese and proceeded to take issue with the latter's allies, the Ethiopian Christians. Lebna Dengal's reign became clouded in the aftermath of these events. His empire was immediately devastated by the worst Muslim invasion the country had yet seen. An uprising started in the southeastern regions, impoverished by the breakdown of a commerce that had been allied to the recently departed Egypto-Indian fleets. This revolt was immediately supported from Arabia by the Turks; it lasted from 1523 to 1543. In 1541 a Portuguese troop, small but well-armed and led by Dom Cristóvão da Gama, brought unexpected help through Mitsiwa (Massawa). The *imām* Aḥmad Grāñ, "the Left-Handed," leader of the Muslim invasion, was killed in the Battle of Waina Dega on February 21, 1543. Though his armies were at first immense, many in them had perished, and, as a result, the southeast provinces were left even poorer than the Christian provinces they had devastated. In spite of that victory, Galawdewos (Claudius; ruled 1540–59) suffered a few last assaults from Harer and himself fell in battle. By then, other peoples—the Galla and Somali—were dividing the spoils of the ancient Muslim provinces that had once extended from Harer to Zeila.

Minas (ruled 1559–63), Sartsa Dengal (ruled 1563–97), and Susenyos (ruled 1607–32) had to face three new problems: Turkish attempts on the coasts of the Tigre, from whence they were dislodged definitively by Ethiopian victories in 1578 and 1589; a great migration of the Galla, who expanded over almost the entire empire; and Roman Catholic propaganda by missionaries who, sent by St. Ignatius of Loyola, arrived with the Portuguese soldiers. The missionaries wished to convert the country to the Catholic (Chalcedonian) faith, and they were so successful that the emperor Susenyos thought he could proclaim the country's conversion in 1626. This transformation was all the more natural, because the Egyptian Coptic Church had itself recently attempted a short-lived union with Rome (1595–1610). Unreasonable demands of the missionaries in Ethiopia quickly raised against themselves and the Emperor great numbers of Christians who adhered obstinately to their ancient faith; and, in 1632, in the face of raging civil war, Susenyos was forced to revoke the union with Rome and to abdicate.

Arrival of Jesuit missionaries

Gonder Ethiopia (1632–1855). This period of Ethiopian history is comparable to the Western Renaissance, of which it was, in certain aspects, the inheritor. Architects, inspired by Indo-Portuguese models introduced by missionaries, produced splendid castles and abbeys. Miniatures were influenced by Western engravings, which had arrived by way of India—those of the Mughals as well as those of the Portuguese. The art of war evolved under Portuguese influence, and armies were formed that would later defeat the Egyptians in 1879 and the Italians in 1896. But this 16th-century Ethiopia, provoked by the Catholic missionaries, closed itself to the Latin West and threw itself into the arms of the Egyptian Coptic Church. It limited its relations to the Muslim powers—Egypt, the Ottoman Sublime Porte, and Mughal India—and to such non-Catholic Westerners as the Lutheran Peter Heylig, the Scottish James Bruce, and the governors of the Dutch Indies.

Fasilides (Basilide; ruled 1632–67) expelled the missionaries in 1633. He founded a new capital, Gonder, and began to build its first palace. Yohannes I the Just (ruled 1667–82) continued his work. Iyasu I (ruled 1682–1706), made a saint by the church, further developed and beautified the city, which, with 100,000 inhabitants, became the second largest city (after Cairo) in Africa. Iyasu brought about humane reforms in the administration, quelled controversies among rival theological factions, and, by his wars, provisorily re-established his authority over the provinces of the south, from which Christianity had been swept away by the pagan Galla. Baqaffa (ruled 1721–30) died too soon to effectively pursue that great work, but his widow, the empress Mentewab, watched over it with wisdom for some 40 more years during the reigns of her children and grandchildren, upon whose advent she, almost each time, exercised a regency. Of these the most brilliant was Iyasu II (ruled 1730–55), who, with Mentewab, had the splendid Abbey of Kusquam built. Iyasu II's armies, however, suffered a costly defeat at Sennar, from which the Fung Muslims had recently swept the Nubian Christians.

Founding of Gonder

To free himself from Mentewab's tutelage, Iyasu II tried to ally himself with the Gallas; he married the daughter of one of their chiefs and introduced Galla warriors into the capital, provoking troubles that became more accentuated after his reign, under Iyoas (ruled 1755–69). The notables appealed to one of the most able Tigre warriors of the empire, Mikael Sehul, to expel the newcomers. Mikael took advantage of the situation to impose himself on the sovereigns; he had the resisting emperor Iyoas assassinated, removed Mentewab, and then himself named the emperors he wished. Tekle-Haimanot II (ruled 1769–77), who was at first submitted to this tutelage, was freed from it in May 1771 thanks to the battles of Sarbakussa, in which the Scotsman Bruce participated. But the empire, still caught between the rival ambitions of the Galla and the Tigre, saw the development, in its government and mores, of a feudal spirit that had been previously unknown and the vestiges of which would continue into the 20th century. Affected

by this decadence, the end of the Gonderian age is known as the "time of the princes," a period in which, at the whim of the new Galla lords or of the Tigre, princes who succeeded one another were from the Solomonid line but had no army, riches, or power.

Reunification (1855–1935). Ethiopia's revival, aided by the church, began simultaneously in the Gonder region and in the province of Shewa, which was kept isolated from the previous troubles by its kings, who descended, as did the Gonderian emperors on another branch, from the line of Lebna Dengal. Ethiopian anger toward the feudals was awakened by the propagation of Islām, undertaken by certain Galla lords, and at the same time by the inability of these leaders to defend against unprovoked attacks that Egypt, already master of the Sudan, was launching against Ethiopia. This anger found a leader in Kassa, a chief of partisans organized against the Egyptians. It was Kassa who gathered enough troops to overthrow the phantom of empire that the feudals had maintained around Gonder.

Anger toward the feudals

Tewodros II. In 1855 Kassa had himself proclaimed emperor under the name of Tewodros II (Theodore; ruled 1855–68). He blamed the ancient Gonderian splendour for the national decadence; he thus despoiled the city of its treasures and established his residence in the austere central mountains. During this period the Western powers were competing to establish themselves in this eastern part of Africa. A misunderstanding with the British government resulted in a punitive expedition led by Lord Napier in 1868; Tewodros chose to kill himself in besieged Magdela rather than fall into the hands of British soldiers.

Yohannes IV. Yohannes (John) IV (ruled 1872–89), an ardent religious reformer, hastily converted the Muslim and pagan populations to Christianity. He defeated Egyptian colonization attempts at Gundet and at Gura (1877). Although he managed to have his rights to the northern part of his empire respected, he failed, through both diplomacy and arms, to prevent the establishment of the Italians at Mitsiwa, where they were introduced by the English to provide aid for the defeat of Mahdism in the Sudan. Furthermore, the Ethiopians were themselves in conflict with Mahdists, who laid waste to Gonder in 1888. Yohannes fell under their blows while winning the Battle of Metemma in 1889.

Menelik II. King of Shewa from 1864, then emperor of all Ethiopia upon the death of Yohannes IV, Menelik II (ruled 1889–1913), in the south, carried out defense policies similar to those of Yohannes in the north, stopping or restraining foreign colonization by arms or by diplomacy. As early as 1871 he attempted to reconquer all the southern provinces that had escaped imperial authority after the 16th century. The Egyptians, established since 1870 on the African coasts of the Gulf of Aden, had taken Harer in 1875 but had less success at Aussa. In 1887, the same year that he founded the new capital, Addis Ababa, Menelik retook Harer. But he came up against the British in the territories that later became the Somaliland Protectorate and against British manoeuvres (despite an Anglo-French treaty of 1888 protecting Ethiopia) attempting to install Italians at Zeila and to open access to Harer for them. In 1889 Menelik II concluded the Treaty of Uccialli with the Italians, which allowed a certain Italian penetration into Eritrea; but he insisted, nevertheless, that on paper, because of a difference between the Ethiopian and Italian wording of this agreement, he had been cheated. He protested, denounced the treaty, and addressed a circular letter to the great powers, defining the ancient frontiers of his empire and proclaiming them sacrosanct. The Italians invaded Tigre, and Menelik II repelled them, finally defeating them in 1896 at Adowa. France, Great Britain, and Italy immediately signed treaties with him (1896–97), limiting their ambitions. After 1892 Menelik entered into a cooperative agreement with France that included plans for a Franco-Ethiopian railroad linking Obock to the Upper Nile along one of the great medieval commercial routes—a project that would have contradicted British plans for linking Cairo to the Cape. With the Mahdists paralyzing the British, Menelik, in 1898, undertook with France to conquer the Upper Nile. But the Franco-Ethiopian expedition was unsuccessful. In addition, in the following weeks English successes against the dervishes permitted the British to dislodge the French captain Jean-Baptiste Marchand, who had been trying to meet the Franco-Ethiopian expedition, from Fashoda (Kodok). As a result, the Franco-Ethiopian railway would not go past Addis Ababa (which it reached in 1917), and, in 1902, an Anglo-Ethiopian treaty forbade the Ethiopians to make free use of the abundant annual floods of the Blue Nile (Ābbay River), the waters of which are indispensable to the prosperity of the Sudan and of Egypt. Menelik was ill in 1906 when, in a tripartite treaty, the Italians, British, and French gave themselves new rights to his empire, and he had to accept a vague treaty between his government and Italy in 1908 for the delimitation of Somalia. This last treaty later gave Italy a pretext for claiming part of the Ogaden region and launching aggression in 1935 (see below).

Agreements with foreign powers

Menelik died in 1913, seven years after the death of Ras Makonnen, who was to have been his successor. Gravely ill, the Emperor had allowed a regency to be instituted in 1907, until his grandson, Lij Iyasu, could reign. But after May 1911 Lij Iyasu, although only 15, rejected all tutelage. He showed himself to be incapable of governing; and, by his tendencies toward Islām and by his alliance with Muḥammad ibn ʿAbd Allāh Ḥasan (the "Mad Mullah"), who had revolted against the British and Italians in Somalia, he seeded defiance among his subjects during a period—World War I—when the Germans and the Turks were trying to gain control of Ethiopia. In 1916 the church and the nobles decided to remove Iyasu, who had not yet been made emperor, and to make Menelik's daughter Zauditu (Zawditu) empress with the assistance of a regent—Ras Tafari, the son of Ras Makonnen, who was proclaimed heir apparent to the throne.

The regency and early reign of Ras Tafari (Haile Selassie I). Zauditu (ruled 1917–30) was crowned on February 11, 1917. Lij Iyasu's reversal saved Ethiopia; tied by Iyasu to the Germans and the Turks, it would have been dismembered by the victorious allies as an enemy power—already the Treaty of London (1915) had promised Ethiopia to Italy—but the proclamation of 1916 had placed it under Ras Tafari's regency in time. Thanks to Tafari's diplomacy, Ethiopia was admitted to the League of Nations (September 28, 1923), which authorized it, in 1926, to protest energetically against a new plan for division of influence drawn up by the British. In order to modernize the country, the Regent had to remove certain feudals and, above all, to regain for his country the outlets to the sea that had been lost to the Italians, French, and British on the former Ethiopian coasts of the Red Sea and the Gulf of Aden. Great Britain offered Ethiopia an outlet through Zeila or Berbera. Italy offered the use, duty free, of the port of Aseb (Assab), a solution which was accepted by a treaty of August 1928 but without follow-up. On April 2, 1930, Zauditu died; Tafari, already crowned king (*negus*) in October 1928, became emperor the next day. On November 2 he was crowned under the name Haile Selassie I.

The crowning of Haile Selassie I

Upon his accession to the throne, Haile Selassie gave his country a constitution (July 16, 1931) and began a vast program of modernization, immediately sanctioned by the opening of an Ethiopian parliament. But international factors accumulated dangers for the country. These included the consequences of the Great Depression, the results of postwar bargaining among the Allies, and Mussolini's desire to avenge the disaster of Adowa and to appropriate Ethiopian territory. Although Ethiopia was Christian, civilized for centuries, and a member of the League of Nations, it could not escape these appetites.

The Italian invasion and occupation (1935–41). Mussolini's plans for conquest emerged in 1928–30, when the Italians began to remove from their maps the roughly outlined frontier between Ethiopia and Somalia, which had been agreed to in 1897 and 1908 but which now seemed too narrow for Italian ambitions. In the international press they paid for an intense campaign against Ethiopia, which did not have the means to defend itself.

The dispute was brought before the League of Nations; the Assembly initiated discussions between the two parties on what seemed to be the basis of the problem: the drawing of the frontier in the Ogaden. But the documents of the treaty of 1897, the basis of the frontier definitions of 1908, had been lost by the Ethiopians, and the Italians asserted that their copies of these same documents were, inexplicably, not to be found. Discussions between December 1934 and March 1935 led to an impasse, and, in spite of the principles of the League of Nations, even the British foreign secretary, Sir Samuel Hoare, and the French premier, Pierre Laval, were resigned to the establishment of an Italian protectorate over Ethiopia (the Hoare–Laval division plans). Italy had massed considerable troops on the Ethiopian frontiers, while the Emperor, through a desire for peace, had retired his troops to about 20 miles (30 kilometres) behind these same boundaries. On October 3, 1935, without a declaration of war, the Italians invaded Ethiopia from Eritrea and Somalia. Sanctions placed on Italy by the League were ineffective. Ethiopia resisted for seven months, but Addis Ababa was taken on May 5, 1936. On June 30 the Emperor personally defended his nation's cause before the League. But Rome had already proclaimed Ethiopia's absorption into its empire, and the League, in taking action, excluded Ethiopia from its debates. (There was an unexpected consequence, however. Egypt, caught between two Italian colonies, one of which controlled the waters of the Blue Nile, became disturbed to the point that the British accorded it independence in August 1936.)

The Italians executed Archbishop Petros (the Coptic *abuna*, sent from Egypt as head of the Ethiopian Church), massacred the monks of the great monastery of Debre Libanos, and, after an attempted coup, decimated the population of Addis Ababa during a three-day period in February 1937. Certain regions of Ethiopia, however, were never conquered. Alliance with Germany in World War II led Italy to invade the British territories of Somaliland, but the British and the free French immediately replied by aiding the Ethiopian partisans and thus delivering Ethiopia. Emperor Haile Selassie returned to his liberated capital on May 5, 1941.

Liberation of Ethiopia

Ethiopia after the liberation. Ethiopia was at first regarded by Great Britain as a conquered Italian colony. It was returned to the Emperor's authority only after agreements in January 1942 and December 1944, except for the Ogaden, which the British kept under military administration until 1948, and Eritrea, which was not reattached to the empire until 1952 and then in a federative form, after a referendum organized by the United Nations; it was finally reunited in 1962. Because it had been removed from official maps as a free nation since 1936, Ethiopia was given incomplete rights in the postwar peace negotiations.

Postwar revival. It took strong efforts by the imperial government to revive Ethiopia after the war. To return the nation to its place in international life, the Emperor appealed to the United States for aid; early in 1945 he met Pres. Franklin D. Roosevelt in Cairo. Several months later he ratified the United Nations charter at San Francisco. A mutual defense treaty with the United States was signed in 1953. The founding of Ethiopian Airlines, sponsored by the U.S.-based Trans World Airlines (TWA) in December 1945, provided the country with vital internal communications and facilitated contact with the rest of the world.

Constitution of 1955

Celebrating the anniversary of his coronation in a resurrected Ethiopia, Haile Selassie I in November 1955 proclaimed an extension of the constitution of 1931; the first general elections to the Chamber of Deputies were held in 1957. An attempted coup d'etat in December 1960 was easily put down. Technical developments, primarily in the highway network, air communications, and electrification, made use of Ethiopia's inexhaustible hydroelectric resources. Education was improved, based on the use of English, together with Amharic, the national language, as a language of instruction; after 1966 French, the principal foreign language before the Italian conquest, officially took the third rank. A university was founded in Asmera (1958), and the university college in Addis Ababa was raised to university status (1961). Considerable work in sanitation was accomplished, primarily in the towns.

Religious reforms. The constitutions of 1931 and 1955 affirmed the traditional religious freedom accorded Ethiopian citizens. By an agreement of 1959, the traditional subordination of the Ethiopian Church to the Coptic patriarchate was reduced to a simple honorary attachment. The Ethiopian Church won international religious prestige that allowed it to push other Eastern "non-Chalcedonian" churches into regrouping; a conference was held in Addis Ababa in 1965, presided over by the Emperor. Renewal of relations with the papacy, broken off since 1632, was sanctioned in June 1969 by a meeting between the Emperor and the Pope in Geneva and, later, by a visit of the Emperor to the Vatican (November 1970).

International relations. In international politics Ethiopia especially dedicated itself to promoting African unity and progress. The charter of the Organization of African Unity, prepared by several conferences, was proclaimed at Addis Ababa in May 1963; the Ethiopian capital had become the principal centre of inter-African conferences.

Regional problems. The regional problems were the legacy of earlier foreign colonization. Thus, the independence of Somalia (1960), formerly two colonies, resulted in complications for Ethiopia. The expansion of the Somali people into the Horn of Africa was a result of the depopulation of the Ethiopian Muslim provinces during the wars of the 16th century. The newcomers had moved southward at the expense of more ancient inhabitants, Gallas and Zanj Bantus. International attempts to obtain, by agreement between Ethiopia and Italy, the delimitation of a precise frontier between Somalia and Ethiopia before Somalia became independent were in vain; Italian representatives again took recourse to arguments founded on the mysterious disappearance of the 1897 treaty documents. From this point there developed among younger Somalis the myth of the greater Somalia, which could be created only by taking certain territories from neighbouring countries, including the Ogaden region from Ethiopia. An intense campaign of anti-Ethiopian propaganda among the Somalis was the prelude to armed confrontation (February–March 1963). After several years, patient diplomacy on the part of enlightened persons on both sides seemed to have restored peace, but the problem was not settled.

Somalia–Ethiopia difficulties

A problem in Eritrea stemmed from analogous causes. In 1962 the Eritrean parliament spontaneously demanded that the province's territory, placed in federative status in 1952, should regain provincial status within the empire. This land, before the Italian colonization, had been one of Ethiopia's oldest provinces, dotted throughout with ancient Christian sanctuaries as well as with historic battle sites. Foreign colonization, propagation of Roman Catholicism, and immigration of Muslims had scarcely affected the Ethiopian character of the region. After 1966 local discontent emerged in Eritrea, which is particularly deprived of natural resources and the needs of which neither Ethiopia nor the occupying Italians had been able to fulfill. A revolt broke out, led by a group known as the Eritrean Liberation Front.

(J.P.Do.)

The fall of the monarchy and after. In 1974 dissatisfaction with the long-established order in the empire of Ethiopia finally amassed sufficient force and popular backing to initiate a process of fundamental change. The movement began in mid-February, when strikes and demonstrations by students, taxi drivers, and trade-union organizations paralyzed life in Addis Ababa. The civilian demonstrators were joined by units of the armed forces, motivated by a number of considerations including a demand for increases in pay. After a series of events that included changes in the Cabinet and widespread action against corrupt officials, the movement culminated on September 12 in a proclamation ending imperial rule and in the deposition and imprisonment of Emperor Haile Selassie. The throne was offered by the Provisional Military Administrative Council—composed of 120 members

of the military forces and the police—to Haile Selassie's named heir, Crown Prince Asfa Wossen. According to the council, the king would act as chief of state but would have no constitutional powers.

These changes were accompanied by measures against the old ruling group of aristocratic families and palace appointees. On November 23, 59 of these, including the last two prime ministers and 18 generals, were summarily executed, and on December 4 a military tribunal began trying former officials on charges of corruption and maladministration. In a policy statement on December 20, the council announced that it would turn Ethiopia into a socialist country with a one-party system, collective farms, and direct government control over all property useful for economic progress.

The monarchy was abolished entirely in March 1975. The proclamation issued by the military council annulled the appointment of Crown Prince Asfa Wossen as king-designate, cancelled the titles of prince and princess, and stated that the Ethiopian people should decide what type of government they wished.

Drought and famine. Thousands of persons—at least 50,000 and perhaps as many as 100,000—died in 1973, particularly in Shewa, Welo, and Tigre provinces, of the effects of a severe drought. By mid-1975 a still worse drought was being predicted, this one centring on the Ogaden region, bordering Somalia. About 500,000 persons were affected, according to official sources.

Eritrean revolt. In December 1974 the Eritrean Liberation Front and a second, more Marxist-oriented guerrilla group that had broken away from it, the Popular Liberation Front, intensified their struggle against the central government. Early in 1975, open warfare between government troops and guerrillas broke out. It continued sporadically thereafter, with considerable loss of life. Popular opinion in Eritrea appeared to be turning more and more toward the guerrillas, and Eritrean independence was attracting more and more supporters. Loss of Eritrea was abhorrent to the government in Addis Ababa: it would mean to Ethiopia the loss of its only seaports and its only petroleum refinery. (Ed.)

BIBLIOGRAPHY

General works: J. DORESSE, *Au pays de la Reine de Saba: L'Éthiopie, antique et moderne* (1956; Eng. trans., *Ethiopia*, 1959, reprinted 1967), with numerous illustrations, and *Histoire de l'Éthiopie* (1970); J. SPENCER TRIMINGHAM, *Islam in Ethiopia*, ch. 2 (1952, reprinted 1965), an excellent summary of Ethiopian history; EDWARD ULLENDORFF, *The Ethiopians*, 2nd ed. (1965), a sometimes debatable treatment; ESTELLE SYLVIA PANKHURST, *Ethiopia: A Cultural History* (1955), well illustrated; RICHARD K. PANKHURST, *An Introduction to the Economic History of Ethiopia from Early Times to 1800* (1961).

Specific periods: J. DORESSE, *L'Empire du Prêtre-Jean*, vol. 1, *L'Éthiopie antique*, and vol. 2, *L'Éthiopie médiévale* (1957), with a description of ancient history based directly on recent discoveries of monuments (including detailed, critical bibliographies); DAVID MATHEW, *Ethiopia: The Study of a Polity, 1540–1935* (1947), excellent for the period from the Gonderian epoch to the reign of Yohannes IV; THOMAS E. MARSTON, *Britain's Imperial Role in the Red Sea Area, 1800–1878* (1961); LEONARD S. WOOLF, *Empire and Commerce in Africa* (1920, reprinted 1968); G.N. SANDERSON, *England, Europe and the Upper Nile, 1882–1899* (1965), three essential works for the colonial expansion around Ethiopia in the 19th century; MARGERY F. PERHAM, *The Government of Ethiopia*, 2nd ed. (1969), a summary of recent history with a detailed study of Ethiopian political institutions; J. DORESSE, *Histoire sommaire de la Corne Orientale de l'Afrique* (1972; Eng. trans., *History of the Horn of Africa from Antiquity to Modern Times*, in prep.), a work on the ancient history of the Ethiopian provinces of the southeast and the recent links between Ethiopia and the Somali Republic; S.H. LONGRIGG, *A Short History of Eritrea* (1945), somewhat out of date, should be supplemented with G.K.N. TREVASKIS, *Eritrea: A Colony in Transition* (1960).

(J.P.Do.)

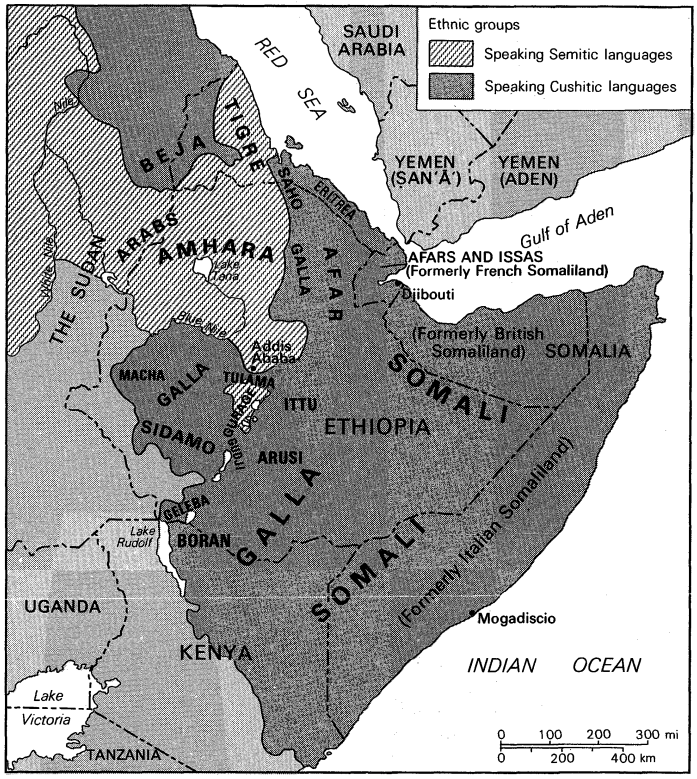
Ethiopian and Somalian Cultures

Ethiopia is a large and ancient state in Africa, with 26,000,000 inhabitants and a history of independent sovereignty extending back at least 2,000 years. Discounting the brief Italian intrusion (1935–40), it was the only traditional empire to survive the colonial partition intact. From those early centuries to the present, Ethiopia has been ruled by leaders drawn almost exclusively from the two dominant ethnic groups: the Semitic-speaking Tigre and Amhara. Ethiopia is thus essentially the political expression of the cultural nationalism of these two closely related peoples, who derive from a fusion of local Cushitic-speaking stock with South Arabian immigrants who settled along the Red Sea coast in the 1st millennium BC.

Ethnic and political origins

The Tigre occupy the northern part of the central Ethiopian Plateau, which consists today of the provinces

From J. Trimingham, *Islam in Ethiopia*



Distribution of Ethiopian and Somalian cultures.

of Tigre and Eritrea and contains the ancient capitals of the empire: Aksum, Gonder, and Lalibela. For over half a millennium, however, power has lain mainly with the Amhara, who live in the southern part of the plateau in the provinces of Begemdir and Simen, Gojam, and Shewa —the seat of government since 1889.

The central Ethiopian Plateau is a generally well-watered and pleasant land producing the indigenous cereal teff (*Eragrostis abyssinica*), as well as wheat, barley, and, in the drier regions especially, maize and millet. Cattle and other stock are reared, the land being tilled by ox-drawn plow. In the lower and hotter southern regions, the false banana, known locally as ensete (*Musa Ensete*), is the main crop, and coffee—which is indigenous to Ethiopia and may be named after a local tribe, the Kafa—is produced as a cash crop. There, for the most part, the hand hoe replaces the ox-drawn plow in cultivation. This ensete-growing region is the home of the Gurage (a Semitic-speaking but partly Cushitic people), of the Cushitic Sidamo-speaking peoples, and, to the southwest, of a few other, more distantly connected peoples, whose linguistic affiliation remains a matter of controversy.

By far the most important Cushitic group, however, is that of the Galla, who number some 10,000,000 and form the largest ethnic unit in northeast Africa. They occupy most of the southern provinces of Ethiopia, with the related Somali on their eastern and southern flanks, and seem destined to play an increasingly significant role in the political development of Ethiopia.

The Galla

The pastoral nomadism that forms their traditional way of life is best preserved today among the Boran (Borana) Galla, who live in the hot dry lowlands of southern Ethiopia and northern Kenya. It is from this region that the many different subdivisions of the expanding Galla nation began their invasion of Ethiopia in the 16th century, sweeping relentlessly in wave after wave into the Christian highlands, where the embattled Amhara and Tigre were unable to check their advance. Those Galla who moved into the rich central highlands, such as the Macha and Tulama, abandoned their nomadic economy and became sedentary cultivators; others, such as the Arusi and Gudji, in the less favoured areas sought to combine both modes of livelihood. Many became Christians, adopting Amhara culture; others embraced Islām, although this was not necessarily incompatible with adopting other aspects of Amhara culture.

The inability of the Christians to withstand the Galla invaders (over whom, however, they later re-established their ascendancy and finally incorporated as subjects in the empire) shows the extent to which Ethiopian fortunes had been reduced by the ceaseless wars of the period with the surrounding Muslim Sidamo and other largely Cushitic sultanates.

The Cushitic Somali became fervent Muslims and put increasing pressure on the southeastern flank of the Galla, thus helping to sustain the latter's continued thrust into Ethiopia. After a long period of turmoil, adjustment, and the re-emergence of Amhara ascendancy, under the forceful rule (1889–1913) of Menelik II, Ethiopia assumed its modern shape. The coast was conceded to the British, French, and Italian colonizers.

Thus, despite their crushing defeat at Adowa by the Ethiopians in 1896, the Italians were allowed to retain the largely Muslim and partly Cushitic territory of Eritrea, which was not returned to Ethiopia until after World War II.

The Somali

Notwithstanding their strong sense of cultural nationalism, the Somali had not previously formed a single political unit, and they were partitioned among the Ethiopians, Italians, British, and French. The French also acquired the closely related Cushitic-speaking Afar (or Danakil) to form their tiny colony around the port of Djibouti. In 1960 the British and Italian parts of the Somali nation became independent and joined to form Somalia (formally known since the coup of 1969 as the Somali Democratic Republic), leaving about a quarter of the total Somali population of 4,000,000 still under alien rule in the neighbouring areas of eastern Ethiopia, northern Kenya, and the minuscule French Territory of the Afars and Issas. Formerly known as French Somaliland (the Issas are a Somali clan), this arid land contains about 100,000 inhabitants, mainly nomads or urban workers.

Traditional ethnic and religious rivalries are thus perpetuated today by the coexistence of a state based on Somali identity (but not yet including the whole nation) and the ancient state of Ethiopia, which, like most of its more recently formed neighbours, includes within its borders a variety of different peoples and tribes. In the three countries the greater number of inhabitants happen to be Muslim in religion and (like the Somali) Cushitic in ethnic affiliation.

TRADITIONAL CULTURE PATTERNS

General characteristics: ethos. The world views of the Christian and Muslim peoples of the region are uncannily alike, and, even in particular aspects of belief and ritual, they reveal many striking resonances. Both believe strongly in a morally significant afterlife and are thus capable of accepting present misfortune and illness with fatalistic resignation. Both are also equally adept at seeking alternative mystical explanations of distress and of resorting to supernatural agencies to seek redress. Otherworldly fatalism is thus tempered by a healthy pragmatic concern for present well-being. Those Galla and Sidamo who retain their traditional cosmologies (and few have remained entirely unaffected by Islām or Christianity) are more firmly anchored in the present and entertain few if any hopes of eternal bliss or fears of eternal damnation.

Differences of outlook and attitudes

The most obvious contrasts in cultural outlook and ethos follow the division between cultivator and nomad. The Amhara–Tigre farmers of the highlands are sturdy, canny peasants whose ready deference to their social superiors conceals hostility and suspicion. Their strongly individualistic and shrewdly calculating attitudes suit their hierarchical but far from closed status system and kinship organization.

Specialist crafts, such as weaving, leatherworking, and ironworking, are traditionally despised, and their practitioners are associated with the evil eye. Other artisan work, unskilled manual labour, and even trade are also considered degrading. Specialist minority groups—such as the Dorse weavers or the Cushitic Falashas (the "Black Jews"), who do a considerable amount of ironwork and pottery—are thus able to establish ethnic monopolies. Similarly, the traditionally disparaged Gurage have acquired a leading role as manual workers in Addis Ababa, and, until quite recently, trade has tended to be monopolized by Muslims.

The ethos of the nomads reflects their egalitarian social structure. Their aggressive self-assertiveness contrasts strongly with the deferential respect of the Amhara peasantry. Pragmatic individualism is tempered by the wider demands of the kinship group. Intense and violent competition rages over access to the sparse resources of the environment—grass and water—on which life depends. Enmities and alliances tend to be ephemeral and shifting, the definition of friend and foe constantly changing. In this, the much divided Somalis, whose constituent fractions are regularly embroiled in savage feuds, contrast with the Galla, whose component groups place a high value on internal peace and count the killing of one of their own fellows as a sin. Yet all these warrior nomads—whether Somali, Galla, or Afar—hold a similar heroic view of life, in which prowess in battle and raiding is the essential manly virtue. The weak and vulnerable receive a certain condescending compassion that is associated with the idea that they may possess special (compensating) mystical powers. In a world in which essentially might is right, however, true prestige is accorded only to those who are manifestly strong and successful.

Descent systems

Social structures. *Territorial, family, and kinship organization.* Groups based upon descent from a common ancestor play some role in the social organization of all the peoples mentioned. Generally, it is patrilineal descent (traced through male ancestors on the father's side of the family) that is of most significance in the inheritance of property and status and in group formation. It is weakest among the Christian Amharas, who regard kinship links traced either through women or men as equally binding. This bilateral kinship system is employed to build up loosely defined clusters of kin each associated with a particular parish or part of a parish. The church-centred local community forms the primary unit in the Amhara administrative system and consists of a series of scattered hamlets rather than a clearly defined village. Through their bilateral kinship ties, its members hold (in principle at least) inalienable use rights (*rist*) in the parish's arable land. These landholding kinship groups are exogamous: descent from a common ancestor over a span of seven generations constitutes a barrier to marriage. Amhara kinship connections, with their important property and status implications, thus ramify widely and enable the poorer peasants to indulge at least vicariously in the fortunes of their distantly related higher placed kinsmen. Clans as such do not exist.

Among the other mainly Cushitic-speaking peoples (including the Semitic-speaking Gurage), descent is more strictly patrilineal, and conventional kinship groups such as patrilineages and clans occur at various orders of social grouping. Where sedentary cultivation is practiced, as among the central and northern Galla, the Gurage, and the Sidamo, there is a tendency for particular clans, or their constituent lineages, to be associated with particular localities. Communities may contain the members

of several separate descent groups, but one of these is identified with a given locality.

The extent of clan and lineage development and ramification is largely a function of the size of the tribal units involved. The Galla, for instance, comprise at least eight major tribes, ranging from the most traditional and largely pastoral Gudji and Boran in the south to the strongly Amhara-influenced, cultivating Tulama and Macha around Addis Ababa. In the large units, kinship is only an ancillary principle of association, supplementing more important ties based on shared membership of an age or generation class, or "set." According to the so-called *gada* system, all the Galla male children of the same generation formed an indissoluble fraternity; and, as members of this fraternity, they moved through the various stages of life and positions open to them. The generation sets held in turn the statuses of bachelor-warriors and later of married elders, each set occupying a given grade for a period of approximately eight years. Each generation eventually supplied the lawgivers and rulers of a given Galla tribe and then retired to make way for its successors.

Galla generation sets

This form of government was essentially democratic and republican since there was a constant rotation of officeholders and each generation set elected its own leaders who, when the time came, would rule the tribe as a whole. The traditional political system was altered radically, however, as the Galla expanded northward and many groups changed their location and economy. Political changes also were influenced by the larger Amhara centralized system and by the adoption of either Islām or Christianity, which were often invoked to legitimize new structures of authority.

The *gada* organization provided the strong sense of cohesion that even large nomadic groups such as the Boran were able to achieve. Age-set organization and descent are also significant principles in the local organization of the Sidamo people. But it is among the Afar and certainly the Somali that descent becomes the most significant principle of social and political allegiance. Ties based on common residence in a given area count for less among the northern pastoral Somali than is true of any other group in northeast Africa, and patrilineal kinship is nowhere more important or more heavily utilized in group formation. The Somali nation possesses a single, all-embracing genealogy in which every member can trace his descent often over more than 30 named generations from the legendary eponym Samaale.

Patrilineal descent, given specific range and content by contractual treaties, provides the key to the traditional Somali political and legal system. Such treaties result in the formation of distinct politico-legal units, containing a few thousand male kin (and their dependents) who have agreed to meet all liabilities in concert. On this basis, if the parties involved in certain deaths and injuries wish to avoid feuding, they can collectively adjudicate their differences through a system of mutual compensations that aim at a kind of balance of payments. So while lineages of the nomads are not identified with any locality, whenever the need arises scattered kinsmen rally together to attack their foes or defend their interests.

Political organization and stratification. In terms of stratification and degree of hierarchically organized government, the republican Somali and Galla stand at one extreme, with the Amhara–Tigreans at the other; the Afar and Sidamo lie between these two poles. Among the Somali pastoral nomads, decision-making is in the hands of ad hoc assemblies of elders, in effect all the adult men involved in a given context. Some Somali clans have traditional clan heads, usually styled "sultans," who are normally little more than *primus inter pares*, symbolizing the general (but not always actualized) unity of this level of grouping. Parallel, though slightly more formalized, arrangements exist in the *gada* system of the Galla.

Among the southern Somali, who, in contrast to their northern kinsmen, are largely sedentary cultivators in the zone between the Shebelle and Juba rivers, a slightly more stratified political system has developed. Political units there are essentially fixed territorial units, and each has a clearly defined representative hierarchy from the village elders to the tribal chief (known by the hybrid Italian–Arabic title of *capo qabila*) at the top. Before their liberation by the Italian colonizers, imported Swahili slaves and local subject Bantu groups were employed by the stronger southern Somali tribes as agricultural labourers, the superior Somali leading the life of pastoral aristocrats. Although the Somali "nobles" or freeborn were more hierarchically organized than the camel-herding nomads, decision-making was still loose and flexible and very different from the proceedings in the more strongly centralized Bantu kingdoms.

The political system of the Afar represents a further step in the direction of stratification. Throughout Afar society runs a cleavage between noble, or "red," groups and subject, or "white," groups, a type of division that also occurs among the Beja pastoralists of Eritrea. The greatest political centralization is achieved by the Afar sultanate of Aussa, a lonely oasis in the Dankali Desert.

Whatever new light archaeology or new historical discoveries may shed on these matters, there can be no doubt that the major centralized political system of the region is that of the Amhara–Tigreans—one of the longest lived states in the known history of mankind. Like so many others, its origins lie in conquest and cultural and ethnic diversity, and it has been dominated throughout its history by the familiar struggle for supremacy between the contradictory pulls of the centre and of the periphery. The emperor's title "king of kings" exultantly expresses the triumph of the central power over the many parochial seats of authority within an empire that is essentially a confederation of kingdoms and principalities. So traditionally the emperor, the "Conquering Lion of Juda"—presently the ruler of the state of Ethiopia—is essentially a war leader, and his court a peripatetic military headquarters with most of his courtiers and ministerial advisers bearing military titles. As God's vice regent on earth and styled the "Elect of God," the emperor rules by divine right, enjoying a special relationship with God that is commemorated in alleged descent from King Solomon and finds expression in imperial ritual and ceremonial. There is traditionally no fixed rule of succession, however. The claim to Solomonic descent is in principle shared among all the Amhara–Tigreans through their wide-reaching bilateral kinship system. This, of course, is the source of the system's flexibility, which enables successful local leaders to aspire to and sometimes achieve the emperorship.

The significance of the empire

Thus supreme power lies with him who can grasp and hold onto it, and to this extent, despite its strongly hierarchical appearance and the inevitable tendency toward aristocratic entrenchment, the system is a partly open one. The successful claimant, however, also requires the support of the Ethiopian church and its head, the *abuna*, who solemnly enthrones and anoints the monarch and administers the oath of fealty that binds his subject princes and leaders to him. Feudal titles are dispensed anew by each imperial incumbent. These titles usually carried with them rights known as *gult*, which entitled those who held them to exact tribute and labour service from the local peasantry, and each lord maintained his own court and subdignitaries in the style of his ultimate master, the emperor. Especially before the establishment of modern communications and military equipment, immense personal energy and skill in statecraft were required of those who held the throne. Their authority was constantly menaced by ambitious rivals and by the many religious, ethnic, and other lines of division within the empire. What is surprising is not that Ethiopia frequently dissolved into a series of warring kingdoms and principalities but that it survived as a unitary empire into the modern era.

The spectrum of political types and degrees of centralized statehood with which this discussion began provides a convenient outline of the main traditional political systems of the region. It omits, however, to mention one of its most widespread features: the existence of despised occupational groups, or "castes," associated with hunting,

fishing, and such crafts as leatherworking and ironworking. Such traditionally despised, and mystically dangerous, endogamous groups occur among all the peoples of northeast Africa and are treated similarly whether they live among the hierarchical Amhara or the democratic Somali. Although, as is usually argued by their more powerful hosts, such pariah communities may derive in part from an early (or autocthonous) population, they have almost certainly also been frequently enriched by dropouts from the very ethnic groups that hold them in such contempt.

Economic systems. Village settlements are not found among the Amhara nor among those who have been strongly influenced by their culture and way of life. The basic settlement unit in the rural areas is the nuclear or extended family (that is, the husband-and-wife family or its extension to kinsfolk, respectively) living in a few circular mud and wattle houses. The typical parish contains a core of some 20 to 30 household heads descended from a common ancestor associated with the parish and holding *rist* landrights in it. These farmers owe fealty to the feudal lord who has been allocated *gult* rights to the same estate. A local steward is responsible to the feudal lord as tax collector and organizer of corvée labour (labour extracted in lieu of taxes). Rights to land use are inherited from either parent and can be acquired by marriage. In the standard form of marriage, equal amounts of wealth (usually cattle) are contributed by each set of parents, and divorce, which is frequent, entails the redistribution of these resources. Children, nevertheless, always retain property rights through both parents, even in the case of the many less formal unions that are contracted.

Settlement patterns

The Amhara pattern of dispersed settlement is replicated fairly closely among the *ensete*-growing Sidamo as well as among most of the settled, cultivating Galla. The Cushitic-speaking Konso with their clearly defined and traditionally fortified villages and the southern Somali with similarly distinct settlements are both unusual. Southern Somali villages are based upon communal water ponds to which access is strictly controlled by membership in the local community and by participation in its blood-compensation arrangements. Among all these groups, land is not traditionally transferable except to other members of the group or to those who join it as clients. Its use entails social and politico-legal obligations, and, when it is abandoned, it reverts traditionally to the local community.

Among the Cushitic pastoralists, settlement patterns are naturally more fluid and flexible. The Boran Galla, who are less mobile and far ranging than the Somali in their movements, have generally separate grazing encampments for each type of livestock. The Boran rank their animals in the following order: lactating cows, dry cows, lactating camels, dry camels, and sheep and goats. The most senior brother in an extended family herds the first category of stock in the grazing regions best suited to it, and other brothers look after the other animals in descending order of seniority and prestige.

The pastoral Somali nomads, on the other hand, have basically only two types of herding unit. The first consists of the camels, which in the dry seasons can go without water for up to 20 days or more. These are in the charge of young unmarried men and, in the dry seasons particularly, seek good grazing, sometimes hundreds of miles from the wells to which other stock—sheep and goats, cattle and milch camels—are compelled to cling closely. These latter are herded by the family unit, consisting of a man and his wife or wives and their unmarried daughters and young sons (boys over the age of seven are out with the grazing camels learning the techniques of camel management). The family moves from pasture to pasture carrying its readily transportable tents and other equipment on burden camels, which are not, however, ridden. Pasture is not owned, and rights to wells are asserted with an intensity that increases in direct proportion to the scarcity of water and the energy expended in utilizing it. In the wet seasons, when lush pasture regions become centres of intensive grazing and settlement, the two herding units move closer together, and social life becomes more expansive and relaxed. With milk and meat in abundance, this is the time for feasts and collective rituals, the season of marriage negotiation and weddings.

Except among the Christians, all these peoples practice polygyny in one form or another, each wife and her children usually forming a separate domestic and economic unit. Marriage payments, largely in livestock among the pastoralists, are probably highest in the case of the pastoral Somali. Among these Somali, however, a bride also brings a considerable dowry in the form of a flock of sheep and goats to provide milk for the children whom she is expected to bear her husband.

Finally, a striking feature of Ethiopian economics is the complex and overlapping system of regional markets held in a given centre on a particular day of the week following a pattern that still obtains in parts of Europe. Saturday is the most popular day. Although such large centres as Gonder or Addis Ababa have daily markets, the weekly market is usually much bigger and has its own distinctive festive atmosphere.

Markets

Religion and aesthetics. The Cushitic tradition, traces of which occur even among such long and deeply Islāmized peoples as the Somali, seems best preserved today (in at least one of its original or early forms) among the Galla Boran. Here Waqa, the God of sky and earth and the creator and sustainer of life, is worshipped in prayer and sacrifice as the guardian of social morality and as the source of all things, good and bad. God's special agents on earth are the sacred dynasties, or lineages, of priests (*kallus*) who still live among the Boran and to whom all the Galla in ancient times used to send emissaries on pilgrimage. The pilgrims came to receive the blessing of the *kallu* priests, or "anointing fathers," who thus made sacred the whole traditional Galla social system. Today, among the Macha and other Galla who now live in the Ethiopian Highlands, these traditional national priests have been replaced by new spirit-possessed charismatic leaders (also called *kallus*) who express the ethnic identity of their tribesmen in the context of the Christian Amhara-dominated Ethiopian state.

The Ethiopian Orthodox Church maintains that Christ had a single, divine nature. It follows the Alexandrian rite, its liturgy being celebrated and recorded in the ancient Ethiopian language Ge'ez (from which Amharic is derived). It is led by the *abuna*, an Ethiopian priest who before 1948 was appointed by the Coptic patriarch in Alexandria but since has been appointed by the local church and emperor. The *abuna* presides over a church that, with its numerous places of worship, its richly endowed monasteries, and its ample priesthood, holds perhaps a third of the land of Ethiopia. Its all-pervasive character is such that one of every five male Christians is estimated to be in orders, and priests are expected to be married, only monks and nuns being celibate. Unordained, but not necessarily unlearned, ritual experts called *debtaras* play a crucial role in dispensing mystically efficacious cures. There is a vast hierarchy of saints and angels, chief among them being the Virgin Mary, St. Gabriel, St. Michael, St. George (patron saint of Ethiopia), and the local saints Takla Haimanot and Gabra Manfas Keddus. On the darker side, a complementary host of demons and evil spirits, many connected with a form of witchcraft (*buda*) and able to possess people and cause illness and even death, are widely feared. The protective cult of the saints, which is officially approved and encouraged, is vast and all-embracing. Every Christian has a special relationship with several saints, and the observance of saints' fasts and feasts bulks large in the Christian calendar. Spirits of every origin and provenance are accepted within the Christian cosmology, where they are naturalized in the continuous process of cultural exchange between the dominant and subject peoples. Thus, for instance, the Galla fertility spirit Atete is readily assimilated to the Virgin Mary, and vice versa (see also EASTERN CHRISTIANITY, INDEPENDENT CHURCHES OF: *Ethiopian Orthodox Church*).

Saints and spirits

Exactly the same pattern is repeated on the Muslim

side, where the cult of saints is equally well developed and the Islāmic cosmology coincides to a remarkable degree with that of the Amhara Christians. Thus among the Somali, who posthumously canonize their own lineage ancestors, saints are petitioned to remedy every distress and anxiety and are venerated as essential mediators between man and the Prophet Muḥammad and God. Again, the process of Islāmization closely parallels that of Christianization: in the case of the Arusi Galla, for instance, the Prophet himself and numerous other Muslim saints are assimilated to traditional spirits and ultimately to God (Waqa).

If the great traditions of Christianity and Islām generously open their arms to assimilate the many local cultures of the region, many elements from the local cultures in turn find their way, by the back door as it were, into the world view of the two major religions. This is very clearly seen in the mystery spirit-possession cults, which attract women and certain underprivileged categories of men particularly and which have a strongly "underground" character. Christians are especially susceptible to harassment by Muslim and pagan spirits, just as Muslims are equally open to attack by Christian demons or pagan sprites. Indeed, one of the salient features of religion in this part of Africa is the emphasis placed on spirit possession and the evil eye as explanations of misfortune that elsewhere would be ascribed to witchcraft or sorcery. The other striking feature deserving notice is the stress placed on food taboos and other commensal restrictions that are applied with particular stringency to maintain social distance from hunting and low-status craft groups. The despised and rejected community, on the other hand, may entertain similar attitudes toward those who consider themselves their superiors. Thus, just as other Ethiopians look down on the religion of the Jewish Falasha and on the ironwork and pottery trades that they follow, the Jewish Falasha themselves seek to remain apart in order to preserve their own ritual purity. Those who have close contact with, or live among, non-Jews are treated with disdain.

The presence for centuries of three world religions in the region, each with its own literate tradition, has preserved a more obvious and tangible literary heritage than in most other parts of Africa south of the Sahara, but the bulk of this literature remains largely untranslated and unknown to the outside world. More accessible are the splendid monuments of the Christian past: the celebrated rock-hewn churches of Lalibela, which were largely constructed during the reign of the 12th-century emperor after whom the city is named; the glorious castles and churches of Gonder with their famous, magnificently decorated ceilings; and, beyond this tradition, the older Aksumite antiquities. Islām too has left comparable memorials in some of the earliest mosques and tombs in such ancient cities as Harer, Zeila, and Mogadiscio (Mogadishu) but, for religious reasons, has left no illustrations comparable to those adorning Ethiopian manuscripts and church walls. Outside these literate traditions, there exists a less well explored heritage of oral literature, both in prose and poetry.

MODERN DEVELOPMENTS

The extension of educational facilities and other modernizing influences, increasingly widely disseminated by better road systems and the diffusion of the transistor radio, are beginning to make some impact on even the remotest parts of the region. On the whole, the Somali area outside Ethiopia has been more intensely and widely subject to such influences than the non-Amhara provinces of the empire. Until recently, modernization in general and education in particular have tended to be monopolies of the Amhara–Tigrean elite, which has thus sought to ensure its self-perpetuation. There have certainly been improvements in the general standard of living, as in health and other social service facilities, in recent years, but the most significant area of change has undoubtedly been the political.

Political parties have not yet been allowed to develop in Ethiopia, but the old style of imperial palace government has been profoundly modified by the innovations of Haile Selassie's reign, particularly by the establishment of an elected Parliament and a ministerial government, both of which show every sign of increasing political ambitions. In the provinces, parallel reforms have been introduced with varying effectiveness. The non-Amhara peoples of the empire are increasingly part of an essentially bureaucratic administrative system in which the impact of the central authorities and the facilities that they have to offer are much more strongly and continuously felt than in the past. Thus the remote Galla pastoralist or Sidamo cultivator tends to look to Addis Ababa as the distant but nonetheless ultimate centre of his universe.

Government reforms

The effect of increasing tribal interaction within Ethiopia has been to intensify ethnic sensibilities. Partly on this account and partly in response to the requirements of the urban situation, one effect of modern urbanization is to stimulate the growth of a vast array of urban associations, many being tribally or regionally based. Although they exist primarily to provide social security for their members in towns, they are also often equally concerned to secure progress and developments for their provincial homes. Here the urban associations of the Gurage in Addis Ababa, which have successfully accomplished major road-building operations, have exerted a pioneering influence. In the absence of political parties, the Amhara authorities seem prepared to extend a limited tolerance to ethnic associations of this kind as a safety valve for the expression of some degree of ethnic nationalism. Here the activities of the Galla association, allegedly involved in attempted coups, have clearly strained the bounds of tolerance. Despite these indications of the potential for divisive ethnic and regional separatism (which the existence of the Somali state inevitably encourages), other countervailing and crosscutting allegiances bind both Muslim and Christian together in what often seems an uneasy acceptance of the status quo. The absence in Ethiopia of clearly defined alternative interest groups, which could realistically claim legitimate leadership on some basis radically different from the present, seems to exert a powerful constraining influence.

There is no doubt that the Somali dispute—which arises from the fact that some Somalis live in Ethiopia and other states while their brothers have achieved self-determination in the Somali Democratic Republic—has on the whole helped to encourage Ethiopian integration. This is all the more remarkable given the fact that the Somalis are Muslims and that the majority of Ethiopian subject peoples (and most notably the Galla) share the same Cushitic identity. For its part, Somalia has no traditional elite, although the southern cultivating Somali (who speak a very distinctive dialect) are not customarily held in high esteem by the northern nomads. Indeed, only the relative size of lineages and their fortuitous proximity to centres of European settlement or, as in the case of the most impoverished nomadic areas, their need for alternative sources of livelihood have given some groups advantages over others in their access to modernity and its benefits. The traditional pastoral nomadic life has continued little changed and has adopted such technological innovations as it can adapt to its uses rather than being changed radically itself. Thus modern firearms, transport, and communications are employed in the nomadic life wherever appropriate; and the traditional ebullient republicanism has readily lent itself to the acceptance of the modern political party system when this form of politics was presented to the Somalis in the early 1950s.

From its inception as a social club in 1943, the Somali Youth League rapidly became the principal vehicle for the expression of Somali nationalism and was the government party from independence in 1960 until the military coup of October 1969 closed Parliament and outlawed party politics indefinitely. Its long success depended upon the delicate balance that it was able to strike (though not always entirely successfully) between the competing interests of the main lineage and clan

Somali nationalism

divisions within the nation. Every group's ambition was quite simply to obtain a seat in Parliament and to do this preferably within the framework of the organization that, being in power, had the most patronage to dispense. The principle of clan balance was so widely (if tacitly) accepted that not only were all senior government appointments distributed between the different constituent units, but this rule was also often followed even in appointing embassy staffs.

The present military regime, led by the Supreme Revolutionary Council, is probably more sincerely and unanimously devoted to the destruction of "tribal nepotism" and inefficiency than some previous civilian administrations, but its constituents judge its actions in terms of the logic of lineage advantage, and no amount of antitribal legislation, however severe, is likely to affect this very much. The present regime encourages the modernist secularizing tendency that appeals to many of the young foreign-educated elite and that deplores the wasteful excesses of the local Muslim saint cults and the general obscurantist conservatism of traditional Islām. This new radicalism (which some find compatible with a fundamentalist interpretation of Islām) favours women's education, their at least partial emancipation, and monogamy. Similar views are held by many of the younger Christian elite in Ethiopia. Finally, the strong martial tradition of both the Ethiopian and Somali cultures is today well represented in the dominant role played by the military in each country.

BIBLIOGRAPHY. EDWARD ULLENDORFF, *The Ethiopians* (1965); and MARGERY PERHAM, *The Government of Ethiopia*, 2nd ed. (1969), give good introductory accounts of the Ethiopian peoples and cultures in their historical setting. J.S. TRIMINGHAM, *Islam in Ethiopia* (1952), is a work of penetrating scholarship analyzing the relations between Muslims and Christians over the centuries. RICHARD PANKHURST, *An Introduction to the Economic History of Ethiopia from Early Times to 1800* (1961), deals with the economic relations of the various communities; while CHRISTOPHER S. CLAPHAM, *Haile-Selassie's Government* (1969), provides a uniquely informed account of the contemporary political scene from the viewpoint of the capital. There are as yet no fully satisfactory books on the social institutions of the Amharas and Tigreans. DONALD N. LEVINE, *Wax and Gold: Tradition and Innovation in Ethiopian Culture* (1965), however, gives a useful account of many aspects of rural Amhara culture; and RICHARD PANKHURST, *State and Land in Ethiopian History* (1966), provides valuable information on the crucial question of land ownership and tenure. Of the other Semitic-speaking groups, W.A. SHACK, *The Gurage: A People of the Ensete Culture* (1966), provides the only detailed account of an *ensete*-growing community. On the Cushitic-speaking peoples of Ethiopia, F.C. GAMST, *The Qemant: A Pagan-Hebraic Peasantry of Ethiopia* (1969), is a short study of an Agau group related to the Falashas; HERBERT S. LEWIS, *A Galla Monarchy: Jimma Abba Jifar, Ethiopia, 1830–1932* (1965), recounts the political history of a settled Galla group on the Gibe River; K.E. KNUTSSON, *Authority and Change* (1967), describes the religion of the Macha Galla in the context of the historical development of Galla social institutions; and G.W.B. HUNTINGFORD, *The Galla of Ethiopia* (1955), with its excellent bibliography, surveys all the existing literature on Galla ethnography and offers brief accounts of the Kafa and Janjero states. ENRICO CERULLI in *Folk-Literature of the Galla of Southern Abyssinia* (1922), indicates the range of traditional Galla oral literature. A. PAUL, *A History of the Beja Tribes of the Sudan* (1954), gives a readable account of the Beja way of life.

I.M. LEWIS, *The Modern History of Somaliland from Nation to State* (1965), relates the political history of the Somali people; the same author's *A Pastoral Democracy* (1961), analyzes in detail their nomadic economy and social institutions, amplified by *Marriage and the Family in Northern Somaliland* (1962); B.W. ANDRZEJEWSKI and I.M. LEWIS (comps.), *Somali Poetry* (1964), displays the riches of Somali oral literature; and I.M. LEWIS, *Peoples of the Horn of Africa* (1969), surveys the literature (in French, Italian, German, and English) on the customs and social institutions of the Somali, Afar, and Saho (contains an enormous annotated bibliography).

(I.M.L.)

Etna (Mount)

Mount Etna is an active volcano on the east coast of Sicily (Greek Aitne, from *aithō*, "I burn"; Latin Aetna; Sicilian Mongibello). It is the highest active volcano in Europe, its topmost point in 1971 being 10,703 feet (3,263 metres). Like other active volcanoes, its height varies: in 1865, for example, the volcanic summit was 170 feet higher than it was in the early 1970s. Etna covers an area of 600 square miles (1,600 square kilometres); its base has a circumference of about 93 miles. It is built upon marine marl and plastic clay of the Pliocene Epoch (up to 7,000,000 years ago), below which are older Eocene sandstones and limestones (of up to 54,000,000 years ago).

Its geological characteristics indicate that Etna has been active since the end of the Tertiary Period, somewhat more than 2,500,000 years ago. The Greeks created legends about the volcano, saying that it was the workshop of Hephaestus and the Cyclops or that underneath it the giant Typhon lay, making the Earth tremble when he turned. The ancient Poet Hesiod spoke of Etna's eruptions, and the Greeks Pindar and Aeschylus referred to a famous eruption of 475 BC. From 1500 BC to AD 1669 there are records of 71 eruptions, of which 14 occurred before the Christian Era. The most violent historical eruption was in 1669 (March 11–July 15), when about 990,000,000 cubic yards of lava were thrown out. The eruption took place along a fissure that opened above the town of Nicolosi, widening into a chasm from which lava flowed and from which solid fragments, sand, and ashes were hurled. The latter formed a double cone more than 150 feet high, named Monti Rossi. The lava flow destroyed a dozen villages on the lower slope and submerged the western part of the town of Catania. Efforts to divert the lava stream away from Catania were made by workers who dug a trench above the village. Historically, this seems to have been the first attempt to divert a lava stream. The effort was initially successful, but 500 heavily armed men, sent by the inhabitants of a neighbouring village who feared that the diverted lava stream might descend on them, prevented further work. Later, increased volcanic activity destroyed both villages, and the lava stream, which partly destroyed Catania, reached as far as the sea. Between 1669 and 1900, 26 more eruptions were reported. During the 20th century there have been eruptions in 1908, 1910, 1911, 1918, 1923, 1928, 1942, 1947, 1949, 1950, 1955, and 1971.

Central and lateral eruptions

Etna has had more than one active centre. A number of subsidiary cones have been formed on lateral fissures extending out from the centre and down the sides. The present structure of the mountain is the result of the activity of at least two main eruptive centres. One, no longer active, can be seen in the Valle del Bove on the east side of the mountain, a great abyss three miles wide and 2,000 to 4,000 feet deep. During the 1669 eruption, the volcanic summit collapsed, and a large circular crater, 8,000 feet in diameter, was formed. Eruptions may take place at one or more central mouths accompanied by lateral activity outward from the mouths, or the eruptions may be entirely lateral. Most of the eruptions in the 20th century have been lateral, except for that of June 29, 1955, when a moderate central eruption occurred. The one beginning on April 5, 1971, was at first near the centre but thereafter was confined to lateral activity.

Among Etna's better known ancient eruptions was that of 396 BC, which kept the Carthaginian army from reaching Catania. An eruption of 1381 sent a lava flow as far as the sea. The eruption of 1852 took many lives. That of 1928 cut the railway that runs around the base of the mountain and buried the village of Mascali. That of 1971 threatened several villages with its lava flow and destroyed some orchards and vineyards.

Life on the slopes

The mountain has three zones, one above the other, each exhibiting its own characteristic vegetation. The lowest zone, sloping gradually upward to perhaps 3,000 feet, is fertile and rich in vineyards, olive groves, citrus plantations, and orchards of cherries, apples, pistachios, and hazelnut trees. Its volcanic soil often yields five harvests a year. Several densely populated settlements, notably the city of Catania, are found on the lower slopes, but they become less frequent as the height in-

creases. Above, the mountain grows steeper and is covered with forests of chestnut, beech, oak, pine, and birch. At heights over 6,500 feet the mountain is covered with ashes, sand, and fragments of lava and slag; there are a few scattered plants such as *Astragalus aetnensis* (local name spino sento), which forms typical bushes almost a yard high, while some alpine plants manage to survive even near the top. Algae have been found near the steam outlets at 9,800 feet.

Etna has been studied systematically since the middle of the 19th century, and three observatories have been set up on its slopes; they are located at Catania, Casa Etnea, and Cantoniera.

The ascent of Etna is best undertaken in summer or autumn, seven to eight hours being required from Nicolosi. The easiest route is by the road from Catania through Nicolosi up to the height of 6,171 feet, a distance of about 21 miles. From the end of the road a cableway runs to the building called Torre del Filosofo at 9,564 feet, a Roman ruin believed to commemorate the ascent of the mountain by the emperor Hadrian.

BIBLIOGRAPHY. D. ANDRONICO, *L'Etna* (1967), includes information on flora and fauna; G. IMBO in the *Catalogue of the Active Volcanoes of the World Including Solfatara Fields*, pt. 18, *Italy* (1965), provides a topographic and morphological description of the mountain. E. RECLUS, *Les Volcans de levre* (1904) and W.S. WALTERHOUSEN, *Der Aetna* (1880) include information on the physical environment.

(G.I.)

Etruscan Language

The Etruscan language was spoken by close neighbours of the ancient Romans. The Romans called them Etrusci or Tusci; in Greek they were called Tyrsenoi or Tyrrhenoi; in Umbrian, an Italic language, their name can be found in the adjective *turskum*. The Etruscans' name for themselves was *rasna* or *raśna*.

The Etruscans lived in Italy in the region of modern Tuscany, in an area bounded by the Arno River on the north, the Tiber River on the southeast, and the Tyrrhenian Sea on the west. At one time they controlled most of an area extending south from Milan through Marzabotto and Sarsina to the Adriatic Sea north of Ancona, and to the southwest their rule extended as far as Capua, Naples, and Pompeii. For the history of the Etruscans and Etruria, see ITALIC PEOPLES, ANCIENT.

Records and scholarship. The Etruscan language is known mainly from epigraphic records originating in the Tuscan area and dating from the 7th century BC to the first years of the Christian Era. There are some 10,000 of these inscriptions, mainly brief and repetitious epitaphs or dedicatory formulas, as well as votive or owner's inscriptions on paintings in tombs and accompanying engraved figures on small artifacts such as metal mirrors. There are, however, some remarkable exceptions to the general brevity of the inscriptions, and there are important differences in their origins. The longest single text, of 281 lines (about 1,300 words), now in the National Museum at Zagreb, is written on a roll of linen that had been cut into strips and used in Egypt as a wrapping for a mummy; a clay tablet found at Capua contains some 250 words; a stone slab from Perugia has two adjacent sides elegantly engraved with an inscription of 46 lines (some 125 words); a bronze model of a liver found at Piacenza, which probably represents the Etruscan microcosm in a form used for instruction in divination, has some 45 words; and a heavy rectangular block found on the island of Lemnos in the northern Aegean has an engraving of what is probably a warrior with one inscription of perhaps 18 words surrounding the head and another of 16 words in three lines on an adjacent side. In 1964 two inscriptions on gold tablets, one in Phoenician and the other in Etruscan, were unearthed at Pyrgi.

Inscriptions

Despite many attempts at decipherment and some claims of success, the Etruscan records still defy translation. While the possibility always remains that an imaginative conjecture or a brilliant inference will suddenly provide the key to the mystery, this now seems remote. The etymological method of investigation, which ultimately depends upon the recognition of presumed cognates from related languages, seems to have failed because no clear and certain relationship between Etruscan and any other language has ever been established. The procedure sometimes called the combinatory method now appears to be the most efficacious if not indeed the only useful one. It requires, first, that note be made of anything unusual in the provenance of the object on which Etruscan writing is found (such as that the mummy wrapping came from Egypt and the Lemnos inscription from the Aegean) and likewise of anything unusual in the object itself (*e.g.*, that it is a bronze replica of a liver or the representation of a god or mythological figure). Finally, each word and phrase and formula is compared with every recurrence of the same element or elements elsewhere, and all variations in the physical and the linguistic contexts are recorded. By this means it has been possible to assign some words to grammatical categories such as noun and verb, to identify some inflectional endings, and to assign meanings to a few words of very frequent occurrence.

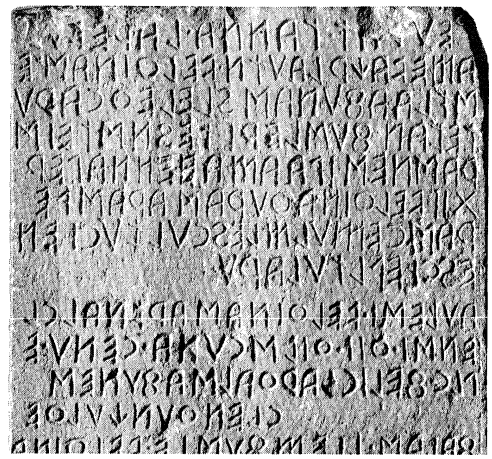


Figure 1: Etruscan inscription from a section of the "Cippus of Perugia." In the Museo Archeologico Nazionale dell'Umbria, Perugia, Italy.
Alinari

The problem of Etruscan origins is insoluble until the language can be translated. While nothing at all is certain other than the existence of Etruscans in Italy, some Etruscan writing in Egypt, and an Etruscan inscription on Lemnos, the weight of all the evidence seems to favour a non-Italic but certainly Mediterranean place of origin of the Etruscan people. It is unlikely, therefore, that the Etruscan language is genetically related to any language or language family existing in an area remote from the Mediterranean. On the other hand, it does not follow that Etruscan must be related to a language or language fragment that can be found in the Mediterranean area.

Linguistic relationships

The Etruscan alphabet. Etruscan is written in an alphabet probably derived from one of the Greek alphabets. It is of very great importance that Etruscan is written in a recognizable alphabet related to the Greek and Semitic because sound values can be assigned with some degree of precision to each symbol. Etruscan writing proceeded from right to left and in earliest times had no word division or punctuation. In about the 6th century BC a system of points, or dots, consisting of four, three, or two dots inscribed vertically, was introduced to mark word boundaries and, in some instances, apparently, to indicate syllables and possibly abbreviations.

There were four vowels in Etruscan, *i*, *e*, *a*, and *u* or *o*, and symbols in the alphabet for *p*, *t*, *c*, *m*, *n*, *l*, *r*, *z* and for the equivalents of the Greek *phi*, *theta*, and *chi*, which in Etruscan as in classical Greek were the aspirated stops *ph*, *th*, *ch* (pronounced as *p*, *t*, *k* with an added brief puff of air). There were two sibilants, written *s* and *ś*, for which the precise pronunciation is uncertain; two front fricatives, *f* and *v*, articulated either with the two lips (bilabial) or with the lower lip approaching the upper

Figure 2: Etruscan alphabet on the edge of a writing tablet from Marsiliana d'Albegna, 7th century. In the Museo Archeologico, Florence.
By courtesy of the Soprintendenza alle Antichita, Florence

front teeth (labiodental); and an *h*, which nearly always occurs at the beginning of words and is used to represent, inconsistently, the rough breathing of Greek (*e.g.*, Greek *Hēraklēs*, Etruscan *hercle* or *ercle*). There were also a *k* and a *q*, of which the precise pronunciation is unknown. A marked tendency to make all vowels in a word similar or identical (qualitative vowel harmony) is characteristic: Greek *Klutaimēstra*, which if transliterated directly into Etruscan would be *cluthemestha*, actually occurs as *cluthumustha* and *clutmsta*.

Both historical changes and dialectal differences can be observed. Diphthongs became single letters. Thus Greek *Aiwas* became Etruscan *aivas*, *eivas*, and *evas*, successively; *au* alternated with *a; eu* (like *ai*) became *e* (Greek *Kleopatra* is Etruscan *clepatra;* Greek *Poludeukēs* is, with Etruscan vowel harmony, *Pultuce*). Among consonants the most noticeable changes are *c* to *ch* to *h* (*e.g.*, *casri* becomes *chasri*, *caspr* becomes *haspr*); similarly, *p* changes to *ph* to *f* to *h* and *t* to *th* to *h*. Throughout the history of Etruscan, a first syllable usually remains unchanged, whereas later syllables tend to weaken or lose vowels, at least in writing. Older Etruscan *lavtun* "family" becomes in later Etruscan *lavtn;* other examples are *mutana* changing to *mutna*, Greek *Adonis* written *atunis* and then *atuns*, Greek *Alexandros* appearing as *elchsntre*. The consonant cluster of *elchsntre*, while extreme, is not untypical of Etruscan spelling; words thus written have led some to suggest that a very economical spelling system may have been used that was far removed from the reality of pronunciation, requiring the introduction of lightly stressed vowels in actual utterance. (For a short history and a chart of the Etruscan alphabet, see ALPHABETS: *Etruscan alphabet.*)

Grammatical characteristics. The minimal unit of meaning seems to have been a verbal root, such as *zic* or *zich*, meaning "write." The suffixing of any vowel or certain consonants (*c* or its variant *ch*, *t* or its variant *th*, *l*, *r*, or *n*) produced a noun. The vowel *u* was used to form a gerund that, without further change, could be used as an agent noun; thus *zicu* meant "writing," then, further, "scribe," and then the equivalent of the Roman name Scribonius. From these verbal nouns, denominative verbs could be made; thus from *zicu* plus *-ce*, a past tense of perfective affix, was made *zichuche* "he wrote, he has written."

Because of the large number of names occurring in the inscriptions, the noun declension system can be understood reasonably well. Similar to the process of word building is the construction called *genitivus genitivi*, or "genitive of the genitive," in which several possessive suffixes may be added to a word in succession. Thus, the simple genitive of *larth*, a proper name, is *larthal* "Larth's"; a second genitive suffix added gives *larthalisa* "of that which is Larth's."

The *genitivus genitivi*

There is apparently no grammatical gender in Etruscan. In late Etruscan an *-i* suffix marks some women's names; still later, apparently (or possibly in a different dialect area), *-ia* is similarly used. Although these suffixes appear to be the same as the final elements in some words designating exclusively female functions, such as *puia* "wife" (in one occurrence spelled *pui*, if it is the same word) and *ati* "mother" (in one occurrence *atiu*, if it is the same word), there is no evidence for any syntactic use of gender, and there is no formal marker that can be shown to have marked gender consistently.

Case endings do not differ from singular to plural; in the singular they are suffixed directly to the word stem, and in the plural they are added to the stem, along with one of the plural markers *ar, er, ur*. There is no distinctive nominative (subject) case marker, the word stem or, in some cases, the root alone serving as the nominative. A final marker *-s*, however, does appear to have been added in some instances of a probable nominative case.

The repetitive nature of most Etruscan inscriptions is such that very few distinctively different verb forms are available for analysis. Indeed, probably the only really certain verbal suffix is *-ce*. It must not be assumed, however, that the paucity of the verbal data from inscriptions reflects an impoverished verb system in the language; indeed, judging from the variety of verbal stems to which the recurring *-ce* is added, it is more likely that the Etruscan verb had a more complicated structure than the noun.

Vocabulary. Since the language is undeciphered, meaning can be assigned with certainty to only a few Etruscan words that occur very frequently in the texts. Some kinship terms are sure—among these are *ati*, "mother," *clan* "son," *śec* "daughter," *puia* "wife." Less certain but probably correct are words designating members of the larger societal organization: *lavtn* "family," *zilc* "official," *maru* "official," *spur* "city," *rasna* or *raśna* "Etruscan, Etruria." A pair of dice certainly have on them the names of the numbers from one through six. Although the order of these numbers has been and still is disputed, the arrangement most generally accepted is this: *thu* "one," *zal* "two," *ci* "three," *śa* "four," *mach* "five," *huth* "six."

Among the continuing mysteries of Etruscan are the reasons why the Etruscans left no written records of their great civilization other than inscriptions and occasional texts and why the Romans, who knew the Etruscans intimately, transmitted little or nothing to posterity about either Etruscan literature or their language, which must certainly have been spoken, or at least preserved, by some families in Rome long after the period of Etruscan greatness had passed.

BIBLIOGRAPHY. The following sources provide accurate and reliable information on Etruscan. M. PALLOTTINO, *Etruscologia*, 6th ed. (1968; Eng. trans., *The Etruscans*, 1955), the standard work; A.J. PFIFFIG, *Die etruskische Sprache* (1969), an excellent and complete statement of what is known about the Etruscan language; MURRAY FOWLER and R.G. WOLFE, *Materials for the Study of the Etruscan Language*, 2 vol. (1965), which contains the inscriptions with indexes of several kinds for easy reference—there is no grammatical information included.

(M.Fo.)

Euclid

The most prominent mathematician of antiquity, Euclid is best known for his *Elements*. Almost from the time of its writing and lasting almost to the present, this treatise on geometry has exerted a continuous and major influence on human affairs, even though almost nothing is known of its author's life. It was the primary source of geometric reasoning, theorems, and methods at least until the advent of non-Euclidean geometry in the 19th century. Euclid wrote many other works in mathematics, although some of them are lost. It is sometimes said that next to the Bible, the *Elements* may be the most translated, published, and studied of all the books produced in the Western world.

Euclid flourished *c.* 300 BC but no record of his birthplace has yet come to light. Of his life it is known only that he taught at and founded a school at Alexandria in

the time of Ptolemy I Soter, who reigned from 323 to 285/283 BC. Medieval translators and editors often confused him with the philosopher Eucleides of Megara, a contemporary of Plato about a century before, and therefore called him Megarensis. Writing in the 5th century AD, the Greek philosopher Proclus told the story of Euclid's reply to Ptolemy, who asked whether there was any shorter way in geometry than that of the *Elements*—"There is no royal road to geometry." Another anecdote relates that a student, probably in Alexandria, after learning the very first proposition in geometry, wanted to know what he would get by learning these things, whereupon Euclid called his slave and said, "Give him threepence since he must needs make gain by what he learns."

Sources of the *Elements*

Euclid compiled his *Elements* from a number of works of earlier men. Among these are Hippocrates of Chios (5th century BC), not to be confused with the physician Hippocrates of Cos (flourished 400 BC). The latest compiler before Euclid was Theudius, whose textbook was used in the Academy and was probably that used by Aristotle. The older elements were at once superseded by Euclid's and then forgotten. For his subject matter Euclid doubtless drew upon all his predecessors; but it is clear that the whole design of his work was his own. He evidently altered the arrangement of the books, redistributed propositions among them and invented new proofs if the new order made the earlier proofs inapplicable. Thus, while Book X was mainly the work of the Pythagorean Theaetetus (flourished 369 BC), the proofs of several theorems in this book had to be changed in order to adapt them to the new definition of proportion developed by Eudoxus (*q.v.*). According to Proclus, Euclid incorporated into his work many discoveries of Eudoxus and Theaetetus. Most probably Books V and XII are the work of Eudoxus, X and XIII of Theaetetus. Book V expounds the very influential theory of proportion that is applicable to commensurable and incommensurable magnitudes alike (those whose ratios can be expressed as the quotient of two integers and those that cannot). The main theorems of Book XII state that circles are to one another as the squares of their diameters and that spheres are to each other as the cubes of their diameters. These theorems are certainly the work of Eudoxus, who proved them with his "method of exhaustion," by which he continuously subdivided a known magnitude until it approached the properties of an unknown. Book X deals with irrationals of different classes. Apart from some new proofs and additions, the contents of Book X are the work of Theaetetus; so is most of book XIII, in which are described the five regular solids, earlier identified by the Pythagoreans. Euclid seems to have incorporated a finished treatise of Theaetetus on the regular solids into his *Elements*. Book VII, dealing with the foundations of arithmetic, is a self-consistent treatise, written most probably before 400 BC.

Other books of the *Elements* are not on this high mathematical level. In Book VIII, the second of the three arithmetical books, are found cumbersome enunciations, needless repetitions, and even logical fallacies. Apparently Euclid's exposition excelled only in those parts in which he had excellent sources at his disposal.

Translations and editions of the *Elements*

In ancient times, Hero and Pappus of Alexandria and Proclus and Simplicius all wrote commentaries. Theon of Alexandria (4th century AD) brought out a new revision of the work with textual changes and some additions; his version was the basis of all published Greek texts and translations until, early in the 19th century, an important Greek manuscript containing an ante-Theonine text was discovered in the Vatican. Three Arabic translations were made in the middle ages: by al-Ḥajjāj ibn Yūsuf ibn Maṭar, first for the Abbāssid caliph Hārūn ar-Rashīd (ruled 786–809) and again for the caliph al-Ma'Mūn (ruled 813–833); by Ḥunayn ibn Isḥāq (ruled 808–873), in Baghdad, whose translation was revised by Thābit ibn Qurrah (died 901); and by Naṣīr ad-Dīn aṭ-Ṭūsī in the 13th century. Euclid was first made known in the West through Latin translations of these Arabic versions.

The first extant Latin translation of the *Elements* was made about 1120 by Adelard of Bath, who obtained a copy of an Arabic version in Spain, where he travelled while disguised as a Muslim student. Adelard also composed an abridged version and an edition with commentary. Hermann of Carinthia translated Books I–XII from the same Arabic version. Gerard of Cremona (*c.* 1114–87) translated the 15 books of Euclid from the Isḥāq-Thābit version. The first Latin translation to be printed was by Johannes Campanus in the 13th century.

The first direct translation from the Greek without the Arabic intermediary was made by Bartolomeo Zamberti and published in Vienna in Latin in 1505; and the *editio princeps* of the Greek text was brought out at Basel in 1533 by Simon Grynaeus. But the most important Latin translation of this period was by Federico Commandino in 1572. The first edition of the complete works of Euclid was the Oxford edition of 1703, in Greek and Latin, by David Gregory. All texts are now superseded by *Euclidis Opera Omnia* (8 vol. and a supplement, 1883–1916), edited by J.L. Heiberg and H. Menge. The first English translation of the *Elements* was by Sir Henry Billingsley. The many later editions include Robert Simson's in Latin and English, containing Books I–VI, XI, XII, and the *Data*, in 1756; and the definitive *The Thirteen Books of Euclid's Elements* by T.L. Heath, with introduction and commentary (3 vol., 1908; 2nd ed., 1926).

Other extant works of Euclid include two belonging to elementary geometry: the *Data*, containing 94 propositions, which demonstrates that if certain elements in a figure are given, then other things are given—*i.e.*, can be determined; and a book *On Divisions* (of figures), discovered in both Arabic and Latin versions and restored and edited in 1915, which deals with problems of dividing a given figure by one or more straight lines into parts that are equal or that have given ratios to one another or to other given areas. The *Optics* of Euclid is extant in Greek in two forms, one being Euclid's own treatise, and the other a critical revision by the Greek writer Theon. The *Catoptrica* ("Reflections") is not by Euclid but is, rather, a later compilation from ancient works on the subject. The *Phaenomena*, extant in Greek, is a treatise on the geometry of the sphere for use in astronomy and is similar in content to the work, by Autolycus of Pitane, *Moving Sphere*. *The Elements of Music* is attributed to Euclid by Proclus and Marinus, the latter being another Greek commentator. Included in this work are two treatises that probably are not by Euclid: the *Sectio canonis* ("Division of the Scale"), which gives the Pythagorean theory of music with some later additions; and the *Introductio harmonica* ("Introduction to Harmony"), written by Cleonides, a student of Aristoxenus, in which an identifiable tone separates notes on the scale.

Lost works attributed to Euclid

Four lost works in geometry are described in Greek sources and attributed to Euclid. The purpose of the *Pseudaria* ("Fallacies"), it is said, was to distinguish and to warn beginning students against different types of fallacies to which they might be susceptible in geometrical reasoning. The *Porisms*, in three books, was an advanced work of which Pappus gave a summary account. Although the word means "corollaries," Euclid apparently meant a statement that was intermediate in significance between a problem and a theorem. The *Conics*, made up of four books on conic sections, corresponded in content to the first four books of Apollonius' *Conics*, although Apollonius added new theorems to his own treatment. Euclid called the conics by their previous designations, sections of a right-angled cone, an obtuse-angled cone, and an acute-angled cone, respectively; it was Apollonius who first gave them the names parabola, hyperbola, and ellipse, the descriptions of which he derived. Pappus also mentioned the *Surface-loci*, which is made up of two books that probably dealt with loci on surfaces, perhaps also loci which are surfaces, and with conic sections.

A fragment in Latin, *De levi et ponderoso*, which is included in Gregory's edition of Euclid, contains a statement of the principles of Aristotle's dynamics but is not by Euclid. See also RELATED ENTRIES under EUCLID in the *Ready Reference and Index*.

Euclid may not have been a first-class mathematician.

He certainly was, however, a first-class teacher of mathematics, inasmuch as his textbook has remained in use practically unchanged for more than 2,000 years. It was not until the middle of the 19th century that a so-called "away from Euclid" movement began and led to the appearance of a multitude of rival textbooks of geometry giving the substance of Euclid's early books in many different forms and producing thereby a variety of geometrical teaching.

MAJOR WORKS

Euclid's extant works are collected in *Euclidis Opera Omnia*, ed. by J.L. Heiberg and H. Menge, 8 vol. and a supplement (1916), containing the *Elements*, Books I-XIII (*The Thirteen Books of Euclid's Elements*, trans. by Sir Thomas Heath, 1952), *Data* (*Euclid's Data, restored to their true and genuine order*, trans. by R. Jack, 1756), *Euclid's Book on Divisions of Figures*, ed. by R.C. Archibald (1915), *The Optics*, and *The Phaenomena*.

BIBLIOGRAPHY. GEORGE SARTON, *Introduction to the History of Science*, vol. 1, pp. 153–156 (1927, reprinted, 1968), an extensive bibliography; B.L. VAN DER WAERDEN, *Outwakende wetenschap* (1950; Eng. trans., *Science Awakening*, 1954), a readable orientation to Euclid and his place in history; THOMAS L. HEATH, *A History of Greek Mathematics*, vol. 1, pp. 354–446 (1921, reprinted 1965), an authoritative survey of Euclid's works; ALBERT LEJEUNE, *Euclide et Ptolémée, deux stades de l'optique géométrique grecque* (1948); MARSHALL CLAGETT, "The Medieval Latin Translations from the Arabic of the Elements of Euclid, with Special Emphasis on the Versions of Abelard of Bath," *Isis*, 44:16–42 (1953); and GLENN R. MORROW (ed. and trans.), *Proclus' Commentary on the First Book of Euclid's Elements* (1970), three scholarly studies.

(B.L.v.d.W.)

Eudoxus of Cnidus

The greatest mathematician and astronomer in the age of Plato, Eudoxus of Cnidus substantially advanced number theory beyond the Pythagorean tradition by demonstrating the reality of irrational quantities and by his techniques for measuring curved surfaces. Moreover, with his system of concentric spheres, which was the first systematic explanation of the motions of the Sun, Moon, and planets, he re-emphasized the Greek insistence on spherical perfection. He thereby introduced geometry into the science of astronomy and began the necessary interaction between observation and theory that has characterized its development ever since. His contributions are known through ample Greek sources, including commentaries in Byzantine codices, even though none of his writings has survived.

Born about 400 BC, in Cnidus in southwest Asia Minor, Eudoxus, son of Aeschines, learned mathematics and medicine at a school that rivalled for a time that of Hippocrates of Cos. Impressed by his ability, a well-to-do physician paid his way to Athens so that he could study at Plato's Academy, which had been established in 387. He also spent 16 months in Egypt some time during the reign of Nectanebo I (380–363). At Heliopolis, now a Cairo suburb, Eudoxus learned the priestly wisdom, which included astronomy; there he wrote the *Oktaëtēris*, his first major work, concerning a calendar based on an eight-year cycle, perhaps from a study of Venus. Earning his living as a teacher, he then travelled in the region of the Sea of Marmora before returning to Athens, where he became respected throughout Greece as a legislator. The few facts concerning his life are derived largely from the writings of Diogenes Laërtius in the 3rd century AD.

It is generally agreed that Euclid drew heavily on Eudoxus in writing the *Elements*, particularly Books V and XII and sections of Books VI, X, and XIII. Eudoxus' two major contributions to mathematics are the theory of proportion, found in Book V, and the method of exhaustion, in Book XII. The philosopher Proclus attributed the first to him, and Archimedes credited him with the second. It is also possible that Euclid's axiomatic method was first developed by Eudoxus.

Theory of proportion

Around the 5th and 4th centuries BC, the Pythagoreans had observed relationships between certain geometrically measurable quantities, such as that between the hypotenuse and legs of a right triangle (*i.e.*, the square of the hypotenuse is equal to the sum of the squares of the other two sides). But following this development, they also discovered that the lengths of certain geometric figures (*e.g.*, the diagonal of a square with sides equal to one) could be expressed only as an irrational number (*i.e.*, a nonterminating decimal, such as 1.414213 . . . or $\sqrt{2}$). The discovery of irrational numbers about 400 BC meant that Pythagorean geometry, which dealt initially with commensurable quantities only, was inadequate; it had no means for dealing with irrational numbers.

Eudoxus' theory of proportion is treated extensively in Euclid's *Elements*, Book V, in which the definition of equal ratios (No. 5) is the principal source of the modern view of irrational numbers. With this theory, which was a major contribution to number theory, mathematics for the first time could take into account incommensurable quantities (*e.g.*, those whose ratio is not the quotient of two integers, such as the diameter and circumference of a circle), as well as such well-known commensurables as two sides of certain triangles. Thus, following the early Greek solution to the problem of finding areas and volumes bounded by straight lines that were represented by rational numbers, the theory of proportion allowed Eudoxus to deal, by means of rational approximations, with measurements that involved irrational numbers. He demonstrated that irrational numbers could be defined by means of approximations of rational numbers.

More difficult, however, was the problem of calculating areas and volumes bounded by curves. Eudoxus' solution was the method of exhaustion—a modern term, not used by the ancient Greeks. By successfully including in this method not only the irrational but also the concept of the infinitesimally small quantity, he showed how to subdivide continuously a known magnitude (*e.g.*, the length of a straight line) until it closely approached that of an unknown, such as the properties of a curve. According to Archimedes, Eudoxus used this method to prove that the volumes of pyramids and cones are one-third the volumes of prisms and cylinders, respectively, with the same bases and heights—although Democritus probably had already discovered that this was so. Eudoxus also demonstrated that the areas of circles are proportional to the squares of their diameters. His technique was something like, but far more rigorous than, inscribing polygons with increasing numbers of sides within a circle in order to find its area. Because it could be used to compute areas and volumes bounded by curves, the method of exhaustion, which is fully elaborated in Book XII of Euclid's *Elements*, was a forerunner of integral calculus.

Method of exhaustion

Although the geometric method of calculating the distance of the Sun and Moon from the Earth is usually attributed to Aristarchus of Samos (3rd century BC), there is a possibility that it may have been invented by Eudoxus. He also proposed a solution for another problem in astronomy—how to explain mathematically the observed motions of the Sun, Moon, and five planets. In order from the stationary Earth, these seven heavenly bodies were the Moon, Mercury, Venus, the Sun, Mars, Jupiter, and Saturn. Although Eudoxus' own account of his solution by means of concentric spheres (an idea that may have originated with Parmenides of Elea around 500 BC) has not survived, the 19th-century Italian astronomer Giovanni Schiaparelli described its details from an analysis of brief references in Aristotle's *Metaphysics* and from a substantial discussion in the commentary on Aristotle's *De caelo* (Book II), written in the 6th century AD by the Greek philosopher Simplicius.

Because the stars did not appear to have independent motion, the ancients perceived them to revolve around the Earth as a unit, forming a fixed backdrop against which the complicated and irregular motions of the planets were observed and plotted. To explain the motions Eudoxus constructed a model of 27 spheres. On the outer sphere he placed all the fixed stars, to account for their diurnal (daily) east-to-west motion. He devised a set of interconnected spheres for each of the other heavenly bodies—three each for the Sun and Moon and four each for the five planets. Each planet rode on the equator of its inner sphere except for the Moon, which rode on the

System of concentric spheres

middle; each sphere had its own appropriate axial inclination and rotational velocity. The outermost sphere accounted for the diurnal east-to-west motion.

The Moon's inner sphere was set at an angle to the ecliptic (the apparent path of the Sun through the constellations) and rotated east to west once each lunar (nodical) month (now known to be 27.2122 days). The middle sphere that carried the Moon moved west to east along the ecliptic to account for the Moon's phases. The Sun's inner sphere, moving west to east, was inclined (erroneously) to the ecliptic to account for the inequality of the seasons; and its middle sphere moved west to east along the ecliptic once each sidereal year (a round trip to the same place in the constellations).

To represent the retrograde motions of the five planets, Eudoxus depended on the combined rotatory motions of the two inner spheres of each. The resulting motion described an elongated figure-eight, which he called a hippopede. One of the few noncircular curves in antiquity, it represented the planet's "wandering" motion in latitude and longitude. As the planet, attached to the innermost sphere, oscillated along Eudoxus' hippopede during retrogressions, the hippopede itself revolved daily on the outermost sphere, while proceeding bodily once each sidereal year around the ecliptic by means of the next inner sphere (from the outside).

By devising combinations of spheres, Eudoxus accounted for the apparently irregular motions of the celestial bodies in the solar system, thus maintaining the Greek vision of circular perfection, the ancient symbol of wholeness. His system was an admirable attempt to "save the phenomena," the vivid Greek expression, in that it closely represented the observed motions of Mercury, Jupiter, Saturn, the Sun, and the Moon. But it was unsatisfactory for Venus and failed completely for Mars; nor could it account for variations in brightness, orbital period, apparent size of the Moon, or synodic period.

After his death in about 350 BC, Eudoxus' fundamental contributions to geometry and number theory led to further, lasting developments in Greek mathematics, as exemplified by the work of Archimedes. Eudoxus introduced into astronomy the view that uniform circular motion would account for the irregular motions of all the celestial bodies—the view that, with the epicyclic refinements of Ptolemy, prevailed until the time of the 17th-century astronomer Johannes Kepler. And his perception of circular regularity persists in modern mathematics in the periodic character of the infinite series.

BIBLIOGRAPHY. The classic analysis of Eudoxus' system of concentric spheres is by GIOVANNI SCHIAPARELLI, "Le sfere omocentriche di Eudosso, di Callippo e di Aristotele," in *Pubblicazioni del R. Osservatorio di Brera in Milano*, no. 9 (1873), in which he also corrected certain errors made by Simplicius. This is the basis of the study of Eudoxus by THOMAS L. HEATH in his *Aristarchus of Samos* (1913); and by J.L.E. DREYER, *A History of Astronomy from Thales to Kepler*, 2nd ed. rev. (1953), which has the clearest exposition. OTTO NEUGEBAUER, *The Exact Sciences in Antiquity*, 2nd ed. (1957), examines the question of Babylonian and Greek influences on Eudoxus. Two studies by GEORGE DE SANTILLANA, "Eudoxus and Plato: a Study in Chronology," *Isis*, 32:248–262 (1940), and "On Forgotten Sources in the History of Science," in A.C. CROMBIE (ed.), *Scientific Change*, pp. 813–828 (1963), are discussions of the intellectual interaction of Eudoxus with his contemporaries and possible sources of his ideas.

The most detailed studies of Eudoxus' mathematics are in two works of THOMAS L. HEATH: *The Thirteen Books of Euclid's Elements*, 2nd ed. rev., 3 vol. (1956), and *A History of Greek Mathematics*, 2 vol. (1921). MORRIS R. COHEN and I.E. DRABKIN, *A Source Book in Greek Science* (1948), contains relevant excerpts from the writings of Simplicius, Plato, Archimedes, Euclid, Aristotle, and Diogenes Laërtius on pertinent topics concerning Eudoxus.

(R.P.A.)

Eugene of Savoy

Prince Eugene of Savoy, an Austrian imperial general in the Austro-Turkish War of 1683–99 and the War of the Spanish Succession (1701–14), considered to have been the greatest general of his time, was the teacher of Frederick the Great and the only one among the seven great strategists of all time whose campaigns Napoleon considered worthy of study by posterity.

Eugene was born October 18, 1663, in Paris, the youngest son of the comte de Soissons, of the house of Savoy-Carignan, and of Olympia Mancini, a niece of Cardinal Mazarin. His paternal ancestors were the dukes of Savoy, who later became kings of Sardinia, Sicily, and eventually of all Italy, while on his mother's side the family included not only Roman patricians but also hatters, valets, and humble artisans. A rumour once had it that Eugene was actually the son of Louis XIV, the Sun King, who, although he always denied it, was said to have had an affair as a young man with Olympia Mancini. Louis, in any event, was ashamed of this unprepossessing supposed offspring of his. So severely did he restrain Eugene's ambitions that the latter, after spending some 20 inglorious years in Paris and Versailles left France altogether and offered his sword to the emperor Leopold I, who was busy fighting the Turks.

Military career

In his first battle, at the relief of Vienna from siege by the Turks in 1683, Eugene so distinguished himself that the Emperor gave him a regiment of dragoons to command the following year. Nothing could now retard this young genius on his way to world fame. It was more than just an outstanding strategic ability that made him victorious; it was also a contagious élan, which enabled him time and again to lead his men to triumph on the field of battle. With equal impetuosity he climbed the senior military ranks, becoming an imperial field marshal at the age of 29. He fought seven battles of major historical significance. Through the victories of Zenta, Peterwardein, and Belgrade, he once and for all secured Hungary from the Turks; together with the Duke of Marlborough, his great friend, he won Bavaria and, indeed, all of Germany at Blenheim; at Turin he gained northern Italy; at Oudenaarde and Malplaquet, the Netherlands. He lost only Spain, and that because of the political ineptitude of the emperor Charles VI.

Eugene's battles were among the bloodiest encounters of that sanguinary period, his marches among the most exhausting, and his decisions in difficult situations among the most daring and successful known to the history of war. And yet there was something still greater and rarer than his generalship in 24 battles: his wisdom in regarding military victory as a mere instrument for achieving political ends. Eugene not only made conquests but also secured them, and in his creative hands the works of peace quickly blossomed forth.

During a span of 39 years he continued to lead the Emperor's armies from the River Save all the way to Lombardy, through the Tirol back again to Bavaria and the Rhine, down once more to Hungary against the Turks, and back up again to Flanders. He was wounded 13 times. Even as he faced a world of foes before him, he had a world of enemies at his back, nourished by the hereditary curse of Austria: slothful souls and thoughtless minds, low intrigue, envy, jealousy, foolishness, and dishonesty. He served three emperors: Leopold I, Joseph I, and Charles VI. Toward the end of his life, Eugene observed that whereas the first had been a father to him and the second a brother, the third (who was perhaps least worthy of so great a servant) had been a master.

Victory at Belgrade

Belgrade (1718) was Eugene's most famous victory. In order to take the city, he had deployed his forces in a position that was both daring and dangerous. It was to no avail that others implored him again and again to withdraw the troops in order not to give the Turkish army—which was intent on raising the siege and which surrounded Eugene's forces with a superiority of four to one—the chance to undo them all. Furthermore, Eugene's soldiers were being decimated by disease. But he remained in position and, in a terrible battle that began in a thick fog in the dead of night, annihilated the relief army once the fog lifted at daylight. Capturing innumerable prisoners and the entire Turkish encampment, Eugene took the city shortly afterward.

Worn out by the exertions of an active career, Eugene died in Vienna on April 21, 1736. Posterity has endeavoured to add to his military glory the reputation of sci-

entific, artistic, and literary interests. But although he is known to have corresponded with the philosopher Gottfried Wilhelm Leibniz, his literary interests cannot have been overwhelming because not one of the tens of thousands of volumes in his library (most of them preserved in the Nationalbibliothek in Vienna) bears any trace of having been much used; and they give the appearance of having been opened hardly at all in the more than 200 years since Eugene neglected to read them. Regarding his interest in architecture and painting, it may be said that, although he commissioned great artists to build and beautify his palaces, he involved himself in these matters no more than was expected of any great lord of his time. Aside from the enormous, in fact uncanny, talent with which he led the Emperor's armies to victory, Eugene had a weakness for comic ruses. At the siege of Philippsburg, for example, he intended to change the course of the Rhine so that it would flow through the French camp, an idea that caused him to laugh uproariously; at Belgrade, he had the Austrian flotilla, which had lain at anchor above the city, move through various canals so that it suddenly appeared below the city. He also indulged in primitive soldiers' pranks—a failing that made him many enemies among his victims. The only subject that held his attention as much as politics was finances, not only public monies (he was also imperial minister of finance) but also and especially his own—he had arrived in Austria with 25 guilders, and he left an estate of 25,000,000.

BIBLIOGRAPHY. Biographies in English include NICHOLAS HENDERSON, *Prince Eugen of Savoy* (1964); GEORGE F. MACMUNN, *Prince Eugene, Twin Marshal with Marlborough* (1934); and GEORGE B. MALLESON, *Prince Eugene of Savoy* (1888). See also WINSTON CHURCHILL, *Marlborough*, vol. 2 (1934).

(A.L.-Ho.)

Eugenics

Eugenics is the study of human improvement in all aspects by genetic means. Its goal is to increase the proportion of persons with better than average genetic endowment. It draws from the science of genetics for an understanding of the processes of heredity; from psychology for analysis of the part played by differences in heredity and environment in the development of personality, intelligence, and character; from medicine and medical genetics for information on hereditary defects and susceptibility to disease; and from sociology, and more particularly demography, for the rates of births and deaths, the mating habits, and the social and physical factors that determine the relative increase or decline of persons with different characteristics. In its scientific aspect it is essentially the study of trends and causal factors in human evolution.

A companion science, euthenics, is also concerned with the improvement of mankind, but by adjustment of the environment. Obviously both concerns are essential to the well-being of man, and since neither science can stand alone, the term "humanics" has been suggested to embrace both aspects of human improvement.

HISTORICAL BACKGROUND

Early history. The idea of applying knowledge of heredity to the improvement of the human race goes back to earliest times. References to eugenic ideals appear in the Old Testament, and Plato's *Republic* idealizes a society in which there is constant selection for the improvement of the human stock. A British economist, Thomas Robert Malthus, noted the struggle for existence; an English naturalist, Charles Darwin, saw in it the means of evolution. It remained for Francis Galton, a scientist of many talents, and a cousin of Darwin, to see that the theory of evolution implied that man might in part direct his own evolutionary future. Galton's first important work, *Hereditary Genius* (1869), contained the results of his studies of the families of eminent men as evidence for his belief that "it would be quite practical to produce a highly gifted race of men by judicious marriages during several consecutive generations." In 1883 he published *Inquiries into Human Faculty*, in which he coined the term eugenics. In *Natural Inheritance* (1889), Galton pioneered the development and application of advanced statistical methods to the study of man.

While Galton's thinking was much ahead of his time, he retained some of the prejudices of an English gentleman in regard to social class and race. In his studies of the families of eminent men Galton underestimated the effects of a favourable environment. But his faults in this respect were far less than those of many of his followers. He developed and used many of the statistical methods for the study of populations and was the first to recognize the value of the study of twins for research in heredity. He repudiated the Lamarckian view, then widely held, that acquired characteristics were inherited. Although Galton attached religious significance to the eugenic movement, he did not think of it in revolutionary terms, but rather as ". . . the new duty which is supposed to be exercised concurrently with and not in opposition to, the old ones upon which the social fabric depends."

The slow acceptance of eugenical ideas

For all his wisdom and high ability, Galton was not able to gain any wide acceptance for eugenics, largely because much of the scientific and technical foundation was lacking. Psychology was not sufficiently advanced to help him in his studies of individual differences or in assessing the importance of the environment; there was no cultural anthropology to provide an outline of the growth of civilizations; and birth control was so little recognized or accepted that he did not take it into account as a possible factor in the control and distribution of births.

From Galton to World War II. Galton had endowed a research fellowship in eugenics in 1904 and, in his will, provided funds for a chair of eugenics at University College, London University. The fellowship and later the chair at University College were both occupied by Karl Pearson, a brilliant mathematician who helped to create the science of biometry, the statistical aspects of biology. Pearson was a controversial figure who believed that environment had little to do with the development of mental or emotional qualities. He felt that the high birthrate of the poor was a threat to civilization and that the "higher" races must supplant the "lower." His views gave countenance to those who believed in racial and class superiorities. Thus, Pearson shares the blame for the discredit later brought on eugenics in the United States and for making possible the dreadful misuse of the word eugenics in Hitler's propaganda. The English Eugenics Society, founded by Galton in 1907 as the Eugenics Education Society, opposed Pearson's views, but was itself slow in throwing off the prevalent biases of that time.

Eugenics discredited by early racist views

In the United States the American Eugenics Society was founded in 1926 by men whose views were coloured by the times in which they lived. They believed that the white race was superior to other races, and, further, that the "Nordic" white was superior to other whites. They thought of races as "pure" groups, quite separate from each other. They did not know that all races are mixtures of many types, the distribution of genes among the races varying in proportions rather than in kind, as evidenced by the distribution of the various blood groups among all of the races. They did not realize that environments are so uncontrollable that scientists cannot reasonably put forward views on the genetic differences in the performance of different races. Furthermore, they believed that the upper classes had superior hereditary qualities that justified their being the ruling class.

The science of that day supported extreme views on feeble-mindedness and "criminal types." Intelligence tests, introduced in the early 1900s by a French psychologist, Alfred Binet, were thought, contrary to Binet's views, to be measures of innate, genetic intelligence. People whose test performance (*e.g.*, intelligence quotient or IQ) gave them a mental age of 12 or less could be classified as feebleminded or morons, even though the term would then include most of those brought up in deprived environments, such as the southern Appalachian highlands or the blighted rural and urban slums occupied by the descendants of slaves. Criminality was generally considered a concomitant of feeblemindedness. Studies on

degenerate family stocks were taken to prove that hundreds of persons in each of these families were feebleminded or criminal types because of the inheritance they received from a single ancestor five or six generations back. Claims were made that immigrants from southern and eastern Europe, besides being socially inferior, included a remarkable number of criminal and defective stocks. There was much in the intellectual atmosphere of the U.S. that fostered an extreme hereditarian racist view.

Even before 1905 eugenists had been active in the effort to restrict immigration. Their arguments, along with those of others who advocated restrictions for different reasons, culminated with the passage of the National Origins Quota Law in 1924, which limited quota immigrants to about 150,000 annually, with no more than 2 percent of each nationality according to the number of persons of their national origin in the U.S. as of 1890, and providing that beginning July 1, 1929, the quota of any country shall have the same ratio to 150,000 as the number of persons of that national origin living in the U.S. had to the total population of the U.S., as determined from the 1920 census. In later years it became clear that the material the eugenists had presented to congressional hearings had little scientific foundation.

Movement to sterilize defectives

Eugenists at this time laid great stress on the importance of sterilizing defective persons. By 1931 sterilization laws had been enacted by 27 states in the U.S., and by 1935 sterilization laws had been passed in Denmark, Switzerland, Germany, Norway, and Sweden. Most of these laws provided for the voluntary or compulsory sterilization of certain classes of people thought to be insane, idiotic, imbecilic, feebleminded, and epileptic; some applied equally to habitual criminals, moral perverts, or the feebleminded. In most cases the purpose was clearly eugenic, though some laws tacitly permitted sterilization for social rather than genetic reasons. In the U.S. most states did not generally enforce these rather extreme measures, and the number of sterilizations was seldom over 100 per year. Exceptions were California, where sterilizations averaged over 350 cases per year, with a total of 9,931 by 1935, and some of the southern states, with fairly high sterilization rates relative to their populations.

With the advance of science the background for these laws began to be questioned. In 1935 a committee of the American Neurological Association reported on an extensive investigation and evaluation of the "facts and theories which constitute the subject matter of the inheritance of the mental diseases, feeblemindedness, epilepsy and crime." The committee concluded that there was a serious question whether "many of the eugenic proposals now current take into account the newer genetic data" and that "though the avowed purpose of such sterilization invariably starts out by being a eugenic one, it often ends by becoming a social one." They exposed many fallacies in the reasoning behind the sterilization laws and concluded that there was no sound basis for sterilization on the basis of immorality or character defect. They recommended sterilization only with the consent of the patient, or those responsible for him, and only in the case of certain specified diseases or defects, namely: (1) Huntington's chorea, hereditary optic atrophy, familial cases of Friedreich's ataxia, and a few other rare hereditary defects; (2) feeblemindedness common in a family; (3) schizophrenia; (4) extreme cases of manic depressive psychosis; and (5) epilepsy, mainly on the ground of its social aspects. Their recommendation for concerted, coordinated, and planned long-term research was not followed seriously until after 1945, when medical schools began active research into medical genetics.

In Europe the studies of criminal types were discredited, and serious work in eugenics continued in the form of family studies by government institutes. Only in Nazi Germany did eugenics serve racial and political purposes. There, Hitler, exploiting the concepts of the master race and the superman as expounded by some German scholars, developed racism into a powerful political weapon.

Eugenics after 1945. With the development of atomic weapons, the general public and medical men alike became educated to the dangers of genetic mutations produced by radioactive fallout from nuclear explosions. Medical X-rays came under new scrutiny, and the findings of studies on hereditary factors in diabetes and other diseases received new attention. Medicine and public health, professions previously little interested in heredity, recognized a new responsibility. By the 1960s the major medical schools in the U.S. had included geneticists on their research staffs, and a number had set up departments of medical genetics. Couples planning marriage became more aware of their responsibility to potential offspring, and heredity clinics associated with medical schools began to provide counselling to advise couples on the chances of having defective children.

Still other factors arose to increase public interest in heredity. The great increase in births that took place in the U.S. and most European countries after World War II aroused interest in the type of children being born and the kind of homes they were being brought up in. Birth control, given worldwide publicity as a solution to the problem of too rapid growth of populations, raised questions as to the type of people who would use it, and the kind of children they would have. Publicity was given to the practice of insemination of women whose husbands were not able to give them children, and questions were raised about the heredity as well as the legal status of children so conceived, and about the effect of the psychological environment on the marriage relationship and on the child. New scientific findings were widely disseminated, and the public began to take a more balanced view of the contributions made by both heredity and environment to good health, intelligence, and character.

Genetics made unparalleled advances after 1950, opening up a whole new field of the biochemistry of gene structure and gene behaviour. At the same time, the scientific basis for the new eugenics was being further broadened by advances in psychology and in the field of demography. The Institut National d'Études Démographiques in France sponsored important investigations of human populations. In England the Royal Commission on Population made its report in 1949, and the Population Investigating Committee at the London School of Economics gave attention to quantitative and qualitative aspects of population trends in its publication *Population Studies*. In the U.S. a number of the larger universities established facilities for training and research in demography, and several nation-wide surveys on factors affecting fertility were taken. The atmosphere was appropriate to a balanced reconsideration of eugenic problems.

Scientific basis of eugenics broadened

BASE LINES OF MODERN EUGENICS

Genetic defects. It is generally accepted that genes causing defects originate in mutations that occasionally take place in one or another of the many thousands of genes present in pairs in every human cell. Usually the mutated gene is recessive; that is, it does not have a harmful effect unless the other gene in the pair has the same characteristics. The chances of two such genes meeting as a pair depend on the number of such genes distributed throughout the intermarrying population. Thus under random mating, if a particular deleterious gene is carried by one person in a hundred, the chances of the mating of persons with such genes are 1 in 10,000. While the chances of such an unfortunate mating seem slim, in a large population such a ratio results in a considerable number of defective persons.

With an increase in the proportion of carriers of a particular defective gene, there is an even greater increase in the likelihood of a mating in which two such genes will be paired in the same fertilized cell. When this happens, there will be a defect. If the defect is lethal (*i.e.*, results in stillbirth or infant death), the two deleterious genes will be removed from the population. If the defect is minor, but of a sort to make marriage or reproduction less likely, the genes are to that extent less likely to survive. Thus at some point nature establishes a balance in which, for every new deleterious gene brought into circulation, a similar gene is lost from circulation.

Geneticists believe that most people carry at least a few deleterious genes; some estimate the average to be as high

as 8 percent, and at least 2 percent of the babies born carry some genetic defect, major or minor.

Most scientists believe that the proportion of carriers of deleterious genes may now be increasing, though the results in the form of an increase in defect would not be apparent for a number of generations. Greater exposure to X-rays and to fallout from atomic explosions is currently accelerating the rate of mutations, while medicine is making it possible for an increasing number of persons with some kind of genetic defect to live long enough to rear children unfortunate in having that parental defect. Diabetics, for instance, formerly died at an early age, and the genes for diabetes were lost with them. After an insulin treatment was found, these people led longer useful lives, but were not generally able to have children. Now safe delivery of the children of diabetic mothers is commonplace. The deleterious genes remain in circulation, while new ones are constantly added by mutations. Thus, with good medical care, diabetics live normally and often bear diabetic children of their own. The accumulation of certain severe types of deleterious genes could conceivably reach a point that might seriously threaten man's future. To support and maintain defective members in the population is not at odds with the goal of genetic improvement of mankind so long as such persons do not reproduce defective offspring.

The genetic burden

While present social and medical conditions favour the increase of many genes that cause physical defects and susceptibility to certain diseases, other genes that affect mental disease and mental defect are probably diminishing as a result of the low marriage rate and the institutionalization of the carriers. It is probable that efforts at control will be strengthened by increased use of sterilization in some areas, by increased use of more effective means of birth control, and by a more eugenic use of heredity counselling clinics. Each year increasing numbers of defects are being defined, the type of transmission understood, and, in many cases, means found for spotting carriers. The immediate solution to the dilemma seems to lie in educating those persons known to be carrying a seriously defective gene to their societal responsibility, and thereby, hopefully, to convince them to surrender their right to bear children who might be defective and burden them as well as society. The possibility of "correcting" nature's mistakes at their source, the deoxyribonucleic acid molecule (DNA) of the gene, is a hope for the future. Such genetic engineering has been proposed by several prominent scientists and termed "algeny."

Psychological traits. With responsibility for work on mental and physical defects transferred to the professions of medicine and public health, eugenists are now giving major attention to the genetic factors that provide the base for the development of such complexly inherited traits as intelligence, character, and personality. The genetic base for these psychological traits is provided by multiple genes in various combinations. They may be broken up and recombined with other genes from one generation to another, making them difficult to trace in family lines. It will take time to locate these genes and trace the processes involved in their transmission. Fortunately, the efficacy of selection does not depend on a full knowledge of the genetic mechanism. For the present, at least, eugenists will be guided by familial incidence of intelligence and personality and by the survival of family stocks of different types.

Massive evidence from studies on identical and fraternal twins, nontwin siblings, and adopted children reared away from their natural parents indicates that differences in genetic factors play an important part in psychological differences between individuals. These were Galton's findings, and they have been confirmed by larger and more carefully controlled studies. Identical twins are extremely similar; even when reared apart from infancy, they resemble each other in physical as well as mental characteristics to a degree greater than that shown by fraternal twins reared together.

Children tend to be like their parents in intelligence not only because of their common heredity, but also because they usually share a common environment. The evidence on differences in personality is not so clear, largely because these traits are more difficult to measure. All studies indicate, however, that personality, like intelligence, is the product of interaction between environment and the genetic base. This is shown in the extremes of personality, like schizophrenia, which can often be measured.

Environmental and socioeconomic factors. As work in the field of individual differences has progressed, scientists have come to feel that the old distinction between heredity and environment is no longer useful; both are a part of the processes that begin when the cell is fertilized and starts its separate life; both operate throughout the individual's life; and both contribute to making each individual different from every other individual. Genetic factors can be given value only in relation to specific environments; some qualities that are socially valuable in one environment may be harmful in another.

Apparent racial differences in intelligence

Racial groups differ in their average response to intelligence tests, but these tests invariably involve environmental factors that favour those culturally more fortunate. Even when differences in such factors as place of residence, income, or length of schooling can be controlled, the results may still be influenced by factors affecting motivation and behaviour.

In studies of socioeconomic, or other, groups, there is no evidence of genetic differences, except perhaps in the case of comparisons between certain of the professional classes and the rest of the population, in which there is some indication that the genetic factors for response to intelligence tests may be slightly superior. Differences between socioeconomic classes in intelligence-test response are obviously strongly influenced by differences in environment in much the same way as occurs among racial groups. It is as yet impossible to measure the relative genetic potentialities of any large population groups. But differences between individuals within each group can be measured, and it is these differences that are the proper concern of eugenists today.

Until very recently, from 30 to 50 percent of all children died before maturity, and death was an important factor in the survival of particular stocks. The evidence from some areas where people are still living under primitive conditions seems to favour the belief that under conditions of life as it was everywhere before the mid-18th century, the general tendency was toward the survival of more competent stocks. If so, this situation changed in the industrialized countries, when the death rate began an accelerated decline, dropping to 12 or less per thousand by 1940. In the mid-20th century, about 95 percent of all children lived to age 30, and differential deaths had little effect on differential survival.

Social-class differences in family size

While this change was taking place, social-class differences in births were widening with the introduction of birth-control measures, which reached first into the more educated and wealthy classes. With differential births greatly favouring the less educated, who made little effective use of birth control, the inverse relation between births and education or socioeconomic status widened rapidly. In the United States this inverse relationship reached its peak with the postponement of births and marriages in the last years of the Great Depression. For the period 1935–40 women with one to three years of college education had a net reproduction rate 33 percent below that required for replacement of their numbers, and women with four years of high school had a rate 26 percent below replacement, while women at the elementary school level had a rate twice that of the college women and well above that required for replacement.

The upsurge of births after World War II and the spread of birth control among the less educated changed this trend. By 1960, among women 35 to 40 years old, who had practically completed their childbearing, high school women and women with one to three years of college had borne more children than the number needed for replacement. Women with less than eight years of elementary school education had had about the same number of children as those who had completed their childbearing during the depression. This trend seems likely to continue.

Extensive studies made about 1960 on national samples of the U.S. population indicate that the number of children actually desired by couples in the "blue collar" (factory or industrial) group is less than the number desired by couples in the "white collar" (office or business) group. When both the husband and the wife were college graduates, only 7 percent of the last pregnancies were unwanted, while among the couples at grade school level 33 percent had not wanted their last pregnancy. Of the college couples, 92 percent effectively used contraception, while only 69 percent of the grade school couples used contraception, and they used it less effectively.

It is expected that birth control may soon be used as widely and as effectively at the lowest educational levels as at the highest. Should that occur, present trends suggest that educational and social class differentials in births may soon bear a direct relation to size of family. The public generally, and many scientists, have assumed that an inverse relation between size of family and education or socioeconomic class must represent an unfavourable, or dysgenic, trend, even though there is little evidence for genetic differences between classes. If the relation between size of family and educational level becomes direct, with the more educated groups having the greater proportion of children, the trend would be generally taken as favourable, or eugenic. But its greatest significance would be an increase in the proportion of children brought up in above-average home surroundings. The major opportunity for genetic improvement will be found in selection of individuals within each group.

There is some limited evidence that at the present time in the United States the abler and the more intelligent members of any group tend to have the most children. In studies made as early as 1927 on graduates of Yale and Harvard universities, and again on Princeton University graduates in 1939, the men were rated by a committee of classmates for degree of success some years after graduation. In each group the highest-rated married men averaged over two children apiece, while the lowest-rated married men averaged less than one and one-half children. Many studies have indicated that the probability of being childless increases as IQ decreases. There is thus some evidence that in attempting to increase the comparative fertility of the more intelligent and successful individuals, eugenists are working with and not against present trends, as has been claimed by some.

Individual voluntary selection on the increase

A number of changes taking place in industrial societies today are favourable to individual selection. There are more opportunities for finding work appropriate to one's genetic capacities. There is a wider choice of mates in a highly mobile society, which favours the production of novel types and offers greater possibilities if there is selection. The trend toward more education for all classes would help make birth control selective within each class. Also, the continuing improvement of birth control methods should improve voluntary forms of selection at all levels of education. To further these trends, eugenists urge better training for choice of mate and for marriage, a more general individual and community sense of responsibility for the parents who have more than one or two children, and continuing study of all the psychological, social, and economic pressures and rewards that would tend toward a voluntary and favourable selection of births within every occupational group.

EUGENICS PROPOSAL

Recent trends in the distribution of births suggest the possibility of the development in modern industrial societies of eugenic selection of a wholly voluntary and unselfconscious sort, in which the only controls might be certain moderate changes in social institutions, most of which changes are already desired for other than genetic reasons. Recent eugenic thinking seems to be developing along these lines, but the public and many scientists are still disturbed by the older conception of a controlled or authoritarian eugenics that would threaten other social and ethical values.

The possibilities. The most immediate and most publicized procedure that can be eugenically applied is the use of insemination techniques. The insemination of women whose husbands are unable to give them children is quite frequently practiced in the United States and Europe. A proposal was made by Herman Muller, a U.S. geneticist, that the sperm for insemination be collected from men of high attainment, of family stocks as free as possible from defective strains, and that it be kept, frozen and stored, for as long as possible, perhaps beyond the lifetime of the donor. It would then be available for women who desired to assure themselves of superior children. It was Muller's belief that the advantages of this procedure would result in its rapid and widespread adoption, and an organization was formed to carry on his work.

The "sperm bank"

Other scientific possibilities, not yet in the stage of application, are already being discussed by scientists for their possible effect on the genetic qualities of human beings. Changing hereditary factors in the human cell by chemical means is not an impossibility, but it lies in the future. Another possibility is the implanting of a specified segment of genes into a chromosome. Preliminary experiments have been performed on mice, but it will be a long time indeed before such methods can be used for the improvement of human germ cells. A nearer possibility is vegetative, or clonal, reproduction, similar to the reproduction of cuttings from plants. If this not improbable technique becomes practicable with humans, the way would be open for the reproduction, in any number desired, of individuals as similar genetically as are identical twins. Human beings could then be propagated for diverse valuable characteristics of every sort. Few persons would regard this proposal as desirable or even thinkable. The technical difficulties and social and moral objections to the "mass production" of such twin types make its application unrealistic even in the remote future.

For all practical purposes the measures discussed above are premature. They are just beyond the realm of theory, but far short of general acceptance even as theories. They would violate many established social attitudes. They raise many moral and legal questions. The population at large generally regards such proposals with suspicion. Their fears cannot be dismissed lightly. Who can assure them that eugenical procedures will not be used ulteriorly? Who can promise that there is no danger?

The public's fear of eugenics

Authoritarian controls would have to be employed and extremely difficult value judgments would have to be made. Who, for example, would determine a superior quality in man? Who would make the decisions whether certain qualities should be preserved or extinguished? Modern eugenics is little interested in authoritarian controls. Rather, it believes that the social and economic environment can be shaped so as to influence a eugenic distribution of births throughout entire populations in a voluntary and largely unconscious process of selection.

BIBLIOGRAPHY

Classic works: C.P. BLACKER, *Eugenics: Galton and After* (1952); THEODOSIUS G. DOBZHANSKY, *Mankind Evolving: The Evolution of the Human Species* (1962); FRANCIS GALTON, *Hereditary Genius: An Enquiry into Its Laws and Consequences* (1869); THOMAS HUNT MORGAN, *The Theory of the Gene* (1926); J. SUTTER, *L'Eugénique* (1950).

General works: R. FREEDMAN, P.K. WHELPTON, and A.A. CAMPBELL, *Family Planning, Sterility, and Population Growth* (1959); FREDERICK H. OSBORN, *Preface to Eugenics*, rev. ed. (1951); C.F. WESTOFF *et al.*, *Family Growth in Metropolitan America* (1961); C.F. WESTOFF, R.G. POTTER, JR., and P.C. SAGI, *The Third Child: A Study in the Prediction of Fertility* (1963); P.K. WHELPTON, A.A. CAMPBELL, and J.E. PATTERSON, *Fertility and Family Planning in the United States* (1966); P.K. WHELPTON and CLYDE V. KISER (eds.), *Social and Psychological Factors Affecting Fertility*, 4 vol. (1946–54); ROGER J. WILLIAMS, *Biochemical Individuality: The Basis for the Genetotrophic Concept* (1956).

(F.H.O.)

Euler, Leonhard

The most prolific mathematician of the 18th century, Leonhard Euler made decisive and formative contributions to the subjects of geometry, calculus, mechanics, and number theory. He developed methods for solving outstanding problems in observational astronomy and

demonstrated useful applications of mathematics in technology and public affairs.

Born in Basel, Switzerland, on April 15, 1707, Euler's mathematical ability earned him the esteem of Jean Bernoulli, then one of the first mathematicians in Europe, and of his sons Daniel and Nicolas. In 1727 he moved to St. Petersburg, where he became an associate of the St. Petersburg Academy of Sciences, and in 1733 succeeded Daniel Bernoulli to the chair of mathematics.

By means of his numerous books and memoirs that he submitted to the academy, Euler carried integral calculus to a higher degree of perfection, developed the theory of trigonometric and logarithmic functions, reduced analytical operations to a greater simplicity, and threw new light on nearly all parts of pure mathematics. Overtaxing himself, Euler in 1735 lost the sight of one eye. Then, invited by Frederick the Great in 1741, he became a member of the Berlin Academy, where for 25 years he produced a steady stream of publications, many of which he contributed to the St. Petersburg Academy, which granted him a pension. In 1748, in his *Introductio in analysin infinitorum*, he developed the concept of function in mathematical analysis, through which variables are related to each other and in which he advanced the use of infinitesimals and infinite quantities. He did for modern analytic geometry and trigonometry what the *Elements* of Euclid had done for ancient geometry, and the resulting tendency to render mathematics and physics in arithmetical terms has continued ever since. He is known for familiar results in elementary geometry; for example, the Euler line through the orthocentre (the intersection of the altitudes in a triangle), the circumcentre (the centre of the circumscribed circle of a triangle), and the barycentre (the "centre of gravity," or centroid) of a triangle. He was responsible for treating trigonometric functions—*i.e.*, the relationship of an angle to two sides of a triangle—as numerical ratios rather than as lengths of geometric lines and for relating them, through the so-called Euler identity ($e^{i\theta} = \cos\theta + i\sin\theta$), with complex numbers (*e.g.*, $3 + 2\sqrt{-1}$). He discovered the imaginary logarithms of negative numbers and showed that each complex number has an infinite number of logarithms.

Contributions to pure mathematics

Euler's textbooks in calculus, *Institutiones calculi differentialis* in 1755 and *Institutiones calculi integralis* in 1768–70, have served as prototypes to the present because they contain formulas of differentiation and numerous methods of indefinite integration, many of which he invented himself, for determining the work done by a force and for solving geometric problems; and he made advances in the theory of linear differential equations, which are useful in solving problems in physics. Thus, he enriched mathematics with substantial new concepts and techniques. He introduced many current notations, such as Σ for the sum; $\int n$ for the sum of divisors of n; the symbol e for the base of natural logarithms; a, b, and c for the sides of a triangle and A, B, and C for the opposite angles; the letter "f" and parentheses for a function; the use of the symbol π for the ratio of circumference to diameter in a circle; and i for $\sqrt{-1}$.

After Frederick the Great became less cordial toward him, Euler in 1766 accepted the invitation of Catherine II to return to Russia. Soon after his arrival at St. Petersburg, a cataract formed in his remaining good eye, and he spent the last years of his life in total blindness. Despite this tragedy, his productivity continued undiminished, sustained by an uncommon memory and a remarkable facility in mental computations. His interests were broad, and his *Lettres à une princesse d'Allemagne* in 1768–72 were an admirably clear exposition of the basic principles of mechanics, optics, acoustics, and physical astronomy. Not a classroom teacher, Euler nevertheless had a more pervasive pedagogical influence than any modern mathematician. He had few disciples, but he helped to establish mathematical education in Russia.

Euler devoted considerable attention to developing a more perfect theory of lunar motion, which was particularly troublesome, since it involved the so-called three-body problem—the interactions of Sun, Moon, and Earth. (The problem is still unsolved.) His partial solution, published in 1753, assisted the British Admiralty in calculating lunar tables, of importance then in attempting to determine longitude at sea. One of the feats of his blind years was to perform all the elaborate calculations in his head for his second theory of lunar motion in 1772. Throughout his life Euler was much absorbed by problems dealing with the theory of numbers, which treats of the properties and relationships of integers, or whole numbers (0, ±1, ±2, etc.); in this, his greatest discovery, in 1783, was the law of quadratic reciprocity, which has become an essential part of modern number theory (*q.v.*).

Lunar theory

In his effort to replace synthetic methods by analytic ones, Euler was succeeded by J.-L. Lagrange. But, where Euler had delighted in special concrete cases, Lagrange sought for abstract generality; and, while Euler incautiously manipulated divergent series, Lagrange attempted to establish infinite processes upon a sound basis. Thus it is that Euler and Lagrange together are regarded as the greatest mathematicians of the 18th century; but Euler has never been excelled either in productivity or in the skillful and imaginative use of algorithmic devices (*i.e.*, computational procedures) for solving problems.

Euler died in St. Petersburg on September 18, 1783.

BIBLIOGRAPHY. A bibliography of Euler's works, compiled by GUSTAV ENESTROM in "Verzeichnis der Schriften Leonhard Eulers," *Jahresbericht der Deutschen Mathematiker–Vereinigung, Ergänzungsband IV* (1910–13), lists 856 books and articles. An edition of *Leonhardi Euleri Opera omnia*, begun in 1911 and still continuing, will comprise 72 volumes in three series: mathematics, 29 vol.; mechanics and astronomy, 31 vol.; physics and miscellaneous, 12 vol. The texts appear in the original language of publication, in most cases Latin; but introductions, sometimes substantial, are in the language of the editors. For example, in series 2, vol. 11, there is a long historical account, comprising part 2 of the volume, by CLIFFORD TRUESDELL, of *The Rational Mechanics of Flexible or Elastic Bodies*. There is a large body of secondary literature on Euler, chiefly in German and Russian. See, for example, OTTO SPIESS, *Leonhard Euler* (1929). The most authoritative account in English, together with an extensive bibliography, is that by A.P. YOUSCHKEVITCH in the *Dictionary of Scientific Biography*, vol. 4, pp. 467–484 (1971). Less substantial summaries are found in ERIC T. BELL, "Analysis Incarnate: Euler (1707–1783)," in *Men of Mathematics*, ch. 9 (1937, reprinted 1961); and CARL B. BOYER, "The Age of Euler," in *A History of Mathematics* (1968). For Euler's contributions to specialized aspects of mathematics, see such works as CARL B. BOYER, *History of Analytic Geometry* (1956) and *History of the Calculus and Its Conceptual Development* (1959); LEONARD E. DICKSON, *History of the Theory of Numbers*, 3 vol. (1919–23, reprinted 1952); and ISAAC TODHUNTER, *A History of the Progress of the Calculus of Variations During the Nineteenth Century* (1861, reprinted 1962).

(C.B.B.)

Euphorbiales

The Euphorbiales order is a large assemblage of dicotyledonous flowering plants (*i.e.*, it is a member of the class Dicotyledoneae) consisting of seven families and about 315 genera and 8,000 species. Fifteen genera have more than 100 species each; the largest include *Euphorbia* (over 1,600 species), *Croton* (700), *Phyllanthus* (500), and *Acalypha* (430). One of the three flowering-plant groups that have the interesting cactus-like growth form and that inhabit very dry places is a member of this order. Many plants of the order are useful to man. The complex inflorescence (flower cluster) and the variation found throughout the order in the female flower are botanical features of special interest.

GENERAL FEATURES

Size range and diversity of structure. The Euphorbiales order includes plants having many diverse growth forms, including prostrate and erect herbs, shrubs, and trees. Many members are broad-leaved. The shieldlike leaves of *Macaranga gigantea* are among the largest simple leaves (60–100 centimetres [about one yard] across) in the plant kingdom. Many different tree shapes occur in the order. In the genus *Elateriospermum*, the monopodial (or single, centrally stemmed) state is found with a conical crown, the trunk reaching the vertex and the

leaves spirally set on twigs. In the genus *Emblica*, the twigs appear as doubly pinnate leaves (*i.e.*, they are compound, not from a central point, like a fan, but from either side of an axis, like a feather; in this case the subdivision is repeated twice), and the crown is inclined to be flat-topped. Intermediates between these two conditions are frequent. Some species of the genus *Euphorbia* growing in the desert regions of Africa and the Canary Islands closely resemble cactus plants in having fleshy stems of different shapes armed with spines. A number of Australian species are suffrutescent—that is, woody and living over from year to year in the lower parts of the stem just above ground level—and they have small, ramified stems and narrow leaves similar to those of the Ericaceae family (order Ericales), a growth form known, therefore, as the ericoid habit. Several members possess flattened stems or branches known as phylloclades that function as leaves (*Phyllanthus*). A few are lianas. In the genus *Picrodendron*, the bark is peculiar and peels off in long strips. Several kinds of hairs are present in different members of the order—stinging hairs in the genus *Tragia*, glandular in *Acidodendron* and *Macaranga*, nonglandular in *Mabea* and *Phyllanthus*, and numerous others.

Resemblance of some euphorbias to cactus plants

Distribution and abundance. Euphorbiales species are found chiefly in tropical regions but are spread over the whole earth with the exception of the Arctic and cold Alpine zones. Two major centres of distribution are tropical America and Africa. Few species have a very wide range, and some genera have very restricted distributions, such as *Simmondsia* (California and neighbourhood), *Picrodendron* (West Indies), and *Panda* (west tropical Africa). The most widely ranging genus is *Euphorbia*.

Utility. Several species of the order have abundant milky juice (latex) and attendant complex secretory tissue; in certain cases, the juice is economically important. The genera *Mabea*, *Manihot*, and *Sapium* are the source of caoutchouc (a form of rubber), and the highest grade of rubber is obtained from *Hevea braziliensis*. In some cases, the latex is medicinal, but in many species it is poisonous, as in the manchineel (*Hippomane mancinella*) of South America. *Manihot esculenta*—manioc, or cassava—the source of tapioca, is one of the most important tropical food plants, its large tuberous roots being very rich in starch. Tung oil and wood oil are important commercial oils obtained from the seeds of *Aleurites fordii* and *A. cordata*, respectively. The solid oil of *Sapium sebiferum* is used for making candles in China, and it is also considered to be a good substitute for lard. Castor oil is extracted from the castor bean, *Ricinus communis*. *Croton tiglium* is the source of croton oil, and resin is obtained from species of *Euphorbia* and *Croton*. The dried resin-containing latex of *Euphorbia resinifera* of Morocco is a well-known commercial drug, gum euphorbium. The fruits of a few species, such as *Cicca*, *Beccaurea*, *Antidesma*, and others, are edible, as are the seeds of the candlenut, *Aleurites*. Fruit of *Emblica officinalis* is very rich in vitamin C and is used in several medicines. Excellent jams and wines are prepared in the Philippines from the fruits of *Antidesma bunius*. Timber of some species is valuable, such as boxwood (*Buxus*). Candelilla wax is obtained from *Euphorbia antisyphilitica* and jojoba wax from *Simmondsia chinensis*. Species of *Tragia*, *Mercurialis*, *Euphorbia*, *Phyllanthus*, and *Croton* are medicinal. Many are toxic to cattle and sheep; notably, the leafy spurge (*Euphorbia esula*), sun spurge (*E. helioscopia*), and eyebane (*E. maculata*). The African plant *Dichapetalum toxicarium* contains poisonous levels of fluorine (as fluoro-oleic-acid). This plant accumulates and concentrates fluorine from the soil. Species of *Plukenetia* are used as potherbs in the East Indies, *Euphorbia* in Africa, and *Claoxylon* in Malaysia. Many are ornamental plants: the poinsettia (*Euphorbia pulcherrima*), the crown of thorns (*E. milii*), the scarlet plume (*E. fulgens*), the copper leaf and red hot cattail (*Acalypha*), crotons (*Codiaeum*), and others. Trees such as *Daphniphyllum macropodum* are grown for their large and handsome foliage.

Edible fruits

NATURAL HISTORY

Life cycle. Some species are annuals of the wet season (*Micrococca*, *Sebastiania*, and others). Some, such as *Macaranga*, are rapidly growing, short-lived trees. Most members of the order are deciduous, dropping their leaves during seasons of unfavourable growth. Evergreen plants occur in the genera *Daphniphyllum*, *Buxus*, and others. Some, such as *Euphorbia neriifolia*, have leaves only during the wet season. In many other species of *Euphorbia*, the leaves are dropped early, and the stems not only perform the functions of the leaves but also act as water-storage organs. Some species begin flowering early in life, but others bloom only after five or six years.

Ecology. When certain species of *Euphorbia* that have a cactus-like growth habit are encountered in dry habitats, it is very difficult to determine from the external appearance whether the plant is actually a *Euphorbia* or a cactus unless it is in flower. The presence of latex, however, distinguishes the spurge. It is a matter of great interest to find how similar conditions of life have brought forth, in three different and widely separated families, such a close similarity of habit (*i.e.*, the cactoid growth form) as is met with in the families Euphorbiaceae (genus *Euphorbia*), Asclepiadaceae (genus *Stapelia*), and Cactaceae (numerous genera).

Many species prefer open land, but some members form an integral part of the tropical and subtropical forests of the world. Arborescent (treelike) cactoid species of *Euphorbia* are common in savannas. Plants such as *Bischofia javanica*, when growing on seashores, possess a peculiar oblique, condensed, one-sided habit (wind forms) and always show a decreased leaf size owing to the desiccation of buds. *Excoecaria parviflora* and *E. agallocha* prefer tidal marshes and saline flats. Some species are inhabited by ants, as in *Endospermum moluccanum*, in which the ants bite their way into stem sections, remove and eat or carry away the pith, and live inside the cavity thus formed. Other ant plants of interest are *Drypetes*, *Macaranga*, and *Agrostistachys*. Many species of the order are botanical nomads (widely distributed moving weeds having several methods of dispersal)—*e.g.*, *Croton bonplandianum*.

Stems inhabited by ants

FORM AND FUNCTION

Vegetative characters. Leaves of the Euphorbiales order are mostly alternate but are sometimes opposite; they are extremely variable in shape, ranging from simple, in the genera *Euphorbia*, *Buxus*, *Daphniphyllum*, and others, to deeply cut (palmate), in *Ricinus* and *Manihot*, trifoliate, in *Picrodendron*, and palmately compound, in *Hevea*. Stipules (small, leaflike appendages at the base of leafstalks) are absent in the families Buxaceae and Daphniphyllaceae but are usually present in most of the other families. The stipules may be minute and persistent, as in *Picrodendron*; or deciduous (falling); or they may be represented by branched, hairlike bodies (*Jatropha*), glands, or spines (*Euphorbia*). Extrafloral nectaries are present in several members of the Euphorbiaceae family.

Floral characters. Flowers in the order are mostly unisexual, except in the family Dichapetalaceae, in which they are predominantly bisexual. Bisexual flowers are also found in the family Buxaceae. When unisexual, male and female flowers may be borne on the same plant (bisexual plants, the monoecious condition), as in *Hevea*, or on different plants (unisexual plants, the dioecious condition), as in *Daphniphyllum* and *Panda*. The arrangement of the flowers (the inflorescence) shows great variation, and almost all types occur—heads, spikes, racemes, clusters, catkins, panicles, and others. In mixed types, often the first branching is racemose (*i.e.*, having a main axis with flowers borne on short secondary branches, the lower flowers maturing and blooming first) and all subsequent ones cymose (*i.e.*, similar to racemose, but with the terminal flowers blooming first). In bisexual plants seven types of flower arrangement can be seen: (1) male and female flowers in the same cluster without

any order (*Cicca, Bridelia*); (2) one female flower to each cluster of males (*Euphorbia*); (3) male flower clusters toward the base of the leafy twigs, female toward the ends (*Breynia, Emblica*); (4) male flowers at the base of the inflorescence with female in the upper parts (*Ricinus*); (5) female flowers at the base of the inflorescence, male in the upper parts (*Croton, Sapium*); (6) female flowers singly at the ends of the main branches of the inflorescence, surrounded by the males on the smaller side branches (*Aleurites, Hevea, Jatropha*); and (7) female flowers on one raceme, male on another, the female raceme usually paired with a male (*Codiaeum*). The partial inflorescences of *Dalechampia* and *Euphorbia* are so condensed as to give the appearance of single flowers. In *Dalechampia* the inflorescence is very complex. In the centre is a cushion of nine to 14 male flowers subtended by four small bracts. Anterior to this is a cyme of three female flowers in the axil (angle) of a larger bract. All these are enclosed in two large pink or white outer bracts. In *Euphorbia*, the partial inflorescence is a cyathium, which is a flowering group containing a central terminal female flower consisting of only the gynoecium (*i.e.*, the female parts or pistil, consisting of ovary, style, and stigma; there are no petals or other flower parts) and four or five bracts below that unite to form a cup-shaped involucre (whorl) resembling a calyx. Each bract subtends a small cyme of male flowers, the oldest of which is arranged near the centre. Each such cyme has a distinctive joint halfway up the stalk. Between the segments of the involucre are large oval or crescent-shaped glands, often brightly coloured, forming petallike structures. The cymose umbels (flower clusters radiating from a common point reminiscent of an umbrella and maturing from the centre outward) in some species of the Dichapetalaceae family are fused to the leaves (epiphyllous).

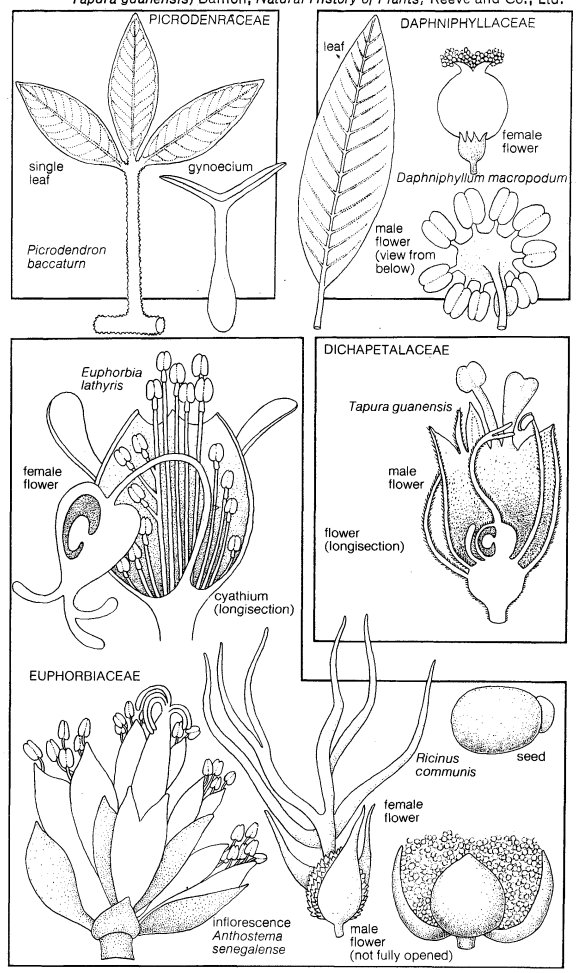
Complex flower clusters

The form of the flowers also shows variation, and the shape and size of flowers of different sex may differ widely, as in *Hevea*. Regular (radially symmetrical) hypogynous (*i.e.*, with a superior ovary—the other flower parts attach at the base of the ovary) flowers occur in almost all the species, except in some of the members of the family Dichapetalaceae, in which the ovary may be inferior. The perianth (the collection of petals and sepals) is mostly in fives or fours, sometimes distinguished into sepals and petals; more often, there is only one whorl of tepals, and, frequently, the flowers are naked, as in several members of the Euphorbiaceae family. When a corolla (the collection of petals) is present, the stamens are usually as many or twice as many as the petals. In naked flowers, the number of stamens is reduced to one (*e.g.*, in the family Euphorbiaceae), or several stamens may be loosely clustered, forming a globose head, as in the genus *Picrodendron*. In *Ricinus* the stamens are branched. In *Phyllanthus cyclanthera* the stamens are united by means of a ringlike common anther (synandrium). Both extrorse and introrse (*i.e.*, opening away from or toward the flower axis, respectively) anthers are present in the group. In the male flowers of several species, a rudimentary ovary is sometimes present. The ovary of female flowers is generally two- to five-chambered, the three-chambered condition being most frequent. Each chamber of the ovary contains one or two pendulous ovules that are usually anatropous (*i.e.*, the body of the ovule is bent back along its supporting stalk) with a ventral raphe (ridge). Placentation (ovule attachment) is mostly axile (along a vertical central axis of the ovary). There are two coats (integuments) for each ovule, and both or either one of them may take part in the formation of the micropyle (a small opening in the ovule). Both of them persist in the seeds of many genera.

Pollination

Cross-pollination is the usual pollination method in the Euphorbiales order. In *Mercurialis* and other genera with inconspicuous flowers, pollination is by wind. Insect pollination takes place in almost all the other members. The flowers being small, insects are attracted to them by the large, brightly coloured bracts, as in *Dalechampia* and *Euphorbia*, or by coloured calyx, as in *Manihot*. Honey glands, present in several members, as on the involucre of the cyathium of *Euphorbia*, are frequently an attraction. Many species are sweet-scented of honey or citronella or have a fetid-wax smell, as in *Elaeteriospermum*, by which means they also attract small flies, beetles, honeybees, wasps, and plant bugs, which pollinate them.

Drawing by M. Moran based on (*Euphorbia lathyris, Anthostema senegalense, Tapura guanensis*) Baillon, *Natural History of Plants;* Reeve and Co., Ltd.



Vegetative and floral structures from four of the families of the order Euphorbiales.

Fruits and seeds. The fruit is a capsule in most members of the families Euphorbiaceae and Buxaceae. They are sometimes prickly, as in *Ricinus*. The fruit may be a drupe (*i.e.*, a fruit with a single-seeded stony pit) or a berry. The capsular fruit often splits into three parts (cocci) from the central column, and at the same time each coccus opens ventrally, letting the seeds escape. Some fruits, such as those of sandbox tree and rubber tree, split with explosive violence, hurling the seeds away. The seeds of some are distributed by animals, either because they have pulpy edible coats or because the fruits themselves are berries. The seeds of several members have a caplike outgrowth called a caruncle. This spongy structure attracts ants, and the seeds carried away by them effect dispersal. The seeds, with the exception of those of the families Picrodendraceae and Dichapetalaceae, contain abundant endosperm (a starchy nutrient tissue for the developing embryo), generally rich in oil. The embryo is usually very large in all but the family Daphniphyllaceae. The cotyledons (seed leaves) are generally flat and emerge above the surface of the ground at the time of germination (epigeal germination) and turn green.

EVOLUTION

The Euphorbiales order appears to be evolutionarily highly specialized, as is evidenced by the very reduced

flowers in many members. It has been presumed that this group evolved from some ancient group intermediate between the family Flacourtiaceae (of order Violales) and the Malvales order. The nonhomogeneity of the order does not support the view that it is monophyletic, however. It is thought that the Euphorbiales order is not a natural assemblage derived from the same ancestors in the distant past but is a synthetic group made up of derivatives from several ancestors brought together merely because of a superficial resemblance and want of knowledge of the true evolution of the groups.

Fossil record

This order does not seem to have fossil representatives below the Upper Cretaceous Period (about 65,000,000 years ago), and such fossils are also few, although such names as *Euphorbiophyllum*, *Crotonophyllum*, and others are occasionally encountered in literature as having been obtained from Upper Cretaceous and Eocene (about 50,000,000 years ago). Tertiary (2,500,000 to 65,000,000 years ago) deposits of *Euphorbia*, *Acalypha*, *Mallotus*, and others are known from various parts of the world.

CLASSIFICATION

Distinguishing taxonomic features. It is impossible to give a set of characters by which the Euphorbiales order can be distinguished absolutely from all other dicots, for there are many exceptions even among the common members. Nevertheless, combinations of certain features are used by taxonomists to set limits to the order and to determine relationships within the groups. Such features include habit, presence or absence of latex, nature of indumentum (hairs on the surface of leaves, stems, and other organs), nature of leaves, presence or absence of stipules, sexuality, nature and arrangement of the flowers, nature and position of the ovary, presence or absence of endosperm, nature of fruit, features of embryo, and others.

Annotated classification. The arrangement of the families presented here is a recent one of wide acceptance.

ORDER EUPHORBIALES

Trees, shrubs, or rarely annual herbs; with simple, stellate, or scaly hairs; leaves simple or rarely compound; stipules mostly present; flowers hypogynous (superior ovary) or very rarely tending to be epigynous (inferior ovary), unisexual or very rarely bisexual, radially symmetric or very rarely tending to be bilaterally symmetric; tepals overlapping or not overlapping in the bud; petals rarely present; stamens numerous to solitary, free or monadelphous (joined by their filaments); ovary with axile placentation; ovules mostly anatropous; seeds with copious endosperm or rarely without it. Seven families, about 315 genera, and 8,000 species.

Family Buxaceae (boxwood family)
Evergreen shrubs or trees with simple leathery leaves; unisexual naked flowers in heads or spikes; stamens 4–6; ovary superior 3-locular; ovules 1–2, pendulous. Five genera and about 60 species distributed in east Asia and North America.

Family Simmondsiaceae (jojoba-wax family)
Evergreen shrubs with opposite simple leaves; flowers without petals, dioecious (*i.e.*, male and female flowers are on different plants), male clustered, female solitary, stamens many; ovary 3-celled with one ovule in each cell. One genus with a single species, *Simmondsia chinensis*, distributed in southwestern North America.

Family Daphniphyllaceae
Leaves simple, crowded, alternate; flowers dioecious in racemes, without petals; sepals of male flowers 3–8; stamens, 6–12; ovary superior, 2-locular with two ovules in each cell. One genus, *Daphniphyllum*, with about 35 species distributed in east Asia and Malaysia.

Family Euphorbiaceae (spurge family)
Herbs, shrubs, or trees; some with milky latex; leaves mostly simple; flowers mostly monoecious (both sexes on one plant); petals rarely present; stamens one to many, free or joined together; ovary superior, 3-locular; seeds often carunculate. About 300 genera and about 7,650 species distributed in tropical and temperate regions.

Family Dichapetalaceae
Small trees or shrubs, sometimes climbing; leaves simple; flowers mostly bisexual; sepals 5, petals 2-parted, free or united with stamens into a tube; ovary superior to inferior, 2- to 3-locular; ovules 2 in each cell. Four genera and 250 species distributed in tropical Africa and America.

Family Pandaceae
Small trees; leaves simple; flowers dioecious; petals present; stamens 10; ovary 3- to 4-locular; ovule solitary in each cell. One genus with a single species, *Panda oleosa*, distributed in west tropical Africa.

Family Picrodendraceae
Trees; leaves trifoliate; female flowers solitary, male in catkins; sepals and petals absent; stamens many, clustered, forming a globose head; ovary superior, 2-locular; ovules two in each cell; fruit 1-seeded. One genus with a single species, *Picrodendron baccaturn*, distributed in the West Indies.

Critical appraisal. There is much disagreement among taxonomists regarding the systematic position of the families included in the Euphorbiales order. Some authorities retain only the family Euphorbiaceae in the order, the remaining families being variously placed—Dichapetalaceae in the order Rosales, Buxaceae (including Simmondsiaceae) and Daphniphyllaceae in Hamamelidales, Pandaceae in Celastrales, and Picrodendraceae in Juglandales. Other authorities treat Euphorbiaceae, Daphniphyllaceae, and Dichapetalaceae as families in the Geraniales order; Buxaceae in Sapindales; and Pandaceae in an independant order, Pandales. Picrodendraceae as a family is not recognized in these systems, and the genus *Picrodendron* is then included in the family Simaroubaceae. Similarly, *Simmondsia* is included in Buxaceae. Three genera of the family Euphorbiaceae —*Micrantheum*, *Pseudanthus*, and *Stachystemon*—have been united by some workers into a special family, Micrantheaceae, but this separation is not widely recognized. Some taxonomists have even elevated many individual genera of the Euphorbiaceae family to the rank of family—*e.g.*, the families Androstachydaceae (from the genus *Androstachys*), Stilaginaceae (*Antidesma*), Bischofiaceae (*Bischofia*), Hymenocardiaceae (*Hymenocardia*), Uapacaceae (*Uapaca*), Peraceae (*Pera*), and others. All these differences of opinion among competent taxonomists probably reflect the polyphyletic (origin from various unrelated ancestors) nature of the order.

Alternative classification schemes

BIBLIOGRAPHY. F. PAX, "Euphorbiaceae," in A. ENGLER, *Das Pflanzenreich*, vol. 42, 44, 47, 52, 57, and 63 (1910–14), a long series of papers, some exceeding 300 pages in length, making an authoritative monographic treatment of the genera of the order; F. PAX and K. HOFFMAN, *ibid.*, vol. 68 and 85 (1919–24), continuing the monographic treatment of Euphorbiales genera; A. ENGLER and K. PRANTL, *Die naturlichen Pflanzenfamilien*, 2nd ed., vol. 19a–19c, 20a (1931), a descriptive taxonomic treatment of flowering plants, including those of the Euphorbiales order, written in German; J. HUTCHINSON, *Families of Flowering Plants*, vol. 1, *Dicotyledons*, 2nd ed. (1959), a good source of general information; A. TAKHTAJAN, *Flowering Plants: Origin and Dispersal* (1969; orig. pub. in Russian, 2nd ed., 1961), treatments of flowering plant classification and phylogeny; J.C. WILLIS, *A Dictionary of the Flowering Plants and Ferns*, 7th ed. rev. by H.K. AIRY SHAW (1966), a compilation of all plant genera, families, and higher divisions, including information on taxonomic placement, members of species, and other botanically interesting matters; older editions also include much information on common names and economic products.

(N.C.N.)

Euripides

An 18th-century English man of letters, Horace Walpole, remarked: "The world is a comedy to those that think, a tragedy to those that feel." Euripides, the youngest of ancient Athens' three greatest tragic poets, following Aeschylus and Sophocles, both thought and felt deeply and found little to smile at. His plays were rich in ethical and social comments and have provided a rich ore for later writers and speakers who have found such comments easy to detach from their dramatic context and enshrine in moral essays, anthologies, or even Christian sermons.

It is possible to reconstruct only the sketchiest biography of this thoughtful man who was the associate of philosophers, the owner of an exceptionally big library, and a stranger to hilarity.

Sources for biography. Among the Euripidean manuscripts, the earliest of which is of the 11th century AD, there is a *Life of Euripides*. Put together sometime in late

Euripides, marble herm copied from a Greek original, c. 340–330 BC. In the Museo Archeologico Nazionale, Naples.
By courtesy of the Soprintendenza alle Antichita della Campania, Naples

antiquity, it is an entirely uncritical assemblage of material selected from previous writers. The earliest and most reliable of these, named Philochorus, wrote a *Life of Euripides* in the first half of the 3rd century BC, well over 100 years after the poet's death. He would have had access to certain official records of the dramatic festivals; apart from that, there are only oral traditions to fall back on. For the most part, the extant *Life* consists of literary gossip, some of which looks like more or less intelligent inferences drawn from the plays themselves; much of it is nothing but jokes made against Euripides by contemporary comic poets, jokes now mistaken for sober history. One well-known biographical "fact," for example, is that Euripides' mother was a greengrocer who sold herbs in the market-place. The great comedist Aristophanes said so in play after play and it was always good for a laugh. But it may not be true. Philochorus, in fact, said that Euripides was of a very good family; Theophrastus (a pupil of Aristotle) recorded that as a lad Euripides was given an honorable part in two religious ceremonies, which would indicate that his mother was not a greengrocer; and Aristotle himself records in the *Rhetoric* that in later life Euripides was involved in vexatious litigation of a technical kind that suggests that he was a man of wealth, or at least could be represented as such—and money was not made in ancient Athens by writing plays any more than by selling greengroceries.

Uncertain biographical accounts

Life. Euripides' mother's name was Cleito; his father's, Mnesarchus or Mnesarchides. Tradition said that he was born in 480 BC, on the date of the Battle of Salamis (a decisive Athenian sea victory against Persia). The fiction is obvious. Since the great tragedian Aeschylus fought in the battle and the other great Athenian tragedian Sophocles led a chorus of boys which celebrated the victory, it would be appropriate to select that glorious day as the time of Euripides' birth. A 3rd-century BC inscriptional record of festival dates reckoned that he was born in 484, which may well be correct. The only other biographical dates that are certain (apart from the dates of some of the plays) are that he first received the honour of being chosen to compete in the dramatic festival in 455; won his first victory (with plays now unknown) in 441; left Athens for good in 408, accepting a very honorific invitation from Archelaus, king of Macedonia; and died in Macedonia in 406. According to the *Life*, traditionally, he died in suitably dramatic circumstances: torn to pieces either by dogs or by women. More worthy of belief is the statement that in 406 BC, when Sophocles next led a chorus into the theatre, he and the chorus wore mourning for Euripides, and in 405 Aristophanes produced *The Frogs*, a play based on the idea that Athens no longer had a great tragic poet. History was to show that Aristophanes was not far wrong.

Known biographical dates

Several anecdotes attest to how greatly honoured Euripides was after his death. The *Life* cites an epitaph composed (as it says) either by Thucydides the historian, or by Timotheus, a musician, presumably for a cenotaph built for Euripides in Athens:

> His bones are laid in Macedon, where he
> Ended his life. His tomb? The whole of Hellas.
> Athens his motherland. His muse gave joy
> To many: many give to him their praise.

In what the sources tell of Euripides' personal history, there is one gap that can hardly be accidental. It is known that Aeschylus was proud of having fought at Marathon (a battle in which Athens repulsed Persia in 490 BC) and Salamis. Sophocles, on the other hand, had once or possibly twice been elected one of the 10 generals of Athens and was later appointed to a commission charged with revising the Athenian constitution. Not a word, however, has survived of any of Euripides' public activity, except that an ancient commentator on Aristotles' *Rhetoric* says that he once served on a diplomatic mission to Syracuse in Sicily (some modern scholars have suspected that his name might be a mistake for that of Hyperides, an Athenian orator).

Private life

Of Euripides' private life, very little can be said. Later tradition invented for him a spectacularly disastrous married life. This, no doubt, was based on jokes made by comic poets and was supported perhaps by the naïve thought that the evil women who appear so often in his plays must have been drawn from unhappy personal experiences. It is known that he married a lady called Melito and had three sons. One of these was something of a poet and produced the *Bacchae* after his father's death and may have completed the unfinished play *Iphigeneia at Aulis*.

Dramatic career. The ancients knew of 92 plays composed by Euripides, of which 67 are known in the 20th century, if only by name. Nineteen plays are extant, if one that is disputed—*Rhesus*—is included. The figure 92 may seem a very large number, but it is perfectly credible. A poet allowed to give a performance at each year's dramatic festival had to produce a tetralogy: three tragedies and a lighthearted production called a satyr play. But Greek plays were short, and a normal tetralogy would not exceed 6,000 verses in length (Shakespeare's *Hamlet* runs to about 8,000). Ninety-two plays would correspond with 23 tetralogies. Spread over an active career of nearly 50 years, they would average one tetralogy every two years—implying an assiduous but not a frantic professional life. Shakespeare's average yearly output, by comparison, was greater. At only four festivals was Euripides awarded the first prize—the fourth, posthumously, for the tetralogy that included the *Bacchae* and *Iphigeneia at Aulis*. As Sophocles won either 18 or 24 victories, it is clear that Euripides was comparatively unsuccessful. More to the point, perhaps, is that on more than 20 occasions Euripides was chosen, out of all contestants, to be one of the three laureates of the year. Furthermore, the regularity with which Aristophanes parodied him is proof enough that his work commanded attention. It is often said that disappointment was one of the reasons for his leaving Athens in his old age. That may be true, though it is purely surmise, but there are other reasons, even more plausible, why an old poet might have left Athens in the 23rd year of the Peloponnesian War.

Euripides is said to have made enemies, to have been prosecuted for blasphemy (like Socrates, a few years later), to have given offence by putting immoral women on the stage, and in other ways to have undermined respectable opinion. In particular, he was accused—and still is, occasionally—of misogyny. Certainly, it is easy to extract from his plays, including those now only partially known, a long list of heroines who are fierce, treacherous, or adulterous, or all three at once. It is significant that

one of the plays in his first tetralogy of 455 BC was *Peliades* (*The Daughters of Pelias*, who, legend related, were tricked by the wicked Medea into dismembering and boiling their father in the foolish belief that he would thereby recover his youth); a few years later he produced *Aegeus*; and in 431 *Medea*. Clearly, he was attracted by the devastating character of Medea, who figured in all three of these works. In the plays there are other such treacherous and destructive women, but misogyny is altogether too simple an explanation. For one thing, in the extant plays there are women who nobly sacrifice themselves for others: *Heraclidae* (*Children* of *Heracles*), *Hecuba*, and *Iphigeneia in Aulis*. What is more, even moderate diligence discloses in the extant plays a list of men who are no more admirable than the fearsome women; none of them indeed so passionate, but most of them more contemptible and some of them more destructive. There are also noble male characters—Theseus (*Suppliants*) and Heracles (*The Mad Heracles*), for example—but not many.

The Euripidean heroine

Also to be noticed as distinctively Euripidean are oppressed heroines who begin by attracting sympathy and finish by being as vindictive as were their oppressors: in *Medea*, for example, the female chorus is at first indignant on Medea's behalf and at the end is even more indignant at what she has done; and in *Hippolytus*, Phaedra is first a loyal wife who abominates her uncontrollable passion for her stepson and will die rather than reveal it and then, when her secret is wormed out of her, writes a lying accusation that will involve others in her own doom. A clue to all this, and to Euripidean tragedy in general, is found in the *Medea*. When the enchantress Medea has destroyed not her husband Jason, leader of the Argonauts, but everyone else who is connected with him, she is seen on the roof, inaccessible, with her dead children in her arms. First the horrified chorus, then Jason, appeal to Heaven and Earth to avenge so inhuman an atrocity. But Heaven and Earth do nothing of the kind: in fact the Sun God (in mythology an ancestor of Medea) sends down a chariot to bear Medea triumphantly to refuge in Athens. Aristotle condemned the ending as "illogical." That it was illogical was Euripides' point.

Euripides' tragic outlook. Euripides' view was emotional, given to despair. Instead of a basic order, hard, unforgiving, but in the main intelligible, he saw a universe much harder for man to live with and much more unpredictable. Chance and disorder and human reason continually thwarted by human unreason and passion and confronted by *theoi* (the powers) who seem indifferent to human reason and morality are the elements of his tragedy, and apparently meaningless suffering is the outcome.

It is often said that Euripides, in portraying "gods" in glaring colours, was attacking the traditional Olympian religion as crude and false. That is probably a misconception. Aeschylus and others had already done so, and Euripides, in this respect, was being traditional rather than a reformer.

The war with Sparta broke out in 431 and continued, year after year, with increasing ferocity and folly. In two of his earlier wartime plays, the *Heracleidae* and *Suppliants*, Euripides represents an ideal Athens undertaking a just war to defend the oppressed; but soon came the *Hecuba* with its twofold picture: first the inhuman and stupid brutality of the conquerors toward Hecuba and the other Trojan captives; then the sickening revenge inflicted by "the mobled (muffled, as with a hood) Queen" herself upon the barbarian who had murdered her son for gold. This series of plays is crowned by the *Trojan Women*, in which, again, the conquerors inflict one outrage after another on the vanquished but were themselves to be destroyed at sea on their voyage home. Like Cassandra (a Trojan princess who foretold the Greek capture of Troy in the legendary Trojan War), Euripides prophesied, but in vain. When he was writing this play his own city was engaged in subduing and destroying an unoffending Aegean island, Melos; immediately afterward Athens despatched a powerful force to conquer Sicily, and hardly a man returned alive. It may not be only out of personal disappointment that Euripides left Athens for Macedonia in 408, when the war had dragged on for another seven years.

Euripides' tragedies are not nicely and logically Aristotelian in structure: Euripides did not feel himself to be living in a nicely logical universe. It follows, too, that certain other of the famous Hellenic virtues are not always to be found in his work. He does not make, for instance, a point of avoiding the sensational, nor of practicing reticence. Euripides' character drawing is rarely rich or subtle; his characters rarely grapple. They debate or they deliver little homilies on matters of philosophic or social interest. The choral odes are apt to be lightweight, prettily decorated, rather monotonous in rhythm, sometimes only tangential to what is going on. It could hardly be otherwise when what is going on is often no more than waiting for the next blow to fall. His power lies in the scarifying situations that he creates and in the total and frightening effect of a play. He is an incomparable knife twister.

The tragicomedies

During the last two decades of his career Euripides began to write tragedies that are not "tragedies" in the modern sense of the word but what might be called romantic drama or tragicomedy because they have happy endings. Extant plays in this style include *Ion*, *Iphigeneia in Tauris*, and *Helen*. Unlike the tragedies, they do not deal with anything of intrinsic importance. If gods appear in them they are not to be taken seriously. Such plays might be described as pure "theatre," of the highest order: in them, turning his back for the time being on the real world that he found so unrelievedly tragic, Euripides reveals himself as one of the greatest virtuosos of the stage.

Plays of the tragicomedy type seem to point forward to the drama of the 4th century BC. It is certain that the fame and popularity of Euripides soon began to eclipse that of Aeschylus and Sophocles. The age of Classical Greece was drawing to a close, to be followed by the cosmopolitan Hellenistic period. One can readily understand that in the new age the essentially political and "religious" tragedy of Aeschylus and Sophocles made less appeal than that of Euripides, with its more obvious emotional, even sensational, effect. It may well be that even their Greek became progressively more difficult to follow than his. Also in the extant "fragments"—that is, citations found in later writers—those of Euripides are more numerous than those of Aeschylus and Sophocles put together.

MAJOR WORKS

Alcestis (438), pro-satyric; *Medea* (431); *Hippolytus* (428); *Heracleidae*, *Andromache*, *Suppliants*, and *Hecuba* (all four before 423); *Electra; Trojan Women* (415); *Helen* (412); *Phoenician Women* (411–409); *Heracles*, *Iphigeneia in Tauris*, *Ion*, and *Orestes* (all four 408); *Iphigeneia in Aulis* and *Bacchae* (both posthumous); *Cyclops*, a satyr play; *Rhesus* (disputable). Substantial papyrus fragments have been discovered of the *Hypsipyle* (411–409), ed. by G.W. Bond (1963).

BIBLIOGRAPHY

Texts: *Euripides*, 3 vol., "Oxford Classical Texts"; Collection des Universités de France: 6 vol. with French translation.

Translations: Convenient translations may be found in *The Complete Greek Tragedy*, 5 vol. (1955–59); and in WHITNEY OATES and EUGENE O'NEILL, *The Complete Greek Drama*, 2 vol. (1938).

Critical works: G.M.A. GRUBE, *The Drama of Euripides* (1941, reprinted 1961); PAUL DECHARME, *Euripide et l'espirit de son théâtre* (1893; Eng. trans., *Euripides and the Spirit of His Dramas*, 1906, reprinted 1968); D.J. CONACHER, *Euripidean Drama: Myth, Theme and Structure* (1967), including a useful bibliography of recent work; GILBERT MURRAY, *Euripides and His Age* (1965); T.B.L. WEBSTER, *The Tragedies of Euripides* (1967), which discusses also the plays known only from fragments; GUNTHER ZUNTZ, *The Political Plays of Euripides* (1963); R.P. WINNINGTON-INGRAM, *Euripides and Dionysus* (1948); WILLIAM RITCHIE, *The Authenticity of the Rhesus of Euripides* (1964); F.L. LUCAS, *Euripides and His Influence* (1963). See also chapters in ALBIN LESKY, *Die griechische Tragödie* (1938; Eng. trans., *Greek Tragedy*, 1965); H.D.F. KITTO, *Greek Tragedy* (1961); and D.W. LUCAS, *The Greek Tragic Poets* (1959).

(H.D.F.K.)

Europe

Among the continents, Europe is an anomaly. Larger only than Australasia, it is a small appendage of the great land mass that it shares with an Asia more than four times its size. Yet the peninsular and insular extremity of Eurasia, thrusting toward the North Atlantic Ocean, provides—thanks to its latitude and its physical geography—a relatively genial human habitat, and the long processes of human history came to mark off the region as the home of a distinctive civilization. In spite of its internal diversity, Europe has thus functioned, from the time it first emerged in the human consciousness, as a world apart, concentrating—to borrow a phrase from Marlowe—"infinite riches in a little room."

All the continents are conceptual constructs, but only Europe was not first perceived and named by outsiders. "Europa," as the more learned of the ancient Greeks first conceived it, stood in sharp contrast to both Asia and Libya, the then known northern part of Africa. Literally, "Europa" is now thought to have meant "mainland," rather than the earlier interpretation, "sunset." It appears to have suggested itself to the Greeks, in their maritime world, as an appropriate designation of the broadening, extensive northerly lands that lay beyond, lands with characteristics but vaguely known, yet clearly different from those inherent in the concepts of Asia and Libya, both of which, relatively prosperous and civilized, were associated closely with the culture of the Greeks and their predecessors. Traders and travellers reported that Europe possessed distinctive physical units, with mountain systems and lowland river basins much larger than those familiar to inhabitants of the Mediterranean region. It was also clear that a succession of climates, markedly different from those of the Mediterranean borderlands, were to be experienced as Europe was penetrated from the south. The spacious eastern steppe and, to the west and north, primeval forests as yet only marginally touched by human occupancy further underlined environmental contrasts. Europe was culturally backward and scantily settled. It was a barbarian—that is, a non-Greek—world, its inhabitants making "bar-bar" noises in unintelligible tongues.

Origins of the "Europa" concept

The Roman Empire, at its greatest extent in the 2nd century AD, revealed, and imprinted its culture on, much of the face of the continent, while trading relations beyond its frontiers also drew the remoter regions into its sphere. Yet it was not until the 19th and 20th centuries that modern science was able to draw with some precision the geological and geographical lineaments of the European continent, the peoples of which had meanwhile achieved domination over—and set in motion vast countervailing movements among—the inhabitants of much of the rest of the globe.

As to the territorial limits of Europe, while these seem clear on its three seaward flanks, they have been uncertain and hence much debated on the east, where the continent merges, without sundering physical limits, with parts of western Asia. Even to the north and west, many island groups—Svalbard (Spitsbergen), the British Isles, the Faeroes, Iceland, and the Madeira and Canary islands—that are European by culture are included in the continent, although Greenland is conventionally allocated to North America. Further, the Mediterranean coastlands of northern Africa and southwest Asia also exhibit some European physical and cultural affinities, and Turkey and Cyprus, while geologically Asian, possess elements of European culture and may, perhaps, be regarded as parts of Europe. Eastward limits, now adopted by European (including Soviet) geographers, assign the Caucasus to Asia and are taken to run southward along the eastern foot of the Urals, and then across the Mugodzhar Hills (Mugodzharskiye Gory), along the Emba River (Vaike Emajogi), and along the northern shore of the Caspian Sea. West of the Caspian, the European limit follows the Kumo-Manych Depression (Kumo-Manychskaya Vpadina) and the Kerch Strait (Kerchensky Proliv) to the Black Sea.

The 4,100,000 square miles (10,600,000 square kilometres) occupied by Europe reveal no simple unity of geological structure, landforms, relief, or climate. Rocks of all geological periods are exposed, and the operation of geological forces during an immense succession of eras has contributed to the molding of the landscapes of mountain, plateau, and lowlands, and bequeathed a variety of mineral reserves. Glaciation, too, has left its mark over wide areas, and the processes of erosion and deposition have created a highly variegated and compartmentalized countryside. Climatically, Europe benefits by having only a small proportion of its surface either too cold or too hot and dry for effective settlement and use. Regional climatic contrasts nevertheless exist: oceanic, Mediterranean, and continental types occur widely, as do gradations from one to the other. Associated vegetation and soil forms also show continual variety, but little is left of the dominant woodland that clothed the continent when *Homo sapiens* first appeared.

All in all, Europe enjoys a considerable and long-exploited resource base of soil, forest, sea, and minerals, notably coal, but manpower, considerable numerically, as well as technically highly qualified, is increasingly its principal resource. The continent, by the 1970s, contained a slightly shrinking fifth portion of mankind, but this represents a population of high skill and initiative. Europe thus supports high densities of population, markedly concentrated in industrialized regions. In manufacture, commerce, and agriculture, it still occupies an eminent, if no longer necessarily predominant, position, and, as agriculture increasingly rationalizes its structure, city life is, everywhere, becoming the norm.

Europe and the world

Europe is pre-eminently the homeland of white peoples. Its early and continuing economic achievements, evidenced by a high standard of living, and its successes in science, technology, and the arts spring from the vigour of its peoples in developing a high civilization, the roots of which lie in ancient Greece and Rome, the Byzantine Empire, and Palestine. Whatever its indebtedness, Europe has always shown its own powers of creativity and leadership: although wracked and exhausted by continued internal conflict, it has nevertheless advanced sufficiently to leave as its heritage the exploration, colonization, and development of other races and regions of the globe, if not always to the benefit of the other races and regions.

This article covers the overall physical and human geography of the European continent. For information on related topics, see the articles EUROPEAN CULTURE SINCE 1800; EUROPEAN DIPLOMACY AND WARS (C. 1500–1914); WORLD WARS; as well as articles on other major topics in the history of the continent and on the history of individual nations. There are separate articles on the major geographical features of Europe, as well as articles surveying contemporary nations and cities; see also CONTINENTS, DEVELOPMENT OF.

This article is divided into the following sections:

I. Geological history
  - Elements and processes in the making of the continent
  - The territorial emergence of Europe
  - The legacy of successive eras

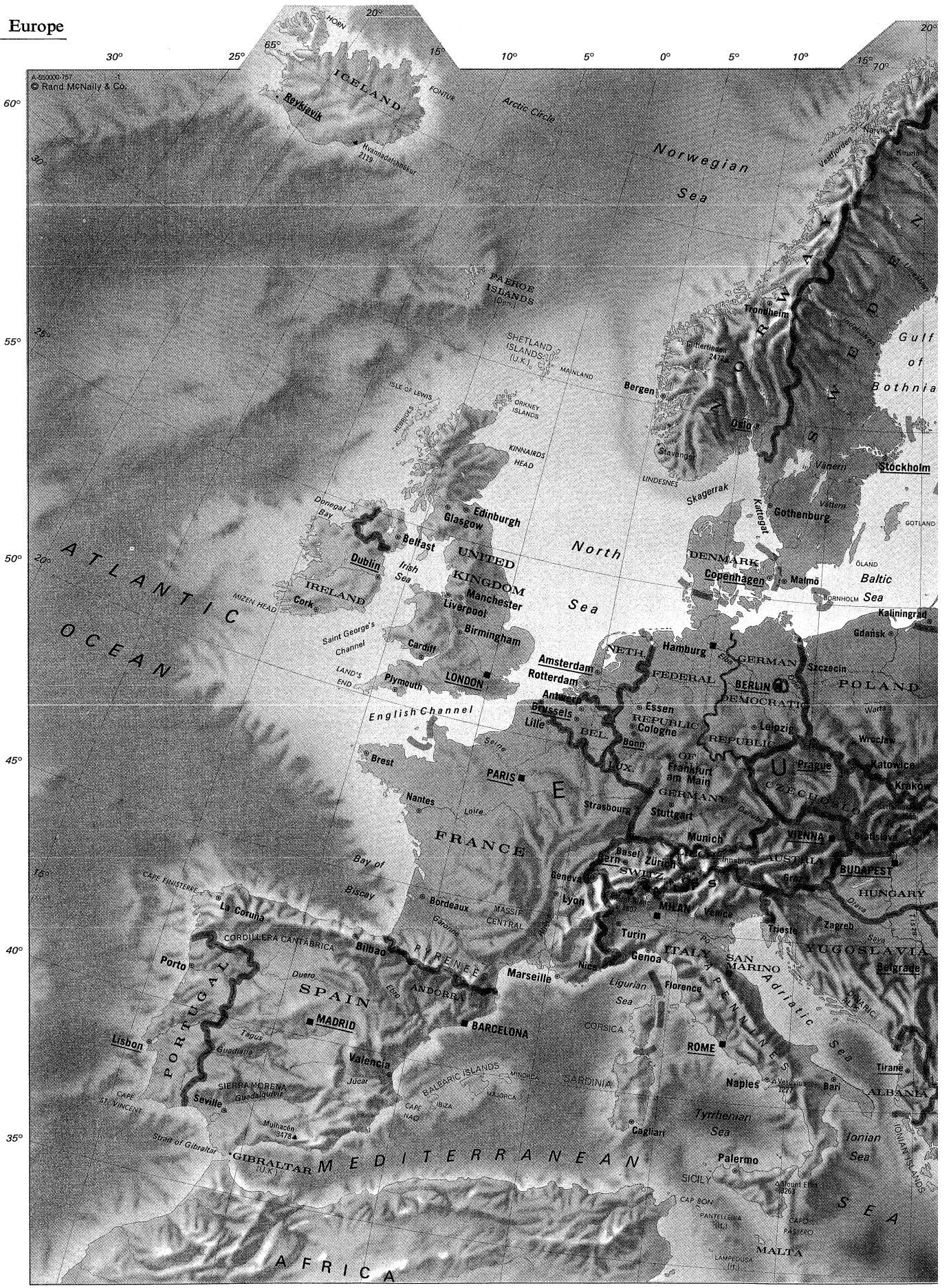
II. Physical geography
  - Relief
  - Climate
  - Drainage
  - Soils

III. Plant and animal life
  - Vegetation
  - Animal life

IV. Natural and human resources
  - Mineral resources
  - Water resources
  - Biological resources
  - Human resources

V. Resource development
  - European industry
  - Power
  - Agriculture and land use

VI. Commerce
  - Transport
  - Internal trade
  - External trade
  - The commercial outlook



EUROPE Size of symbol indicates relative size of town • ◉ ■

Elevations and depths in metres

Barents Sea
CAPE KANIN NOS
Murmansk
KOLA PENINSULA
TIMAN RIDGE
Pechora
URAL MOUNTAINS
White Sea
Arkhangelsk
Onega
Northern Dvina
Kotlas
FINLAND
Lake Onega
Sukhona
NORTHERN
Perm
Kirov
Lake Ladoga
Helsinki
Vologda
LENINGRAD
Gulf of Finland
Tallinn
Rybinsk Res.
Yaroslavl
Gorky Reservoir
Kazan
Ufa
Magnitogorsk
Lake Peipus
Lake Ilmen
Gorky
Volga
Kuybyshev Reservoir
Gulf of Riga
Valdai Hills
Kalinin
MOSCOW
SOVIET UNION
Riga
Western Dvina
Kuybyshev
Orenburg
Jelgava
Nemen
Smolensk
Penza
Oka
PRIVOLGA HILLS
Vilnius
Tula
Saratov
Minsk
Dnepr
CENTRAL RUSSIAN UPLAND
Ural
Gomel
Voronezh
Warsaw
Desna
Kursk
Don
VOLGA
Volgograd Reservoir
Kiev
Kharkov
VOLGOGRAD
CASPIAN DEPRESSION
Lvov
Tsimlyansk Reservoir
Volga
Donetsk
Astrakhan
Dniepropetrovsk
Zaporozhye
Rostov-na-Donu
Kishinyov
Dnepr
Sea of Azov
Caspian Sea
ODESSA
CRIMEA
Krasnodar
Grozny
ROMANIA
Sevastopol
GREATER CAUCASUS
Bucharest
Danube
Black Sea
Varna
BULGARIA
Sofia
ISTANBUL
TURKEY
Sea Of Marmara
Thessaloniki
Aegean Sea
GREECE
Athens
CYCLADES
DODECANESE
RHODES
Iraklion
CRETE
ASIA
25°
30°
35°
40°
45°
50°
55°
60°
65°
70°
75°
80°
70°
65°
60°
55°
50°
45°
40°
35°
0 100 200 300 400 500 600 700 800km
0 100 200 300 400 500mi

VII. Demographic patterns
Population distribution
Extracontinental movements
The outlook for Europe

I. Geological history

The continental framework of Europe is a geologically recent product of earth history, and modern research has outlined the history of the making of the European landmass over a period exceeding 500,000,000 years. This time span starts with the Precambrian, from which the fossils of the first living things are found in the rocks, and includes successive eras known as the Paleozoic (lasting 345,000,000 years), the Mesozoic (lasting 160,000,000 years), and the Cenozoic (which has lasted 65,000,000 years and includes Tertiary and Quaternary periods). Each of these eras is divided into named sequent periods, and the whole of the immense span of the geological record—itself only the latter phase of Earth history—dwarfs that of the human record, which begins towards the end of the Tertiary. In approximations to its present configuration, Europe cannot be traced back before Tertiary times, but some parts of it are made up of rock structures that originated as far back as the Paleozoic, or even the Precambrian, eras.

ELEMENTS AND PROCESSES IN THE MAKING OF THE CONTINENT

Shield, crust, and ocean. The specific molding of the European continental mass resulted from the interplay, over time, of certain basic structural elements and powerful processes, the latter often operating deep within the Earth. Among the elements from which Europe became compounded, special interest attaches to the massive earth blocks, or shields, which form the oldest territorial units of the continent. These consist wholly of the oldest known rocks; the yet older, or primitive, rocks of the earth's crust have disappeared through time. These ancient shield, or platform, rocks of Europe, some of which have stood above the sea throughout much of geological time, were themselves formed up to more than 1,000,000,000 years before the geological record begins. Both major and minor shields can be distinguished: the Fennoscandian (or Baltic) Shield and the Russian Platform are major units, whereas the Tyrrhenian and Aegean shields to the south are only fragmentary. These structures performed two functions: they provided core areas around which younger components of the continent came to be built, and they provided almost immovable obstacles, or forelands, to lateral, horizontal pressures of the Earth's crust.

Crustal layers

These crustal pressures stemmed from the nature of the crust (or lithosphere) itself in and around the region where Europe emerged. The crust is formed of an inner mantle of very great depth, covered by a relatively thin layer of dense rocks, notably basalt, and known as the sima. Above this layer, an even thinner cover of lighter rocks, the sial, underlies the continent but is largely lacking from the surrounding ocean bed; Europe can thus be visualized as a raft of sial rocks, notably granite, buoyed up in the denser sima. A principle of equilibrium, moreover, applied to these crustal processes: the mass of sima, beneath the ocean, and the mass of sial, penetrating deeply from the European land surface until it rests on the sima below, must achieve a state of balance. This operation, isostasy, accounts for continued, small-scale raising or lowering the land surface relative to sea level and has been of vital significance in the molding of Europe.

Another fundamental element is provided by the vast saltwater body, the world ocean, which has always covered much of that part of the crust lacking a sial shell: indeed, at all geological times, sial land areas, in Europe as elsewhere, have been exceptional features. When the volume of water in the ocean surrounding Europe increased, it invaded large land areas. Into these geological seas, later drained, which above all characterized the Mesozoic and Tertiary eras in emergent Europe, masses of sediments and the fossil remains of marine life were deposited, providing an accurate time record.

Endogenetic and exogenetic forces. In addition, both the ocean bed surrounding Europe and the emerged lands comprising the continent have been subjected to strong vertical pressures exerted from within the Earth by way of the very deep mantle enclosing its inner core. These inner forces, summed up by the term endogenetic, are very important to an understanding of the emergence of Europe. They have had various effects; one is that, along tensional fissures, or zones of weakness in the crust, great volumes of magma, or molten rock substance, have been emitted as lava, which has solidified to become the rock (predominantly basalt) known as igneous. Where the magma has not reached the European surface, it has been injected among other rocks to form subterranean sheets, sills, and veins. Such emissions of hot liquid within the crust are believed to account for the great extent of granite there, and indeed the process is known as the granitization of pre-existing rocks. These granitic intrusions, or batholiths (literally: "rocks of the depths") occupy, notably in Finland, areas of considerable size and may become exposed through the removal of rock layers above them. Migmatites, mixed rocks containing some granite, are formed by the same process acting on the margins of granitic bodies and are a common occurrence in Europe. In addition to the sedimentary and igneous rocks of Europe, a final class of rock is known as metamorphic and was formed by such great heat and pressures from within the crust as to change sedimentary and igneous rocks into a new form.

Igneous and metamorphic rocks

In addition to isostasy, noted above, endogenetic forces that have been active in Europe include both epeirogenesis—the continent-making force, exerted vertically through the earth's mantle, which can raise or lower extensive areas—and orogenesis, or the mountain-creating force. This latter process involved not only the ancient shields of Europe but also the equally important structural units known as geosynclines: enormous, elongated, troughlike zones, lying adjacent to the shields, into which sediments—transported waste from the shields—accumulated, over geological time, into great depths, causing the trough itself to sag deeper. The great mountain systems of Europe, with the Alpine System the most recent and pre-eminent example, resulted from powerful and volcanic compressional forces exerted laterally by shield areas against the massive sedimentary content of the geosynclinal basins lying between them, forcing the former troughs upward and folding them into new, elevated mountain systems. This orogenic process took place most markedly in the Paleozoic and Cenozoic eras and involved a dramatic rupturing of parts of the European shields, much volcanicity and metamorphism, and the creation of great masses of basalt and granite. Violent and dramatic as these orogenic processes appear, they took place over many millions of years and were the main force molding the emergent physique of Europe and the important disposition of land and sea. Milder crustal movements, known as warping, also affected the shaping of Europe's surface through their effect on the ancient shields, particularly in the Mesozoic Era, when orogenesis was largely absent.

The operation of exogenetic forces—that is, the endless processes of erosion and sedimentation at the surface, where rocks are exposed to the atmosphere and thus to the effects of climate and weather—provided the finishing touches to the making of the European continent. Over periods of long duration, these forces have worked powerfully on the exposed rocks and landforms of Europe and so have differentiated remarkably the physical geography of the continent. The chemical and mechanical weathering, or wearing away, of rocks was followed by the transportation of rock waste by gravity and the action of glaciers, running water, and winds, eventually depositing the waste offshore (infilling the geosynclines mentioned above), in lakes and inland seas, and on the land surface itself. A final, but most important, exogenetic process active in Europe was associated with the Great Ice Age of the Pleistocene Epoch, which

The role of weathering

lasted for about 2,000,000 years after the close of the Cenozoic Era and formed the initial phase of the Quaternary Period.

Chronological summary. The chronology of the emergence of Europe may be summarized as follows: the first element to be noted is the presence of Precambrian shield areas in the north and east, known, respectively, as Fennoscandia and the Russian Platform. Against these structures, the Caledonian orogeny of the Paleozoic folded and uplifted the Caledonian mountains. The Paleozoic also witnessed the Hercynian orogeny, which created mountain systems in many parts of the present European continent from Iberia and the British Isles in the west, to the Ural Mountains at the eastern margin. Prolonged erosion and deposition from these ancient elevated surfaces continued, while, predominantly during the ensuing Mesozoic and Cenozoic eras, widespread seas, occupying much of the present continent, which had not yet emerged as land, as well as the shield rocks of the Russian Platform, received layer upon layer of sedimentary rocks. The Cenozoic Era brought a new series of orogenic events, known as Alpine, which, in conjunction with later uplifts, bequeathed to Europe the relatively youthful mountain systems that persist in many parts of the continent. Other, less violent, structural changes took place beyond the main regions of mountain-building, and also the whole pattern of land and sea changed drastically. This Cenozoic orogeny is thought to have resulted from an overriding of Europe by the continent of Africa, a process that compressed violently the geosynclinal deposits interposed between the ancient shields of both continents. Other processes important in the making of Europe include continued marine transgressions and sedimentation, minor earth movements and rock warpings, and the effect of glaciation, which, notably in the north, wore down elevated areas and, by the process of isostasy, brought new lands above the sea and molded new regional configurations in the coasts. Finally, earthquakes and volcanic activity continue to mold the European continent in the present day, although the process is relatively quiescent except in local terms.

THE TERRITORIAL EMERGENCE OF EUROPE

The continental shield. The oldest rocks that survive in Europe are the shield areas that formed stable nuclei, or older cratons, around which later rocks accreted. Nothing remains of the earliest continental land, and the Fennoscandian (or Baltic) Shield is merely a relic from Precambrian times. The history of earlier Earth surfaces can only be conjectured. This European shield—part of a vast continental area that included Iceland, Greenland, and the Laurentian or Canadian Shield of North America—owed its emergence to orogenesis in Precambrian times, and it was subsequently folded. In the Cambrian Period the sea, occupying an elongated geosynclinal trough, covered most of Fennoscandia; indeed, all of the western peninsular part of the European continent as we know it was covered by sea, although western Ireland, part of Finland, and the broad mass of eastern Europe, namely the Russian Platform, rose above it.

The Fennoscandian Shield

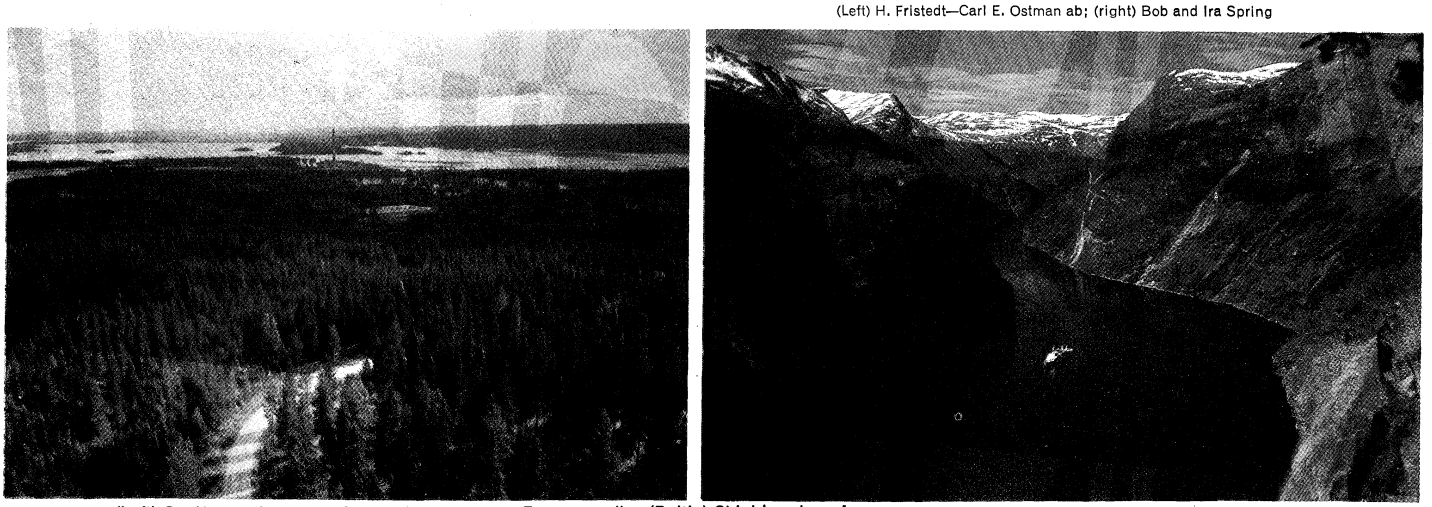
The Fennoscandian Shield presents basement and even surface rocks in Norway, Sweden, Finland, and adjacent Soviet Karelia. In Norway and Sweden they rise to 6,000 feet (2,000 metres). Granites and gneiss, worn down relentlessly over an immeasurable period, predominate, and only a few portions may have escaped marine submergence. Outlying remnants of comparable rocks of Precambrian age lie in western Ireland and northern Scotland. There, as in Norway, mountain folding of the Caledonian orogeny (early Paleozoic Era) occurred. A succession of marine incursions clearly affected parts of the shield, for Cambrian, Ordovician, and Silurian deposits cover the greatest part of the surface of Sweden. Metamorphism has much changed the character of many of the Precambrian rocks, and volcanicity adds complications to the basic geology of Norway and Sweden, Scotland, and Ireland. In Finland, however, the shield surface, where not exposed, is masked by thin Quaternary coverings. While the concept of the Fennoscandian Shield as an ancient, stable core of continent-making has validity, Europe has been exposed to so rich a geological history, including the recent Pleistocene glaciation and isostatic adjustments, that it presents a physical geography of great variety.

The Russian Platform is low-lying, seldom rising above 800 feet (250 metres). It covers most of the European Soviet Union, with projections westward into Romania and Poland, and ends eastward where the Ural Mountains, created by the Hercynian orogeny, are distended north–south. The Russian Shield was worn down over long geological time and invaded by successive seas of the Mesozoic and Cenozoic eras, but it affords a rigid basement, the rocks of which, notably granites, are exceptionally exposed in Kola and the Ukraine.

The role of geosynclines. The first geosyncline relevant to Europe's geological history began in the Cambrian Period, and continued through the Ordovician into the Silurian. Aligned northeast to southwest, and confined between older blocks represented today by the Baltic Shield and relics of the Precambrian rocks of western Britain (then part of an extensive continental area), it became downwarped and filled with shales and coarse sediments, as well as volcanic rocks, to great depths. This process ended at the close of the Ordovician, when the Taconic orogeny presaged the more violent Caledonian revolution of the Silurian Period, when powerful compressional forces, exerted from the shield flanks, created

Caledonian mountain building

(Left) H. Fristedt—Carl E. Ostman ab; (right) Bob and Ira Spring



(Left) Coniferous forests and lakes on the ancient Fennoscandian (Baltic) Shield region of Finland. (Right) Scenic fjord, or sea inlet, winding deep into the mountainous coast of western Norway.

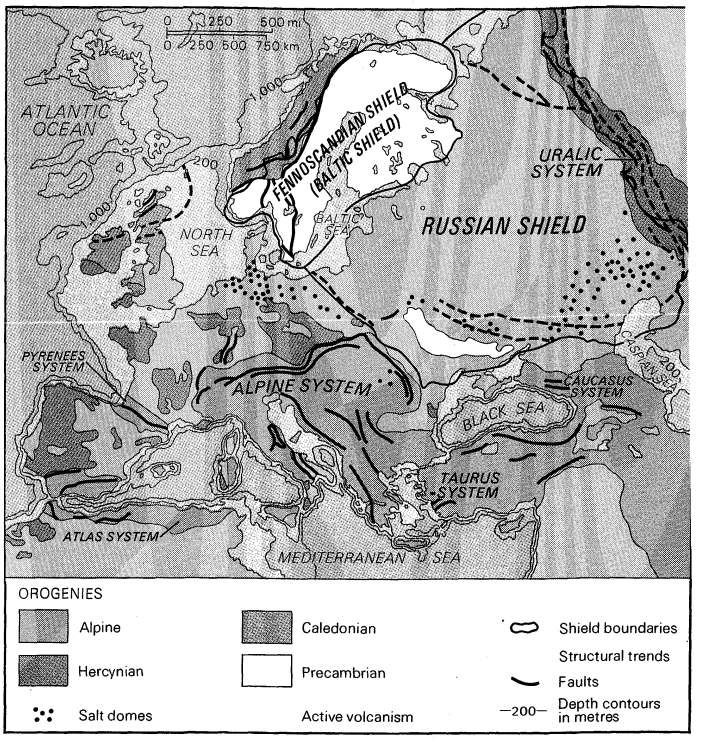
the mountain systems that now cover most of Norway and parts of northwest Britain. Granitization and migmatization dating from this orogeny have also left visible effects on the landscape of these regions. This mountain-building period also produced an uplift of the Baltic Shield, as well as of Arctic and Danish lands. One result of the Caledonian orogeny was that the Hebrides mass, Greenland, and the Baltic–Russian shield grew together to form the Old Red Sandstone continent. A vast new geosyncline, an outstanding feature of the succeeding Devonian Period, formed southward of this dry land: it has been named Hercynian, Variscan, Appalachian, or Altaid, according to region.

In Devonian times the Old Red Sandstone continent consisted of most of the British Isles, its neighbouring seas, parts of northern Europe including the Baltic Shield, and much of the Russian Platform. The Devonian geosyncline, southward of this, edged into Ireland and southwest England and Scotland and stretched deeply into northern Africa and eastward across the present continent beyond the area of the Black Sea.

Alpine mountain building

The Hercynian (or Variscan) orogeny, as violent and probably as lengthy as its Caledonian predecessor, involved Earth movements during Devonian and Carboniferous times, when marine transgression reached its maximum, and culminated in important mountain building late in the Carboniferous Period. Its effects persist widely: mountain ranges were thrown up from Ireland through Britain, Spain, France, Germany, Bohemia, and the Soviet Union. Mineralization and the preservation of coal measures in basin areas were important economic results of this orogeny. Despite all that has happened geologically since Upper Carboniferous times, the lava-penetrated upland and highland massifs of this orogeny survive throughout Europe. Also in the Upper Carboniferous, the sea intruded to divide the Old Red Sandstone land into a Greenland and a Fennoscandian block.

One more major geosyncline is notable. It appeared at the end of the first period of the Mesozoic Era after the deposition of Triassic rocks, including halite (rock salt) and gypsum, on the continental surface of Europe. Known as the Tethys Sea, it originated in areas of subsidence associated with the Hercynian orogeny and covered Alpine areas as well as those of the Mediterranean Sea and lands of today, swinging northeastward across western Siberia and also across the depression west of the Urals in eastern Europe. It lay south of a low-lying desert land carrying deposits of New Red Sandstone, marls, and slates of the Triassic Period, on which further deposits accumulated during the subsequent Jurassic and Cretaceous periods. The compression of the deep sediments of the Tethys between lateral forces exerted by the African Shield against the Paleozoic rocks and structures of this desert land explains the third and final mountain-building period of Europe, known as the Alpine. Starting in the Eocene Epoch of the Cenozoic Era and culminating in the Miocene Epoch, this created the youthful mountain ranges, notably the Alpine system itself, but also those of the Pyrenees, Sierra Nevada, Carpathians, Dinaric Alps, and, just outside the continent, the Caucasus and Atlas ranges.



Tectonic map of Europe.

THE LEGACY OF SUCCESSIVE ERAS

The varying processes described above have bequeathed a geological heritage to contemporary Europe, with each era leaving a legacy of rock layers that differ in both type and distribution.

Precambrian eras. Stretching over an estimated 3,500,000,000 years, the Precambrian eras provided basement rocks for the greater part of the continent: they can be subdivided into an older Archeozoic (or Archean) and a younger Proterozoic (or Algonkian) Era. While most of the Precambrian rocks are deeply covered, they outcrop in Fennoscandia, in Kola and Karelia, in the Ukraine, in Scotland, and, sporadically, in western and central Europe. In age they range between 1,700,000,000 and 3,400,000,000 years, and it is clear that these rocks are remnants of many distinct mountain-building periods and include considerable metamorphic and volcanic components. Five structural units trending northwest to southeast can be distinguished in northern Europe: the oldest is the Saamian, in Kola; abutting this to the southwest is the Marealbian, younger by upwards of 600,000,000 years; the next oldest, the Svecofennian, occupies most of Sweden and Finland and underlies the Baltic Sea and much of European Soviet Union; the Karelian system, stretching across northern Sweden and Karelia, and of about equal size, abuts its eastern flank; finally, on the western flank of the Svecofennian, remnants of the Gothic orogeny are represented by the belt of country southward from Oslo to Skåne Landskap (province), Sweden, as also (probably) by the country westward of it. Outlying relics of Precambrian origin, represented by the Ukraine massif and the oldest rocks of western Scotland, conform with the Baltic Shield in respect to the age and character of their rocks, those of Ukraine being the older. Elsewhere in Europe the Precambrian floor is visible in such elevated areas as the Rhine Plateau, the Bohemian Massif, and the Central German Uplands.

Precambrian rock variety

Precambrian rocks include all types—sedimentary, igneous, and metamorphic—and experienced many cycles of sedimentation, orogeny, granitization, and erosion. Granites, schists, gneisses, and other migmatites are common, as are very thick beds of sedimentary rocks, notably limestones, dolomites, and red sandstone. Glaciations have also left deposits, especially in Finland. These eras have, as yet, yielded few fossils, the commonest being algal forms, and it is inferred that soft-bodied organisms, such as worms, existed but left no record. With a climate broadly cool and humid like that of today but more severe during ice ages, the shield areas must be conceived as having been virtual deserts, although an anthracite-coal deposit in Finland and graphite deposits elsewhere suggest an abundant plant life at one time. It is for its rich contribution of economic minerals—as well as for the many unsolved problems concerning the period—that the Precambrian is of outstanding interest. Ironstone in Arctic Sweden and the Ukraine, graphite, and nonferrous metals are among the varied and abundant sources of wealth the shield areas offered to the enterprise of European peoples.

Proto-Europe of Precambrian times cannot be firmly delineated on a world map. The Caledonian Geosyncline had come into existence, flanked by the edge of the Baltic Shield in Norway to the east and by western Scotland,

including the Hebrides, in the west. There was no Atlantic Ocean, and Greenland, together with the rest of North America, presents a Precambrian history similar and related to that of Europe. The shield areas of Europe, above all in the north and east, at the close of the Precambrian had been worn down to a low plateau, or peneplain, for it is upon this that Cambrian deposits lie.

The Paleozoic and Mesozoic eras. Geological time since the Precambrian amounts to only one-sixth the preceding era. Of a total of 670,000,000 years, the Paleozoic Era occupies the greater part, some 375,000,000 years.

The Cambrian Period. The Cambrian Period, following an ice age, brought an amelioration of climate and considerable marine incursions over the Baltic Shield, the northwest Soviet Union, and parts of central and western Europe. The Caledonian Geosyncline, in fact, extended from the region between the shield areas of the Baltic and northern Britain over much of the Baltic Shield itself and over much of the rest of the continent. Mountain building and volcanic activity was at a minimum, and Cambrian rocks are remarkable for the diversity of fossils of numerous marine invertebrates with hard parts. A thin cover of Cambrian sediments, shales, and limestones overlies the sub-Cambrian floor of the Baltic Shield, arranged, where not subsequently disturbed, in horizontal layers. Lower Cambrian deposits survive in the western parts of the Russian Platform, blue clays (silts and sandstones) are found in the east Baltic lands, while very deep strata were laid down in the geosynclinal areas of Norway and Wales. Cambrian rocks also enter into the structure of Brittany, Bohemia, Germany, Spain, and Arctic areas. Much of Europe as it exists today was left at the end of the Cambrian Period as parts of a wide geosynclinal system swinging from western Scandinavia and the British Isles through western and parts of central and southern Europe, including the Mediterranean area, leaving the eastern part of the Baltic Shield and the Russian Platform uncovered by the sea.

Marine fossils

The Ordovician Period. The succeeding Ordovician Period was marked by a warm climate, more extensive marine transgressions, a diversification of invertebrates (including new forms, notably specimens with calcareous hard parts), intermittent volcanic effusions, and the beginning of the Caledonian orogeny. The Ordovician seas covered the British Isles, the western half of the Baltic Shield, much of the northern Soviet Union, and much of western, central, and southern Europe. Structural and magmatic activity characterized the Caledonian Syncline and its extensions to the south and east, while the Baltic Shield was subject to peripheral undulations. Ordovician rocks occur widely in Europe: in Brittany, the Ardennes, the Black Forest, Saxony, Iberia, Sardinia, the eastern Alps, and Bohemia. Fine-grained limestones of organic origin are the predominant new feature. Iron ores are associated widely with some of these limestones, as in northern Spain, Bohemia, Thuringia, and Brittany.

Limestone and iron ore

The Silurian Period. The Silurian Period is believed to have been one of increasing warmth, although colder at the very end. It witnessed the maximum extent of the Lower Paleozoic seas, which covered the British Isles, the Baltic lands, and much of the European continent, leaving strata of shales, marls, and limestones. The period closed with the Caledonian mountain-building movement, and thus magmatic rocks are also found. Shallow-water invertebrates continue to characterize fossil life, but the Later Silurian period is notable for the emergence of early land plants and the first true (freshwater) fish. The Silurian is devoid of metallic minerals, but it left some thin salt beds that have been put to economic use.

Later periods. The Permian and Triassic periods are noteworthy for the accumulation of chemical salts during phases of semiarid climate. The Carboniferous period witnessed the Hercynian (or Variscan) orogeny, involving the folding of the sediments in a geosyncline newly formed during the preceding Devonian Period. It stands out for its coal-forming forests, which flourished in a moist subtropical climate in swampy ground. The Cretaceous Period was notable for the deep cover of chalk that was deposited widely across Europe and also for the appearance of granite in the Pyrenean area, a hint of the Alpine mountain-building period to follow in the ensuing era.

The Cenozoic Era. It is in the Cenozoic Era, which lasted only some 65,000,000 years, that Europe takes effective and recognizable shape, revealing, successively, ancestral forms of mammals, including apes, and the full development of flowering plants.

The Eocene, Oligocene, and Miocene epochs. Intense compressional forces from the south acted on the deep sediments of the Tethys Sea, pushing them northward against shield obstacles, from the Eocene to the Miocene Epoch. The forces exerted were so powerful they produced complicated folding and even total inversions of rock strata over a wide area from Spain through central Europe to the Balkans and the Caucasus. The first beginnings of Italy are seen with the emergence of islands that would be later contained within the Apennine system and of mountainous structures that joined the Piedmont Alps, via Corsica and Sardinia, with the Atlas Mountains of North Africa. Volcanic outpourings and the formation of adjacent depressions within the Danube Basin and the area that became, as seas withdrew and deposits accumulated, the north Italian Plain also characterized the crustal disturbances of this period.

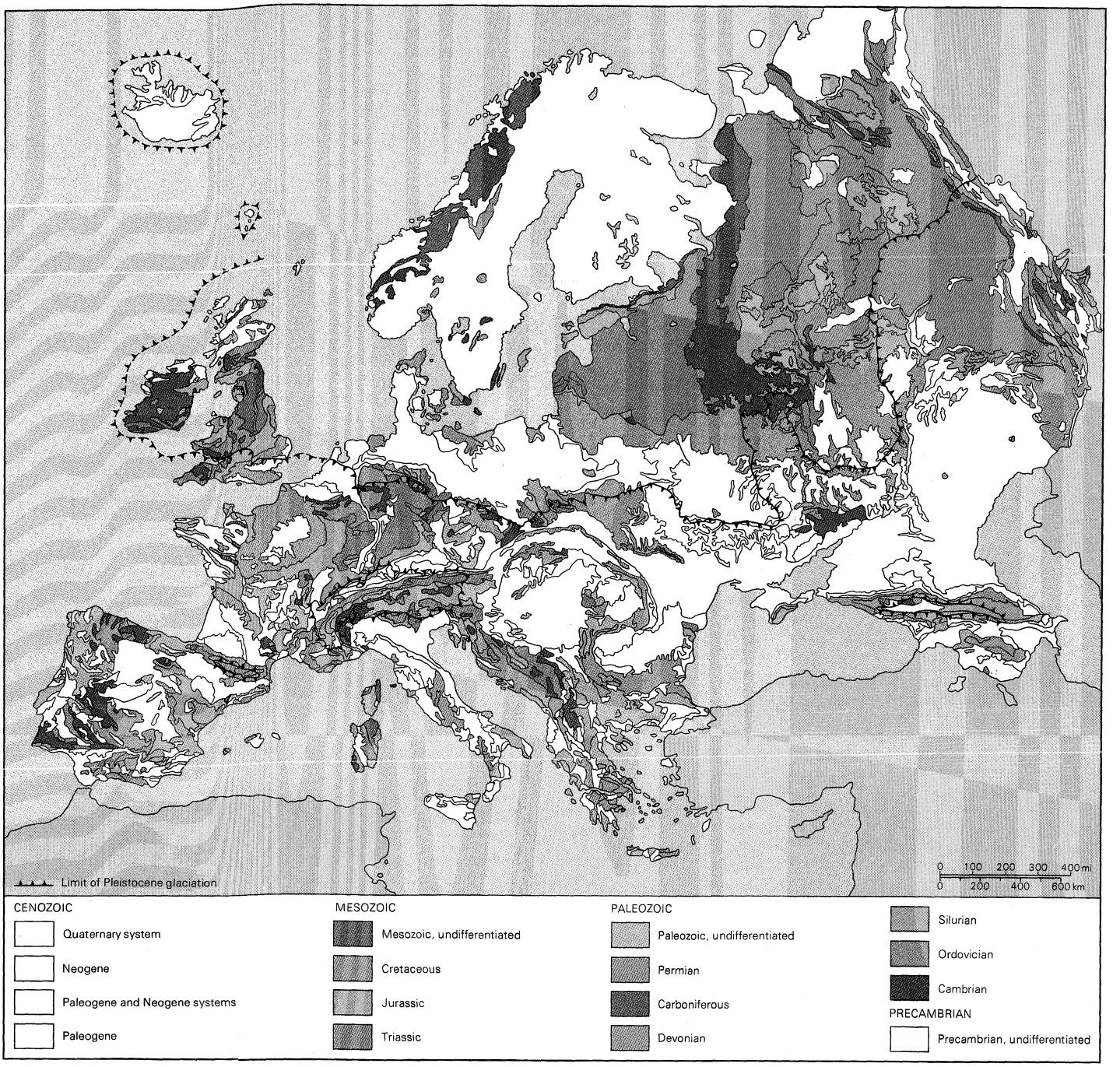
The Pliocene Epoch. Many events in the physical shaping of southern and southeastern Europe occurred during the final Pliocene Epoch. They included the emergence of the Italian peninsula, the creation, by faulting and subsidence, of the islands of Corsica and Sardinia, and the drying of the Hungarian and Romanian plains. A large Aral–Caspian sea survived even into postglacial times, being cut off from the Mediterranean by the Balkan–Caucasus uplift and by dry land where the Black Sea now lies. In postglacial times, after the deep Black Sea had been formed by subsidence, the Dardanelles Strait became cut down, giving access to the Mediterranean. The latter was made up of many distinct basins, notably the Ionian and Aegean, and was studded with insular fragments of either shattered massifs or of Tertiary mountains.

The Quaternary Period. The Great Ice Age, now more usually called the Pleistocene Epoch, occupies the Quaternary Period (2,500,000 years), with the exception of the last 10,000 years, which are called the Holocene or Postglacial Epoch. Although the precise causes of this ice age are controversial, it is known that northern Europe had risen to stand much higher than now and that ice formed to great depths there, as in the rest of the Atlantic landmass and the Alpine areas. The Pleistocene was punctuated by warm interglacial periods separating glacial advances; during its latter part, hominids (early forms of man) occupied niches in the more southerly parts of the continent. They were precursors of *Homo sapiens*, who emerged about 40,000 years ago.

Hominids and *Homo sapiens*

Glaciers are the most powerful engines provided by nature for the transport, by plucking or quarrying, of large masses of rock, and certainly the European glaciers transformed the physique both of their source areas and those to which they moved. Many physical forms of northern and Alpine Europe resulted from glacial erosion, supplemented by weathering, and the areas where the glaciers eventually withered away had their surfaces provided by masses of transported material. Southern Scandinavia, southern Finland, the Swiss Plateau, and the whole North European Lowland were thickly plastered with a variety of forms, including boulder-studded clay, gravels, sands, and the windblown deposits known as loess. And new drainage patterns were formed. The melting of so much ice raised the level of the oceans by an estimated 320 or more feet, while former ice-clad lands, including the North Sea area, began to rise isostatically. It was not until quite late in the Postglacial Epoch that the northern seas of Europe—the Irish, North, and Baltic—took, by stages, their present shape.

The modern geological framework. Although the exposed rocks of Europe are obscured increasingly by the works of man, and while detailed understanding of rock



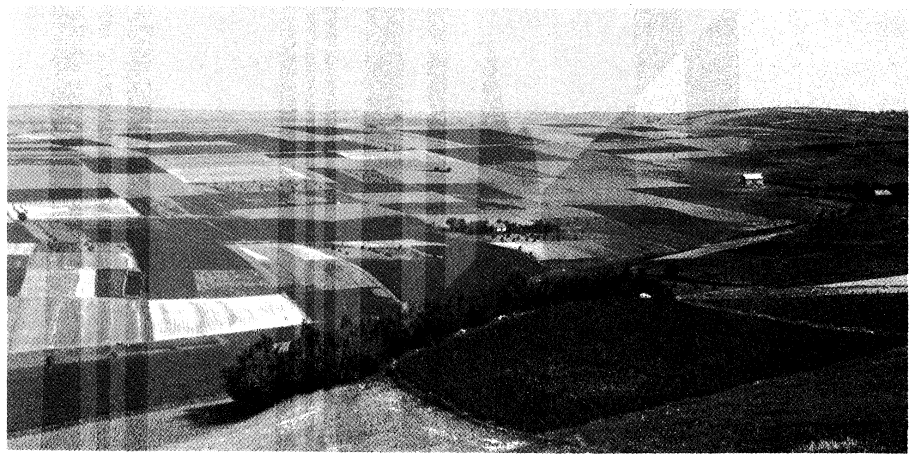
Geologic map of Europe.

Differences between north and south Europe

patterns present challenges even to the expert, the major formations of the continent are clear. In the north lie wide areas of ancient worn-down rocks, stripped of soil by the glaciers but compensated in some measure by the coastal plains created by uplift. In contrast, southern Europe, although incorporating such relics as massifs of Paleozoic rocks, is essentially a youthful world, not yet fully fashioned, as continuing seismic disturbances remind us. Eastern Europe, based on the great Russian Platform, is a stable world still young in surface, since the floor of its shield rocks is deeply concealed beneath Mesozoic and Tertiary deposits, above which glacial material covers the northern half and loess deposits enrich the south. Although in scale this platform is a continental area, river development facilitates access to inland seas both in the north and the south. Ancient rocks, where they lie near the surface, offer mineral wealth, and the former Volga–Ural seas have left a residue of petroleum and mineral salts. For the rest, western and central Europe show great diversity of landforms and landscape as well as varied soil and mineral resources. Alpine ranges in the south and southeast, upraised since their initial orogeny, combine high altitude and relief with scenic attractions and—more importantly—with high precipitation and water dispersion. Highland areas, which are remnants of faulted Hercynian folds surrounded by younger strata, provide another type of wooded landscape, with their contained coalfields. Lastly, lowlands, of outstanding human value, recall their varied origins—former sea and lake basins; lowlands of glacial deposition; parts of eroded synclinal structures; and alluvial and marine plains won from the sea by isostasy or, as exemplified by the Dutch polders, by the triumphant—if precarious—work of man.

II. Physical geography

A contrast exists between the configuration of peninsular (or western) Europe, and east Europe, which is a much



Rolling expanse of the east European plain, consisting of glacial deposits, in southern Poland.
D.C. Williamson

Peninsular and east Europe

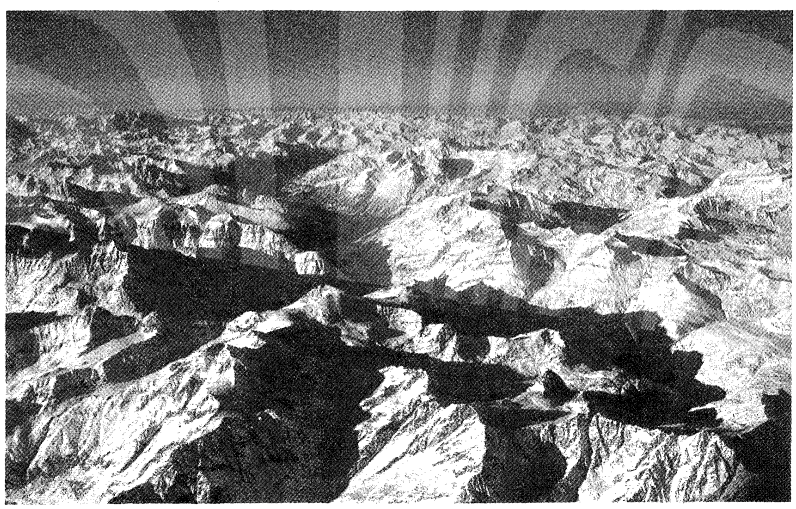
larger and more continental area. A convenient division is made by a line linking the base of the Jutland Peninsula with the head of the Adriatic Sea. The western part of the continent clearly enjoys a high proportion of coastline with good maritime access and often with inland penetration by navigable rivers. Continental shelves —former land surfaces covered by shallow seas—are a feature of peninsular Europe, while the coasts themselves are both submerged or drowned, as in southwest Ireland and northwest Spain, and emergent, as in western Scotland and south Wales where raised former beaches are in evidence. East of the Vistula, Europe's expansive lowlands have something of the scale and character of those of north Asia, but the continent also possesses numerous islands, some—notably the Faroes and Iceland —standing at a distance from the mainland. Fortuitously, Europe has no continuous mountain obstacle aligned north–south, corresponding, say, to the Western Cordillera and the Andes in the Americas, that would limit access into western Europe from the ocean.

RELIEF

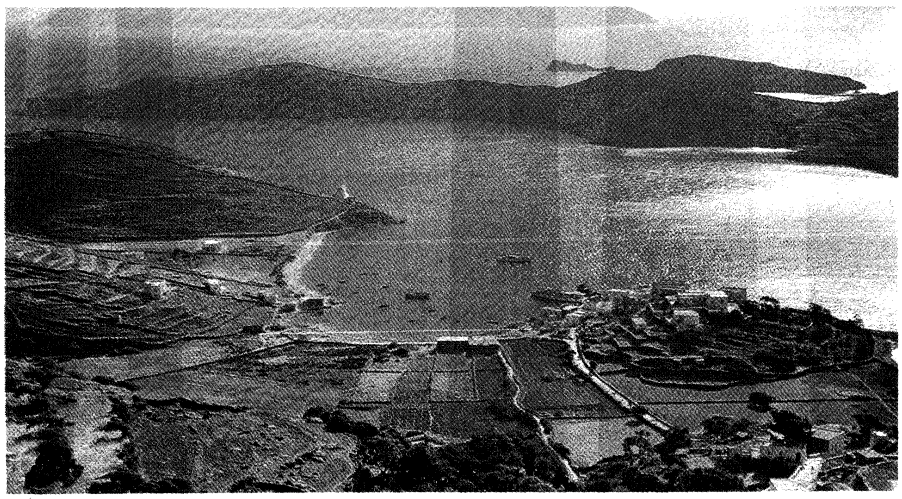
Elevations. Lands lying at high altitude can, of course, be lands of low relief, but in Europe relief tends to become more rugged with altitude. The greater part of Europe combines low altitude with low relief. Only hill masses under 800 feet (250 metres) in height rise gently within the Great Russian Plain which continues northward into Finland, westward into the North European Lowland, and southward in the Romanian, Bulgarian, and Hungarian plains. The North European Lowland, common to much of Poland, north Germany, and Denmark, broadens in western France and continues, across the narrow seas, in southeast Britain and Ireland. The major peninsula of Scandinavia is mostly upland and highland, with relief greatest at the descent to the Norwegian fjords and the sea; eastward and southward the seas are approached more gently. The highest points reached in Norway and Sweden are, respectively, Galdhøppigen (8,097 feet [2,468 metres]), and Kebnekaise (6,965 feet [2,123 metres]). Iceland's highest peak is Hvannadalshnúkur, at 6,952 feet (2,119 metres), while Ben Nevis, the highest summit in Britain, stands at only 4,406 feet (1,343 metres). Greater relief is found in those areas in the heart of western and central Europe where uplifted and faulted massifs survive from the Hercynian orogeny. The worn-down Ural Mountains also belong in this category, and their highest point, Narodnaya, (6,214 feet [1,894 metres]) corresponds approximately to that of the French Massif Central. Altitudes in these areas are mainly between 500 and 2,000 feet, and many steep slopes occur.

The highest altitudes and the most rugged relief of the continent are found further south, where the structures of the Cenozoic orogeny provide mountain scenery. In the Alps, Mt. Blanc rises to 15,771 feet (4,807 metres), and, in the Pyrenees and the Sierra Nevada of Spain, the highest peaks exceed 11,000 feet (3,400 metres). The Apennines, Dinaric Alps, and Balkan Mountains, as well as the arc-shaped Carpathians and Transylvanian Alps, exhibit high altitudes. Above all, in southern Europe—Austria and Switzerland included—level, low-lying land is

(Left) Josef Muench; (right) Harrison Forman



(Left) Mountain ranges of the Swiss Alps in winter, aerial view. (Right) Deep river-cut gorge in the Carpathian Mountains of Romania.



Coastal landscape of submerged mountains and resulting islands and bays characteristic of Greece along the Aegean Sea.
Josef Muench

scarce, and mountain, plateau, and hill landforms dominate. The lowest terrain in Europe, virtually lacking relief, stands at the head of the Caspian Sea many feet below sea level.

Physiographic units. Four topographic units can be simply, yet usefully, distinguished in Europe: coastal and interior lowlands, central uplands and plateaus, the northwest highlands, and the mountainous world of southern Europe.

Lowlands. More than half of Europe consists of lowlands, standing mostly below 600 feet but, infrequently, rising to 1,000 feet. Most extensive between the Baltic–White seas in the north and the Black–Azov–Caspian seas in the south, the lowland narrows westward, lying southward of the northwest highlands; it is divided also by the English Channel and the mountains and plateaus of central Europe. The Danubian and northern Italian lowlands are thus mountain-ringed islands. The northern lowlands are areas of glacial deposition and, accordingly, their surface is diversified by such hills as the Valdai and the North German Heights; by deposits of boulder clay, sands, and gravels; by glacial lakes; and by the Pripet Marshes, a large ill-drained area. Another important physical feature is the southeast–northwest zone of wind-blown loess deposits that have accumulated from eastern Britain to the Ukraine. This *Börde* (German: "edge") belt lies at the northern foot of the Central European Uplands and the Carpathians. Southward of the limits of the northern ice sheets are vales and hills, with the Paris and London basins typical examples. Superficial rock cover, altitude, drainage, and soil have sharply differentiated these lowlands—which are of prime importance to human settlement—into areas of marsh or fen, clay vales, sand and gravel heaths, or river terraces and fertile plains.

The northern lowlands

Central uplands and plateaus. The central uplands and plateaus present distinctive landscapes of rounded summits, steep slopes, valleys, and depressions. Examples are the Southern Uplands of Scotland, the Massif Central of France, the Iberian Meseta, and Bohemia. Routes detour around, or seek gaps through, these uplands, whose German appellation, *Horst* (literally "thicket"), recalls their still wooded character, while their coal basins give them great economic importance. The well-watered plateaus give rise to many rivers and are well adapted to pastoral farming. Volcanic rocks add to the diversity.

Northwest highlands. The ancient, often mineral-laden, rocks of the northwest highlands, their contours softened by prolonged erosion and glaciation, are found in much of Iceland, Ireland, and in north and west Britain and Scandinavia. They include lands of high rainfall, which supplies hydroelectricity and water to industrial cities, and summer pastures for cattle but are of little use for crops. The coasts—notably the fjords of Norway—invite maritime enterprise.

Southern Europe. A world of peninsulas and islands, southern Europe is subject to its own climatic regime, with fragmented but predominantly mountain and plateau landscapes. Iberia and Anatolia (Turkey) are extensive peninsulas whose interior tablelands of Paleozoic rocks are flanked by mountain ranges of Alpine type. The restricted lowlands lie within interior basins or fringe the coasts; those of Portugal, Macedonia, Thrace, and north Italy are relatively large. The Alps furnish much water for electric generating stations, as for major rivers.

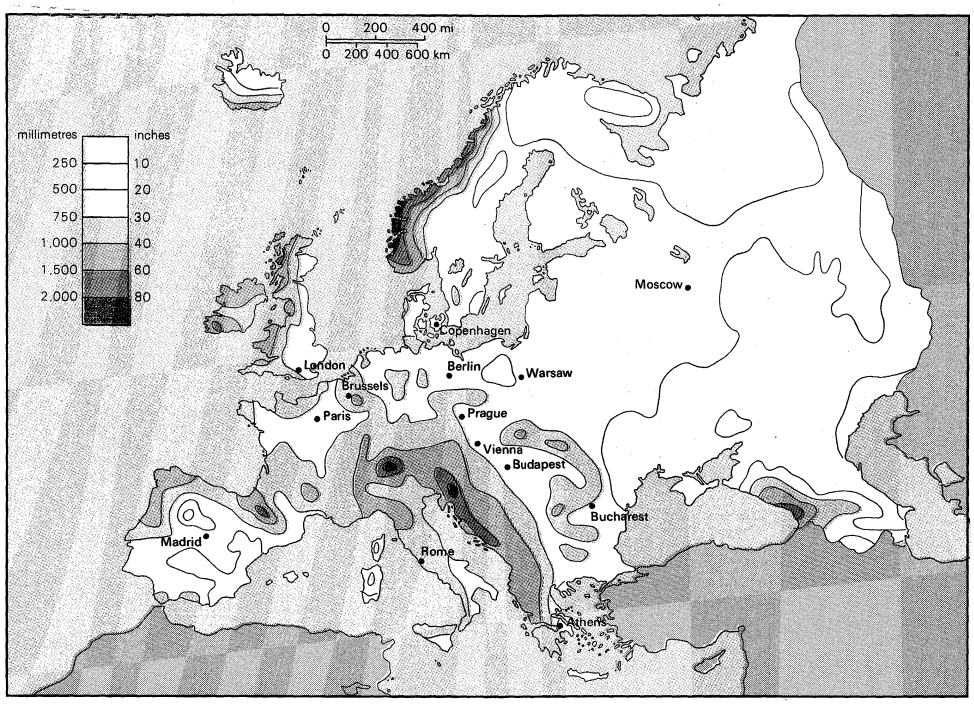
Peninsulas and islands of the south

CLIMATE

As Francis Bacon, the great English Renaissance man of letters, aptly observed, "every wind has its weather." It is air-mass circulation that provides the main key to Europe's climate, the more so since masses of Atlantic Ocean origin can pass freely through the lowlands, except in the case of the Caledonian mountains of Norway. Polar air masses derived from areas close to Iceland and tropical masses from the Azores bring, respectively, very different conditions of temperature and humidity and produce different climatic effects as they move eastward. Continental air masses from eastern Europe have equally easy access westward. The almost continuous mountain belt trending west–east across Europe also impedes the interchange of tropical and polar air masses.

Air-pressure belts. Patterns of some permanence controlling air-mass circulation are created by belts of air pressure over five areas. They are: the Icelandic Low, over the North Atlantic; the Azores High, a high-pressure ridge; the (winter) Mediterranean Low; the Russian High, centred over Central Asia in winter but extending westward; and the Monsoon Low, a low-pressure system over southwest Asia. Given these pressure conditions, westerly winds prevail in northwest Europe during the year, becoming especially strong in winter. The winter westerlies, often southwesters, bring in warm tropical air; in summer, by contrast, they veer to the northwest and bring in cooler Arctic or sub-Arctic air. In Mediterranean Europe the rain-bearing westerlies chiefly affect the western areas, but only in winter. In winter, the eastern Mediterranean Basin experiences bitter easterly and northeasterly winds derived from the Russian High, and their occasional projection westward explains unusually cold winters in western and central Europe, the exceptionally warm winters of which, on the other hand, result from the sustained flow of tropical maritime air masses. In summer the Azores High moves 5–10° of latitude northward and extends further eastward, preventing the entry of cyclonic storms into the resultantly dry Mediterranean region. The eastern basin, however, ex-

Westerlies and the Russian High



Average annual precipitation for Europe.

periences the hot and dry north and northeast summer winds called etesian by the ancient Greeks. In summer, too, the Russian High gives place to a low-pressure system extending westward, so that westerly air masses can penetrate deeply through the continent, making summer a wet season.

It is because of the interplay of so many different air masses that Europe experiences very changeable weather. Winters get sharply colder eastward, but summer temperatures relate fairly closely to latitude. Northwest Europe, including Iceland, enjoys some amelioration due to warm Gulf Stream waters, which keep the Soviet port of Murmansk open all the year.

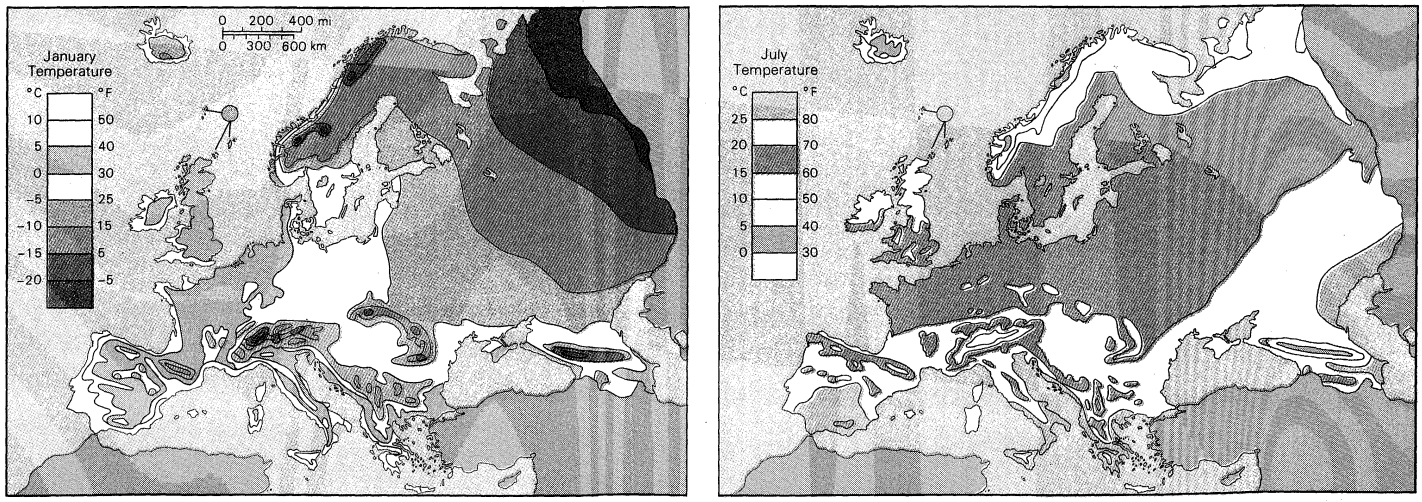
Climatic regions. Four regional European climatic types can be loosely distinguished, each characterized by much local topographically related variation. Further, the great cities of Europe, because of the scale and grouping of their buildings, their industrial activities, and the layout of their roads, create local climates—including a central "heat island" and pollution problems—quite distinct in themselves (see URBAN CLIMATES).

Maritime climate. Characterizing western areas heavily exposed to Atlantic air masses, the maritime type of climate—given the latitudinal stretch of these lands—exhibits sharp temperature ranges. Thus the January and July annual averages of Reykjavík (Iceland) and Coruña (Spain) are, respectively, 32° F (0° C) and 53° F (12° C), and 50° F (10° C) and 64° F (18° C). Precipitation is always adequate—indeed, abundant on high ground—falling the year round. The greatest amount of precipitation occurs in autumn or early winter. Summers range from warm to hot depending on the latitude and altitude, and the weather is everywhere changeable. The maritime climate extends across Svalbard, Iceland, the Faeroes, the British Isles, Norway, southern Sweden, western France, the Low Countries, northern Germany, and northwest Spain.

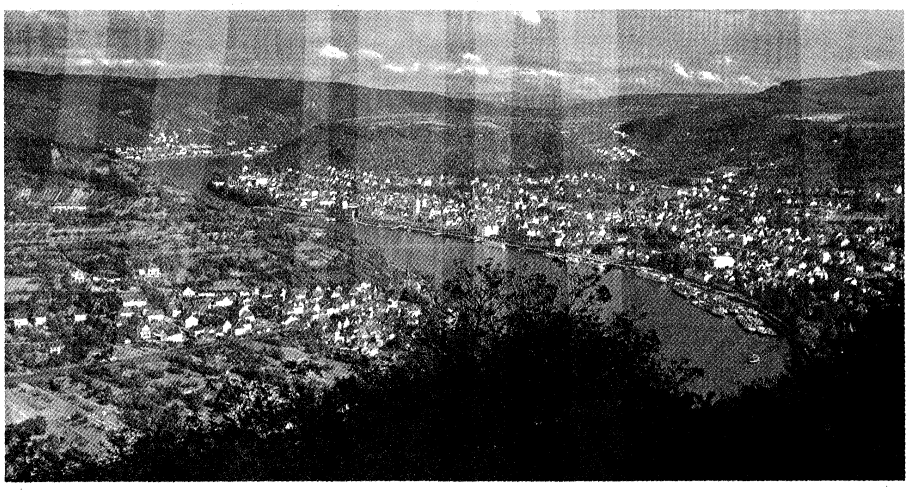
Central European (transitional) climate. The central European, or transitional, type of climate results from the interaction of both maritime and continental air masses and is found at the core of Europe, south and east of the maritime type, west of the much larger continental type, and north of the Mediterranean type. This rugged region has colder winters, with substantial mountain snowfalls, and warmer summers, especially in the lowlands. Precipitation is adequate to abundant with a sum-

Increasing continentality

Adapted from Norton S. Ginsburg (ed.), *Aldine University Atlas* (1970), Aldine Publishing Co., Chicago; copyright © 1970 by George Philip & Son Ltd., London; with permission from the author and Aldine-Atherton, Inc.



Average temperatures for January and July in Europe.



Meander in the Rhine River Valley at Boppard, West Germany.
J. Allan Cash

mer maximum. The region embraces central Sweden, southern Finland, the Oslo Basin of Norway, eastern France, southwest Germany, and much of central and southeastern Europe. The range between winter and summer temperatures increases eastward, while the rainfall can reach over 80 inches (2,000 millimetres) in the mountains, with snow often lying permanently around high peaks. The Danubian region has only modest rainfall (24.2 inches [615 millimetres] per year at Budapest), but the Dinarid mountains experience heavy cyclonic winter, as well as summer, rain.

Continental climate. The continental type of climate dominates a giant share of Europe, covering the European Soviet Union (except the Baltic areas), most of Finland, and northern Sweden. Winters—much colder and longer, with greater snow cover than in western Europe—are coldest in the northeast, and summers are hottest in the southeast; the January to July mean temperature range is high, from 50° to 70° F (10° to 21° C). Summer is the period of maximum rain, which is less abundant than in the west: Moscow's annual average is 25 inches (635 millimetres), while, both in the north and southeast of the Russian Plain, precipitation reaches only between ten and 20 inches annually. In parts of the south, the unreliability of rainfall combines with its relative scarcity to raise a serious aridity problem.

Mediterranean climate. The subtropical Mediterranean climate characterizes the coastlands of southern Europe, being modified inland (for example in the Spanish Meseta, the Apennines, and the north Italian plain) in response to altitude and aspect. The main features of this climatic region are mild and wet winters, hot and dry summers, and clear skies, but marked regional variations occur between the lands of the western and the more southerly lying eastern basins of the Mediterranean; the former are affected more strongly by maritime-air-mass intrusions. Rainfall in southern Europe is markedly reduced in areas lying in the lee of rain-bearing westerlies: Rome has an annual mean of 25.7 inches (653 millimetres), but Athens has only 15.8 inches.

The effects of climate. The local and regional effects of climate on the weathering, erosion, and transport of rocks clearly contribute much to the European landscape, and the length and warmth of the growing season, the amount and seasonal range of rainfall, and the incidence of frost affect the distribution of vegetation. Wild vegetation in its turn provides different habitats for animal life. Climate is also an important factor in the making of soils, while modern European industry and urban life depend increasingly on water supplies, with rivers and lakes continuing to provide important commercial waterways in some areas. The winter freeze in northern and eastern Europe is another effect of climate, and the spring thaw, by creating floods, impedes transport and harasses farmers. The snow cover of the more continental regions is useful to man, however, for it stores water for the fields and provides snow ways for sled users. Regional variations of climate also help determine where crops are grown commercially. In southern Europe the climate supports specially adapted wild vegetation and precludes all-year grass in coastal lowlands, while the practice of moving flocks and herds to pastures seasonally available at different altitudes is clearly adapted to other conditions set by climate. In sum, in only a modest proportion of Europe does climate severely restrict human occupance and land use. These areas include regions of high altitude and relief, such as the sub-Arctic highlands of Fennoscandia and Iceland, and the arid areas of the southeast European Soviet Union and interior Spain.

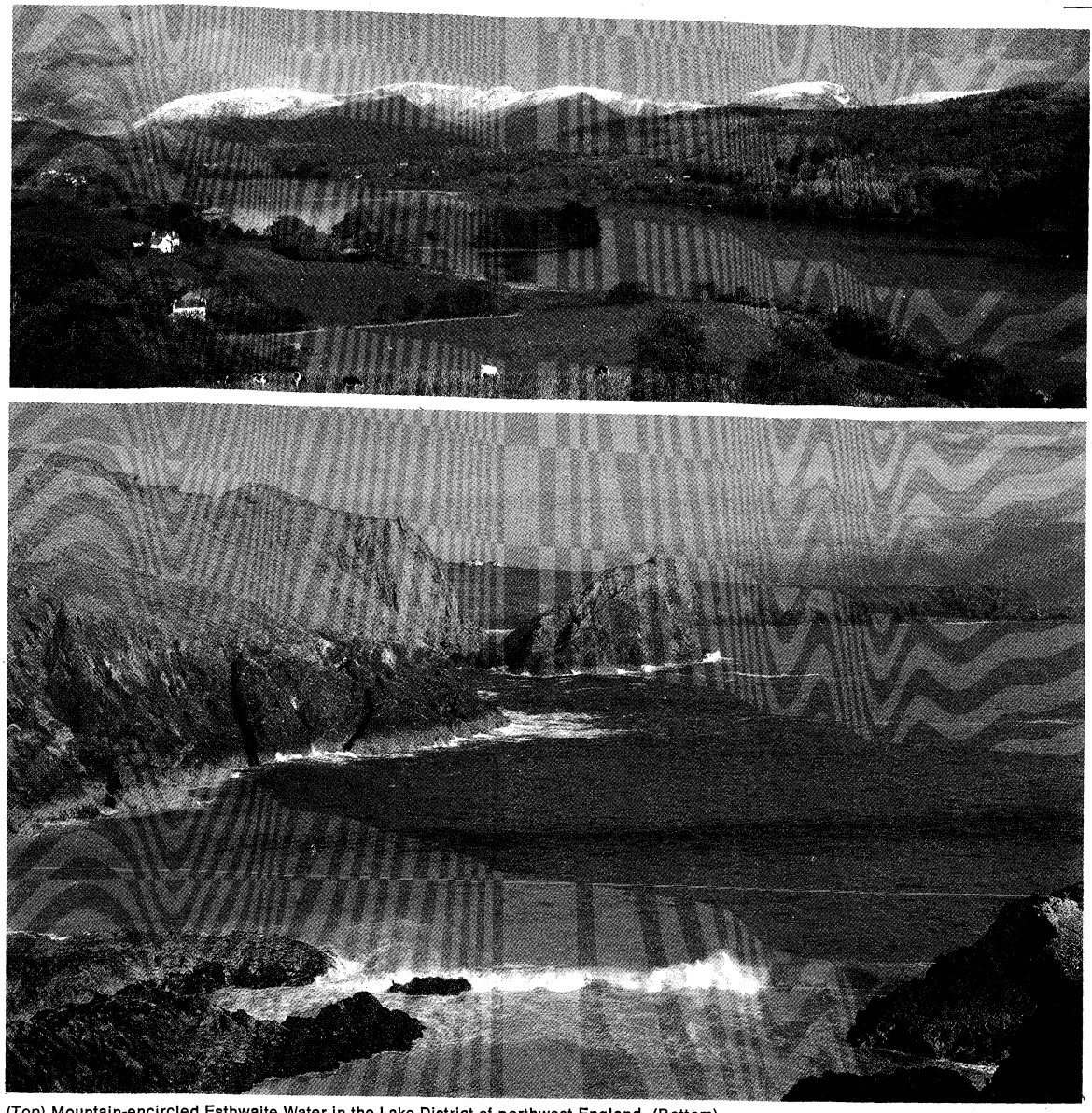
DRAINAGE

Topographic influences. The drainage basins of most European rivers lie in areas originally uplifted by the Caledonian, Hercynian, and Alpine mountain-building periods that receive heavy precipitation, including snow. Some streams, notably in the European Soviet Union, Finland, and Poland, have their sources in hills of Tertiary rocks, while others, including the Thames and Seine rivers, derive from hill country of Mesozoic rocks. Drainage is directly, or via the Baltic and the Mediterranean seas, to the Atlantic and the Arctic oceans and to the enclosed Caspian Sea.

Drainage basins

The present courses and valley forms of the major rivers result from an intricate history involving such processes as erosion by the headstream, downcutting, capture of other rivers, faulting, and isostatic changes of land and sea levels. The Rhine, for example, once drained to the Mediterranean before being diverted to its present northerly course. The courses of many rivers—notably those of Fennoscandia and the North European Lowland—have been shaped since the Pleistocene Epoch. While the Alps, Apennines, and Carpathians provide watersheds, other mountain ranges have been cut through by rivers, as by the Danube at Vienna, Budapest, and the Iron Gate, and the Olt (in Romania). In the Great Russian Plain the rivers are long and flow sluggishly to five seas. In western, central, and eastern Europe, rivers are largely "mature": their valleys are graded, and their streams are navigable. Northern and southern Europe, in contrast, present still "youthful" rivers, as yet ill-graded and thus more useful for hydroelectricity than for waterways. The Atlantic rivers have scoured estuaries widening seaward, while, in the Baltic, Mediterranean, and Black seas, with minimum tidal influences, deltas and spits have been created. The Upper Dnepr, since post-Pleistocene times, has failed to drain effectively the low area of minimal relief known as the Pripet Marshes.

Hydrology. The water volume of, and discharge from, the rivers of Europe are governed by factors that include local conditions of rainfall, snowmelt, and rock porosity. In consequence, the rivers in the western area have more volume and higher discharges in the winter season and



(Top) Mountain-encircled Esthwaite Water in the Lake District of northwest England. (Bottom) Rugged Atlantic coastline, Cornwall, southwest England, with sea stack forming at centre.
(Top, bottom) FPG, (top) Tom Wright, (bottom) Muller-Brunke

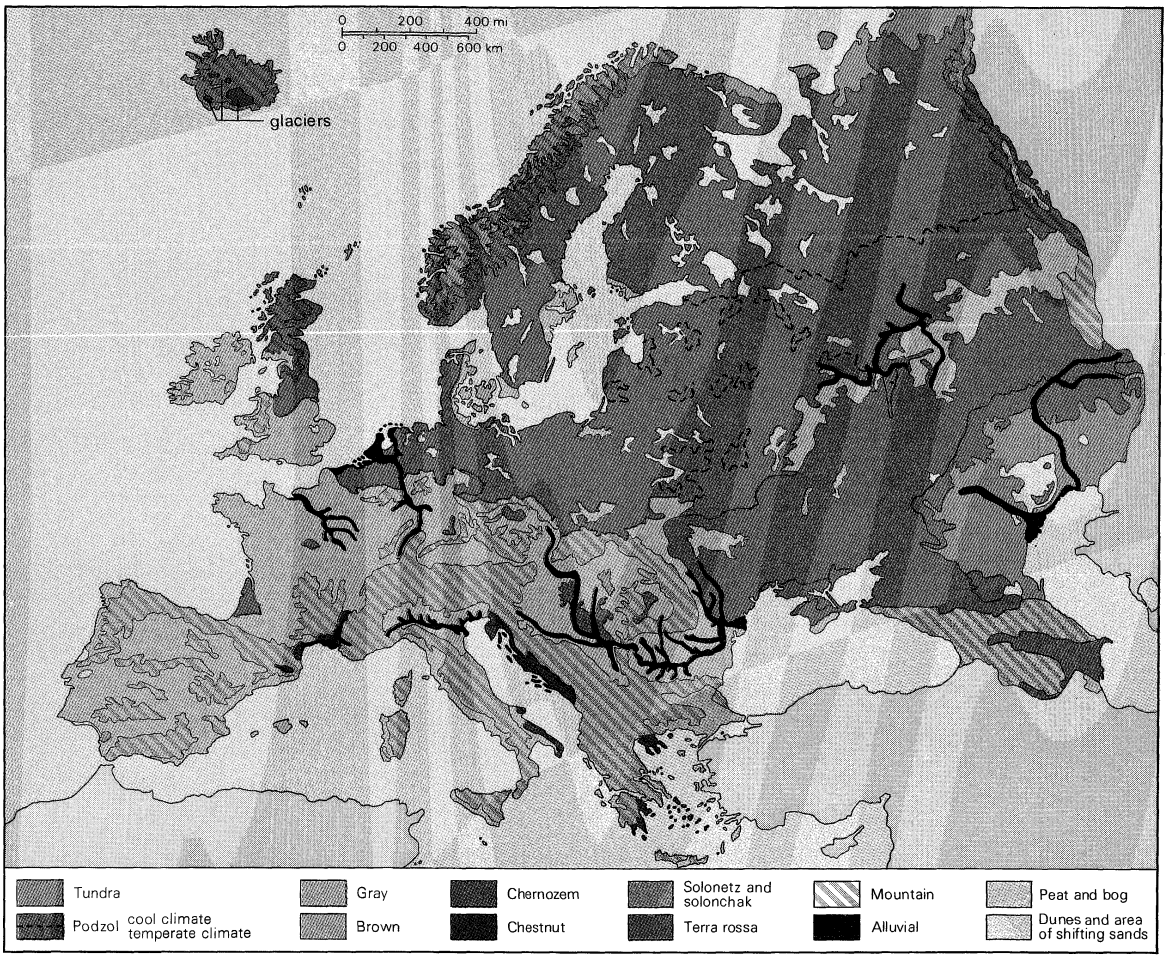
their lowest in summer. In areas of mountainous and continental climate, thanks to the runoff of snowmelt, the rivers are highest in spring and early summer. The longer rivers of the Continent, notably the Rhine and the Danube, have complex regimes, since their basins extend into areas of contrasting climate. Although embanking measures have reduced the problem, flooding is a continued threat. Thus, rivers in the Soviet Union are liable to flood with the spring thaw; oceanic rivers, after heavy or prolonged rain over the whole basin; and Alpine rivers, when the warm foehn wind rapidly melts the snow. In the Mediterranean region some rivers—as in peninsular Greece—tend to dry up in summer through a combination of scant rainfall, evaporation, and porous limestone beds. In the Abruzzi (Italy), however, heavy rainfall, mainly in winter, permeable and porous rocks within the basin, and abundant snow combine to regulate the river regimes.

Flood danger

The Rhône achieves a steady flow throughout the year, deriving a high input from the Cévennes Mountains—which experience heavy winter rain—plus abundant spring and summer snowmelt from the Alps via Lake Geneva. The Rhine and Danube tap supplies from the Alps in spring and summer, and the Rhine, especially, taps areas of winter rainfall maximum. The Volga has its highest water in spring and early summer, thanks to snowmelt, and falls to a summer low. The Saône, lying within the oceanic climatic area, tends to have a good flow all year. The winter freeze of the east only rarely seriously affects the Danube and west European rivers.

Lake systems and marshes. Lakes cover less than 2 percent of Europe's surface and occur mostly in areas subjected to the Pleistocene glaciation. Fennoscandia and the North European Lowland account for four-fifths of the area of lakes; and in Finland lakes cover one-fifth of the surface. The other major zones of lakes lie marginal to the Alpine system, while Scotland, too, has its many "lochs," and Ireland its "loughs." Lakes survive where the inflow of water exceeds loss from evaporation and outflow and should eventually disappear through alluvial accumulation. Their origins lie in the glacial excavation of softer rocks, in the building of dams by morainic material, and in tectonic, or deforming, forces, which may create depressions. This second explanation clearly applies to Alpine lakes; to many of those in the British Isles—including the small but scenic ones of the Lake District of England—and also those of central Sweden. Volcanic crater lakes are found in central Italy, and small lakes of the lagoon type are found along the Baltic and Mediterranean, where spits have lengthened parallel to the coast and hence cut off sea access.

A well-developed zone (*Marschen*) has formed along the low-lying and reclaimed sea marshes of western Ger-



Soils of Europe.

The Marschen zone

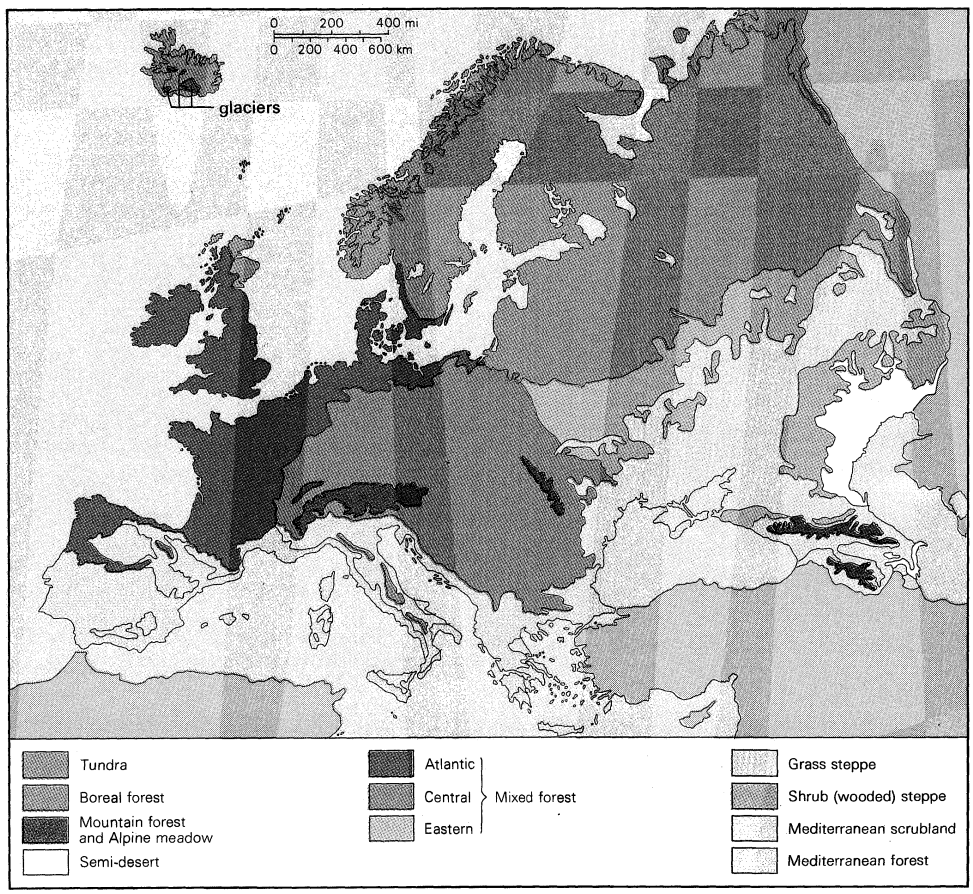
many and The Netherlands, and characteristically the estuaries of Europe's tidal rivers are edged by flat alluvial marshes. Fens, as exemplified by the Dutch polders and the Fenland of England, are made up of either alluvium or peat and stand too low to be drained effectively, except by continuous pumping. The continent's largest marshland is the Pripet Marshes of the Belorussian Soviet Socialist Republic.

SOILS

Regional divisions. The soil patterns of Europe are clearly and zonally arranged in the Great Russian Plain but are much more complicated in the rest of the continent, which exhibits a more varied geology and relief. Tundra soils occur only in the most northerly parts of the Soviet Union and Finland and in high areas of Sweden and Norway; they tend to be acidic, waterlogged, and poor in plant nutrients. South of this zone, and extending around the Gulf of Bothnia and across Finland and the Soviet Union north of the upper Volga, podzols are characteristic. These soils, formed in a coniferous woodland setting, suffer from acidity, the leaching of minerals, hardpan formation and permafrost beneath the topsoil, and excess of moisture; given the climate, they are virtually useless for crops. The larger zone to the south stretches from central U.S.S.R. westward to the British Isles and southward from central Sweden, southern Norway, and Finland, to the Pyrenees, Alps, and Balkan mountains. In this region podzols and brown forest soils have developed in a mixed-forest environment, and these soils, which are very varied, usually have a good humus content. Locally, the farmer recognizes soils of heavy to light texture, their different water-holding capacities, depth, alkalinity or acidity, and their suitability for specific crops. The soils, rich in humus, within this zone that cover loess are excellent loams; lowland clays, when broken down, also exhibit high quality, as do alluvial soils; in contrast, areas covered with dry, sandy, or gravelly soils are more useful for residential and amenity purposes than for farming. In the southeast of the European portion of the Soviet Union, notably in the Ukraine, some soils that have been formed in areas of grass steppe are chernozems (black earths), deep, friable, humus-rich, and renowned for their fertility. In the formerly wooded steppe lying to the north of the grass steppe in both the south central Soviet Union and the lower Danubian lowlands, soils of somewhat less value are known as degraded chernozems and gray forest soils. At best, chestnut soils—some needing only water to be productive—and, at worst, solonetzic (*i.e.*, highly saline) soils cover areas of increasing aridity eastward of the Ukraine to the Ural River. Lastly, in southern Europe, where the countryside is fragmented by mountains, plateaus, and hills, much soil has been lost from sloping ground through forest destruction and erosion, and a bright-red soil (terra rossa), heavy and clay-rich, is found in many valleys and depressions.

Chernozems and terra rossa

Problems of classification. The origin, nature, variety, and classification of Europe's soils raise highly complex problems: so much is involved—bedrock, drainage, plant decomposition, biological action, climate, and the time factor. Man, moreover, has done much to modify soils and, with increasing scientific knowledge, to render soils of greater and continuing value by drainage, crop rotation, and the input of suitable combinations of chemicals. In such ways, naturally poor soils can—as Denmark has shown—be made productive. The practice of an enforced "resting" of soils, by leaving fields fallow to recuperate, began to disappear with the agricultural revolution of the 18th century, and agronomic science continues to show how the best results can be got from specific soils and also how to check soil erosion. Europe's arable land



Vegetation zones of Europe.

lies mainly in the lowlands, which have podzols, forest brown, black earth, and chestnut soils, although the upper level of cultivation, as of animal husbandry, rises southward. New land is won from the sea, and this more than offsets coastal losses through erosion, but the losses to urban expansion and to such competitors for level land as airfields, on the other hand, are becoming increasingly serious.

III. Plant and animal life

VEGETATION

Major vegetation zones. "Natural," "original," and "primitive," as epithets applied to the vegetation of Europe, have no precise meaning unless they are related to a specific time. It is, nevertheless, possible to envisage continental vegetation zones as they formed and acquired some stability during postglacial times, although they are only rarely recalled by present-day remnants.

The tundra. Tundra vegetation, made up of lichens and mosses, occupies a relatively narrow zone in the extreme north of the European Soviet Union and Fennoscandia, although this zone is continued southward in the mountains of Norway. Vegetation of a similar kind occurs at altitudes of 5,000–6,000 feet in the Alps and the northern Urals.

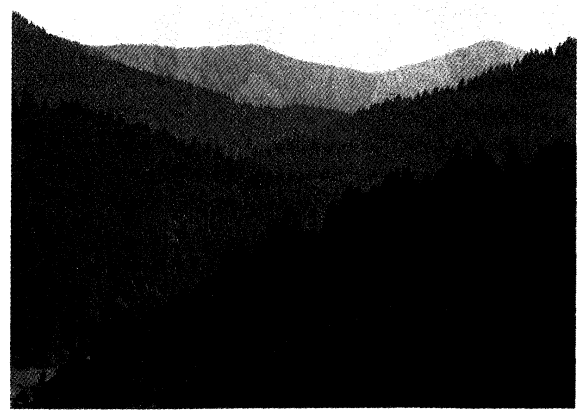
The boreal forest. Southward, the virtually treeless tundra merges into the boreal (northern) forest. The more northerly zone is "open," with stands of conifers and with willows and birch thickets rising above a lichen carpet. It is most extensive in the northern European Soviet Union but continues, narrowing westward, across Sweden. South of this zone, and with no abrupt transition, the "closed" boreal forest occupies a large fraction —mainly north of the upper Volga—of the Soviet Union proper and Fennoscandia. Conifers, thin-leaved and resistant to cold, together with the birch and larch, predominate.

The mixed forest. The northern vegetation may superficially suggest its primeval character, but the zone of mixed forest that once stretched across the continent from the British Isles to central U.S.S.R. has been changed extensively by man. Only surviving patches of woodland—associations of summer-leaf trees and some conifers, summarily described as "Atlantic," "central," and "eastern"—hint at the formerly extensive cover.

The Mediterranean complex. In southern Europe, Mediterranean vegetation has a distinctive character, containing hard-leaf forests and secondary areas of scrub, especially maquis, which is made up of trees, shrubs, and aromatic plants. Such scrub is scattered because of summer drought, particularly in areas where the soil is underlain by limestone or where there is little, if any, soil.

Steppe and semidesert. Three further vegetation zones are confined to the southeastern European Soviet Union, although the first two—the "wooded steppe" and the "grass steppe"—extend into the Danubian lowlands. Finally, semidesert vegetation characterizes the dry lowland around the northern shores of the Caspian Sea.

Marc Riboud—Magnum



Steep hillslopes of the Black Forest region, southern West Germany.

The shaping of vegetation zones. *Climatic change.* The primeval vegetation of Europe began to take shape as the climate ameliorated following the disappearance of the Pleistocene ice sheets. The microscopic study of pollen grains preserved in datable layers of peat and sediments has made it possible to trace the continental spread, in response to climatic improvement, of forest-forming trees. The double barrier of the Alpine mountains and the Mediterranean Sea had checked the retreat of trees at the onset of the Great Ice Age, and there were relatively few indigenous species to return northward from unglaciated refuges. In the first postglacial climatic phase (the boreal), spruce, fir, pine, birch, and hazel nevertheless established themselves as far north as central Sweden and Finland. During the succeeding climatic optimum (the Atlantic phase), which was probably wetter and certainly somewhat warmer, mixed forests of oak, elm, lime (linden), and elder spread northward. Only in the late Atlantic period did the beech and hornbeam spread into western and central Europe from the southeast.

The Alpine barrier

During postglacial times, therefore, when *Homo sapiens*, in small numbers, was finding homes within Europe, the continental surface was thickly clad with trees and undergrowth, except where tree growth was precluded by extreme cold, high altitude, bad drainage, or exposure to persistent gales. Even those relatively attractive areas where windblown loess had been deeply deposited are now known to have had woods of beech, hawthorn, juniper, box, and ash, as did also limestone plateaus. The Mediterranean peninsulas also had evergreen and mixed forests rooted in an ample soil.

The role of man. From prehistoric times onward, with ever increasing force, man, seeking optimum economic use of available resources, has acted as a vigorous agent of vegetation change. The effects of grazing animals may well explain why some heathlands (*e.g.*, the Lüneburger Heide in north Germany) replaced primeval forest. By fire and later by axe, forest clearance met demands for homes and ships, for fuel, for charcoal for iron smelting, and, not least, for more cultivation and pasture. The mixed boreal forests suffered most because their relatively rich soils and long and warm growing season promised good returns from cultivation. The destruction of woodlands was markedly strong when population was growing (as between *c.* AD 800 and 1300). It was later intensified by German colonization east of the Rhine and reached maximum scale in the 19th century. In southern Europe—where naval demands were continuous and sources of suitable timber sharply localized—tree cutting entailed, from classical antiquity onward, serious soil loss through erosion, increased aridity, floods, and marsh formation. Further north throughout the continent, as present distribution of arable land shows, former forests were reduced to remnants; only in the north and below the snow line of Alpine mountains have forests of large and continuing commercial value survived. These coniferous forests of Sweden, Finland, and northern U.S.S.R. are "cropped" annually to preserve their capital value. On the positive side may be noted the reclamation of marshlands and the soil improvement of hill grasslands and heaths, their wild vegetation being replaced by pasture and crops; in timber-deficient countries the afforestation of hillslopes, chiefly with quickly growing conifers, belatedly attempts to restore some of the former forests. Another drastic vegetation change brought about by man has been the virtual elimination of the wooded and grass steppes, which have become much needed granaries.

Afforestation in Scandinavia and the Soviet Union

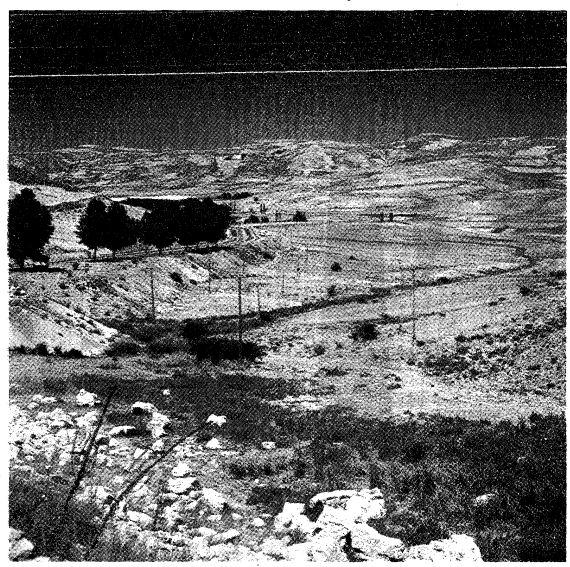
Exterior influences and European survivals. To a surprising degree, European vegetation stemmed from the importation of plants from other continents, although some imported crops—notably citrus fruits, sugarcane, and rice—can only grow marginally in Europe, and then by irrigation. From an original home of wild grasses in Ethiopia, cultivated varieties of wheat and barley reached Europe early, via Southwest Asia and Egypt, as did also the olive, vine, fig, flax, and some varieties of vegetables. Rice, sugarcane, and cotton, of tropical Indian origin, were introduced by the Arabs and Moors, especially into Spain. The citrus fruits, peach, mulberry, oats, and millet reached Europe from original Chinese habitats, and Europe owes maize, tobacco, squashes, tomatoes, red pepper, prickly pear, agave (sisal) and the potato—first grown for fodder but destined to become the cheap staple food for the large families of low-paid workers of the 19th-century—to the Americas. Europe has drawn greatly on eastern Asia and North America for trees, especially ornamental trees, while some acacias and the eucalyptus derive from Australia. Sugar beet, however, was a European discovery, first grown when much of Napoleonic Europe was subjected to maritime blockade.

The forests of northern Europe and the Alpine ranges, although in no sense primeval, represent unchanging land use during the postglacial period. The "closed boreal forest" occupies 1,000,000 square miles, made up of a spruce-fir association (but with stands of pine, birch, and larch) above an undergrowth of mosses and herbs. This large and valuable reserve of timber is of world importance; forests once covered 80 percent of Europe's surface, and they still occupy about 30 percent.

Human adaptations. Clearly, animal life, wild and domesticated, has been adjusted to fit largely man-made patterns of vegetation, which, in turn, reflect age-long attempts to achieve chiefly economic ends. With such endeavours are associated varieties of *modes de vie*, or "means of livelihood." In the mountains, as in the boreal forests, the environment is exploited by winter lumbering and by the transport of felled trees by river after the spring thaw. So, too, agriculture in its many forms—in part for subsistence but commonly for urban markets—is a basic occupation of the lowlands, long cleared of extensive forests or steppe vegetation. In Mediterranean Europe, rural life, based on horticulture and arboriculture rather than on large-scale cultivation, as well as on the rearing of sheep and goats and wheat cultivation, continues, little changed in many areas. For such deeply rooted fruit-bearing trees as the olive and vine, use is made of sloping, broken, and terraced land. Farming also extends to specialized forms with respect to the subtropical crops that climate, sometimes supplemented by irrigation, permits.

Modes de vie

Bjorn Bolstad—Photo Researchers



The Meseta, an arid region in the interior of Spain.

ANIMAL LIFE

Patterns of distribution. With animals as with plants, the earlier Pleistocene range and variety has been much reduced since man disposed of what nature provided. Wild fauna has been long in retreat since Upper Paleolithic times, when, as cave drawings portray, small human groups held their own against such big game as aurochs and mammoths, now extinct, and also against

such survivors as the elephant, bison, horse, and boar. Hares, swans, and geese were also hunted, and salmon, trout, and pike were fished. Man was, inevitably, a successful competitor for land use. By prolonged effort he won the land for crops and for domesticated animals. And he hunted animals, especially for furs. As population mounted in industrializing Europe, he no less inevitably destroyed, or changed drastically, the wild vegetation cover and the animal associates it housed. With difficulty, and largely on human sufferance, animals have nevertheless survived in association with contemporary vegetation zones.

The tundra. In the tundra some reindeer (caribou), both wild and domesticated, are well equipped to withstand the cold. Their spoon-shaped hooves are useful in finding food in rough ground. Their herds migrate southward in winter and eat lichens and other plants, as well as flesh, notably that of lemmings and voles. Dogs, too, are reared for traction but yield less than reindeer, which also provide meat, milk, pelts, wool, and ivory. The Arctic fox, bear, ermine, partridge, and the snowy owl may appear in the tundra, where, in the short summer, seabirds, river fish, and immigrant birds (swans, ducks, and snipes) vitalize a harsh environment then made almost intolerable by the swarms of midges.

The thinning out of animal species in boreal forests

Boreal associations. In the boreal forests the richness of animal, including bird, life, which persisted during historical times, is now greatly reduced. Among large surviving ungulates are the elk (moose), reindeer, and roebuck, and, among big beasts of prey, the large brown bear. The lynx has been exterminated, but not the wolf, fox, marten, badger, polecat, and white weasel. The sable, much hunted for its fur, only just survives in the northeastern forests of the European Soviet Union. Rodents in the forests include squirrels, the white Arctic hare, and (in the mixed forests) the gray hare and the beaver. Among birds are the black game, snipe, hazel hen, white partridge, woodpecker, and the crossbill, all of which assume protective colouring and are specially adapted to be able to find their food in a woodland environment. Owls, blackbirds, tomtits, and bullfinches may be seen in the forests, and, in meadow areas, geese, ducks, and lapwings may be seen.

The steppes. The fauna of the steppe zones now lacks large animals, and the saiga antelope has disappeared. Numerous rodents, including the marmot, jerboa, hamster, and field mouse, have increased in numbers to become pests, now that nearly all the steppe is under cultivation. Equally plentiful birds include the bustard—who can fly as well as run—quail, gray partridge, and lark. These take on yellowish-grey or brown protective colouring to match the dried-up grass. Eagles, falcons, hawks, and kites comprise the birds of prey; water and marsh birds—especially the crane, bittern and heron—also make their homes in the steppes. Different kinds of grasshopper (cicadas) and beetles are insect pests.

Mediterranean and semidesert associations. In Mediterranean Europe, remnants of mountain woodland harbour wild goats, wild sheep—such as the small mouflon of Corsica and Sardinia—the wildcat, and wild boar. Snakes, including vipers, and lizards and turtles are familiar reptiles, but birds are few. Finally, the faunas of the semidesert areas of the southeast European Soviet Union show affiliations with the grass steppe and the desert between which they lie. Two types of an antelope (saiga and jaran) survive there; as do rodent sand marmots and desert jerboa, and, as a beast of prey, the sand badger. There are also many reptiles—lizards, snakes (cobras and steppe boas)—and tortoises. The saksaul jay and the saksaul sparrow, named after the desert tree, also live there, while scorpions, the karakurt spiders, and the palangid, are insects dangerous to man and camel.

Conservation problems. Pressure on space, hunting, either for sport or to protect crops, the pollution of seawaters and freshwaters, and the contamination of cropland have so reduced many animal species by the closing decades of the 20th century that strong efforts are now being made to preserve those threatened with extinction, in such refuges where they still, precariously, live.

Nature reserves

Nature reserves have been set up in many European countries, with international support from the International Union for the Conservation of Nature and Natural Resources (IUCN) and the World Wildlife Trust (WWT). Seabirds find safe homes, for example, in the Lofoten Islands of Norway, and Farne Islands of northeast England. The snowy owl, which feeds on lemmings, is seen in Lapland, the rare great bustard in the Austrian Burgenland, and the musk-ox in Svalbard. Père David's deer, which had become extinct in China, its native home, was introduced in 1898 at Woburn Abbey, England, where it now flourishes. Nearly half the bird species of Europe, including the egret and the imperial eagle, are represented in the Coto Doñana Nature Reserve, within a setting of wild vegetation in the *marismas* of the Guadalquivir Estuary (Spain); there too the Spanish lynx survives. In Poland the extensive Bialowieza National Park, a wild forest once hunted yearly by the tsars, contains deer, wild boar, elk (moose), bears, lynx, wolves, eagle owls, black storks, the European bison (wisent), and the tarpan, a gray-coloured horse and a survivor from remote days. In Belorussia, in the Belovezhskaya Pushcha, bison are preserved. Italy has its famous Gran Paradiso National Park in the Aosta Valley, which preserved from extinction the Alpine ibex; Austria has a bird refuge in the Neusiedler See, which is the only breeding site of white egrets in western Europe; and the huge and magnificent delta of the Danube is largely left to wildlife. The golden eagle, Alpine marmots, and chamois are to be seen in the German Alps near Berchtesgaden.

Other rare birds are the sandwich tern (at Norderoog Island, western Germany) and the spoonbill and cormorant, found respectively at Texel Island and Norderoog Island, the former in The Netherlands. For ornithologists (as for botanists) Iceland has abounding interest, notably at Stutness, an island in the shallow lava Lake Mývatn. The beautiful wild horses of Réserve Naturelle de Camargue (Rhône Delta), the wild ponies of the New Forest (England), and the Barbary apes, maintaining a politically indifferent foothold on the Rock of Gibraltar, continue undiminished in popular interest.

Thus the European environment, once not so unequally shared by plants, animals, and man, has, with the march of civilization, been subjected to man's attempt at mastery. Favoured by their proximity to the Near East, where crop cultivation and animal domestication first began, the peoples of Europe have fashioned cultural landscapes, at the expense of wild nature, to serve their economic and social ends. Only with difficulty—and sporadically—has wild nature survived, and only just in time has awareness of the cultural losses from the impoverishment of natural vegetation and its animal associates underlined the urgent need for careful protection and preservation of nonhuman nature for communal enjoyment and scientific research.

The war against predators

Against certain predators, notably the rabbit and the anopheles mosquito, war has been waged with good effect, for malaria no longer afflicts the Mediterranean world, and rabbits, competitors for grass, have been greatly reduced and might perhaps be reduced again by the spreading of myxomatosis. On the credit side, too, should be listed the full use made of domesticated animals for pastoral husbandry—on high and rough ground, as well as on farms. The familiar farm animals are selectively bred and reared with some regard to the physical character of their environment as also to market demands and government decisions. In the far north, herds of reindeer are adapted to withstand cold and to find their food in snow-covered ground. In the rough, hilly, scrubland of Mediterranean Europe, the sheep, goat, donkey, mule, and ass have adapted well. The horse, which, in its long history, has drawn chariots, carried mounted knights, hauled the plow, wagons, artillery, stagecoaches, canalboats, and urban trams, is now largely replaced by the tractor, truck, and jeep. He is now reared more for racing, riding, ceremonials, and the hunting of fox and stag, but is still used for farm work, especially in eastern Europe. Distribution maps of animals kept on farms show how widely they enter into

farming: sheep have a special concentration in Great Britain and the Balkan countries, cattle have small place in southern Europe, while pigs are relatively numerous in the north, especially in the highly populated areas of West Germany, Denmark, and the Low Countries.

IV. Natural and human resources

MINERAL RESOURCES

With rocks and structures of virtually all geological periods, Europe possesses a wide variety of useful minerals. Some, exploited since the Bronze Age, are depleted; others have been greatly consumed since the Industrial Revolution. Useful minerals include those that provide energy, ferrous and nonferrous metals and ferroalloys, and those that furnish materials to the chemical and building industries. Europe has a long and commendable prospecting tradition, but, as in the case of North Sea gas, some surprises are still encountered. In relation to the ever-mounting requirements of its economy, however, Europe—the Soviet Union apart—is very dependent on mineral imports.

Recent mineral discoveries

Mineral fuels. *Coal.* Europe commands abundant resources of hard and soft coal, which remains of considerable, if declining, importance as a source of power, in smelting minerals, and for its many by-products. Only exceptionally does northern Europe have coal measures of commercial scale, but Carboniferous coal measures are preserved in Hercynian basins throughout the continent, lying diagonally across Britain, Belgium, The Netherlands, France (especially Lorraine), Rhine-Westphalia and the Saar (German Federal Republic), Upper Silesia (mainly in Poland but also in Czechoslovakia), to the Donets Basin and Urals. There are numerous fields, small but often, as at Pécs and Komló, in Hungary, and in the Arctic fields of Soviet Vorkuta and Norwegian Svalbard, of great locational value. Some, as in southwestern Scotland and southern Belgium, have been either worked out or become uneconomic. Deeper workable seams are sought, in the Ruhr, for example, and, undersea, off Yorkshire, England. Major reserves, encompassing mostly hard deposits of coking, anthracite, and steam coal, lie in the Ruhr, the Pennine fields of England, Upper Silesia, and the Donets Basin. Softer and less calorific brown coal, or lignite, occurs in the two Germanies, the Chomutov fields of Bohemia, and the Moscow-Tula field.

Petroleum and natural gas. Known petroleum and natural-gas reserves are, except in the Soviet Union, Romania, and Albania, wholly disproportionate to Europe's large requirements. The Volga-Ural field dwarfs all others; Romania's reserves from the Carpathian and sub-Carpathian zones, although very much smaller, meet its needs. Many west European countries have located and are exploiting quite modest reserves of petroleum. At Groningen, The Netherlands have a rich source of natural gas, and the United Kingdom has gas and oil beneath the North Sea bed.

Uranium. Sources of uranium for use in nuclear reactors have been discovered in many European countries, including France (centred on the Massif Central); the Soviet Union (in the Ukraine and Estonia); and lesser amounts in parts of central and eastern Europe.

Metallic deposits. *Iron ores.* The largest known iron reserves are the Soviet deposits at Krivoy Rog in the Ukraine Shield, Magnitogorsk, and near Kursk. High-quality ores (of 60 percent iron) from the first two sources are becoming expensive to mine, but Soviet reserves are more than sufficient for the country and its European associates: the Kursk Magnetic Anomaly, located in the south of the Soviet Union proper, has an estimated 15,000,000,000 tons of iron rich quartzites. Deposits in other European countries are small and, except in France and Sweden, inadequate for large-scale heavy industry.

The Kursk Magnetic Anomaly

Ferroalloy metals. The richest ferroalloy deposits occur in the emerged shield rocks of Kola (titanium and molybdenum), the Urals, and the Ukraine. Nickel is also mined at Pechenga, in Kola, and at several Ural sites. The southern Urals have also deposits of manganese, required for basic steel manufacture, but these are dwarfed by that at Nikopol, near the Krivoy Rog iron field, which is the largest and best located in the world. Other countries have virtually no significant nickel or tin reserves and only small manganese resources. There are chromium deposits of some scale in the Soviet Orsk-Khalilevo region and in Macedonian Yugoslavia, which also contains antimony and molybdenum. Wolframite (for tungsten) is mined from Iberian Hercynian rocks. Norway has molybdenum workings, and Finland has deposits that yield titanium, vanadium and cobalt—valuable and scarce alloys for special steels—and ranks as Europe's largest producer outside the Soviet Union.

Nonferrous base metals. With notable exceptions, known European reserves are small, partly as a result of the depletion, for example, of Cornish tin and Swedish copper. Deposits yielding copper, often from copper pyrites, are found in Scandinavia, the southern Urals, and Mediterranean lands. At Bor, Yugoslavia has the largest reserve of copper (low grade) in Europe; its reserves in lead and, especially, zinc are also high. Mercury is obtained near Krivoy Rog, in Yugoslavia, and in southern Spain. Europe has much bauxite, the principal basis for aluminum, with Hungary alone possessing some 16 percent of world reserves. Nepheline, an alternative raw material for aluminum, is worked in Kola.

Bauxite reserves

Precious metals. Europe's once widely available reserves of gold appear largely exhausted. A little silver is still produced, mainly in north European countries, but only in the southern Urals is gold found on any scale.

Nonmetallic deposits. Minerals within this large category are widely available. Clay minerals include fuller's earth—used to cleanse wool and to finish cloth—derived from the decomposition of feldspar; kaolinite, of similar origin, and valuable as china clay, occurs in a pure form in southwestern England: it surpasses coal as a United Kingdom export; rock salt, important in the chemical industry, occurs widely, much of it being precipitated in such geologically ancient salt lakes as the Soviet Lake Baskunchak (lower Volga), which contains 130-feet-thick strata. Europe also has substantial sulfur deposits, and the mining of sulfur in Miocene beds in Sicily gave Italy a virtual monopoly before the opening up of New World deposits in Texas. The carbonate rock, dolomite, like talc, is used as a refractory material, as in lining metal furnaces, and is widespread. Graphite, a crystalline form of carbon used as a lubricant and the basis (with clay) of the "lead" in pencils, is worked in Norway, Finland, and England. Nitrates, for fertilizers and explosives, are made from the air electrolytically in England, Norway, and the Soviet Union, and deposits generating potash and phosphate fertilizers are relatively abundant. The Soviet apatite (calcium phosphate) deposits at Kola are the world's greatest, as are its potash deposits at Solikamsk, in the Urals. Corundum, a very hard abrasive, occurs widely. Building materials for cement and bricks, as well as stone, are abundant, although only regionally available, depending on geological structure. Particular building stones—marble from central Italy, granite from Norway and Scotland—have localized sources. Except in the Urals, precious stones are rare: these mountains also contain the chief European deposit of asbestos.

Building materials

WATER RESOURCES

Surface water. The mountainous and upland areas of Europe collect great amounts of surface water, which supply the rivers and lakes; the lowlands, with lower rainfall, thus receive "exotic" water from the higher portions of their river basins. In the Mediterranean lands, surface water is minimal in summer, exceptions being northern and northwestern Iberia (which gets ample rain), the north Italian plain, which has Alpine rivers, lakes, springs, and summer rain; and the Apennine zone of Italy. In the European Soviet Union, surface water is relatively abundant in the centre and north but decreases south and southeastward, where, fortunately, rivers, collecting water from extensive basins, bring in supplies, and the Volga and Dnepr dams have created enormous lake reservoirs.

The increasing water requirements of thermal power stations and industry and, to a lesser extent, domestic

needs make those little populated and industrialized European highlands, which offer surplus water, indispensable to the lowlands. The pollution of water by effluents containing nonoxidizable detergents from the cities and by those from oil refineries, chemical, and metallurgical plants has reached such proportions in, for example, the Rhine below Basel, the Ruhr region, and Lakes Geneva and Garda as to present serious problems and to incur high reclamation costs. In reaction to water shortages, water is, as in the Thames, recycled many times, a practice that, happily, improves river-water quality.

Groundwater. Europe is relatively well watered, for the water table is normally not far below the surface in the lowlands, and wells and springs are widely available there; underground supplies (groundwater), held particularly in porous rocks, are sporadically utilized by pumping. A growing trend is, artificially, to add to supplies of groundwater and thus integrate surface and underground water; nearly half of Sweden's urban water requirements are thus supplied. High capital costs, rather than lack of water, leave some areas—notably the southeastern European Soviet Union and parts of interior Spain and Turkey—in an arid state. The needs of major European cities and industrial regions involve continuing efforts to collect enough water by impounding surface water, by pumping groundwater, and by encouraging the economy, reuse, and reclamation of water.

Recycling of water

BIOLOGICAL RESOURCES

Some reference to plant, animal, and human resources is needed to complement any discussion of European resources. Reference has already been made (see above *Plant and animal life*) to what remains of Europe's plant and animal heritage, supplemented as this is by such vigorous developments as the breeding of livestock to specific purposes and the acclimatization of trees and plants of economic value, which have taken place throughout its history. The following sections (see below, *Human resources; Resource development; Demographic patterns*) turn attention to the development of the natural heritage and to humans themselves, the creators of wealth and the disposers of animate nature.

HUMAN RESOURCES

Origins of the Europeans. The highly skilled peoples of Europe constitute its principal resource. While they can be distinguished clearly on the basis of nationality, language, religion, standards of living, and political ideology and organization and are of predominantly white skin colour, they can claim no clear-cut racial origin. The continent had a scanty population of human types, now extinct, before *Homo sapiens* appeared some 40,000 years ago, and throughout prehistory it received continual immigrants from Asia. Scientific investigation has thrown little light on the race or races from which European peoples are derived, if "race" is taken strictly in its biological sense. Efforts have been made to characterize their different "ethnic types," but this refers only to selectively defined physical traits of the present population and, at best, has a certain descriptive and statistical value. Fortunately, much more is known about nationality and language, which have had immense social and political import in the continent.

Race and ethnic types

The coming into the continent from the east of peoples, Aryan, or Indo-European, in language, and Bronze Age in culture, and their widespread settlement provide a clear vantage point for the discussion of European origins. Caucasian (white skinned) in ethnic type, they were semisedentary farmers. They moved along such routeways as the loess-covered country, which they could clear of light woodland by fire, and settled in clearings or on the edges of forests.

Linguistic patterns. Most of the European languages of today—notably those that fall into the Germanic, Romance, and Slavic language families—derive from the speech of these Indo-European immigrants. Thus English, German, Netherlandish, (Dutch and Flemish), the Scandinavian languages, Icelandic, and various dialects such as Swiss-German derive from German, originally spoken in, and spread from, the Baltic region. The Romance languages—French, Spanish, Catalan, Portuguese, Italian, and Romanian—are regional derivations from the Latin speech of Roman conquerors, the western portions of whose empire lasted more than 400 years. The great Slavic language family bequeathed numerous national tongues, notably Russian, Ukrainian, Belorussian (White Russian), Polish, Czech, Slovak, Serbo-Croatian, and Bulgarian. Finally, Greek, Albanian, Latvian, Lithuanian, and Celtic are additional Indo-European languages, while the non-Indo-European languages represented in Europe are Finno-Ugrian (to which Finnish, Estonian, Karelian, Lapp, and Magyar belong), and Turkic (Turkish and some groups in the European Soviet Union), Semitic (spoken by Maltese), and the mysterious Basque.

The geographical patterns of Europe's language map were simplified by both voluntary and forced migrations during and after World War II. In general, the national languages follow compact patterns, but such old surviving languages as Basque, Wendish, Welsh, and Breton are of exceptional interest to the linguist.

Linguistic survivals

Political groupings. *The historical legacy.* Europe is unique in that it has given the world so many national languages and literatures; the continent also holds a multiplicity of self-conscious and clearly recognizable national groups and states. National groups have long existed, but only by stages have European nations become self-conscious and highly emotive about their identities and used them vigorously and articulately as the basis of claims to separate statehood. The United States War of Independence (1775–81) and the French Revolution of 1789 created, dramatically, the modern concept of the nation state, although the English, the Scots, the French, and some others had created nation states even in the Middle Ages. Multinational empires—foundering only in the 20th century, with the collapse of the German, Austro-Hungarian, Russian, Ottoman, and, in particular respect to Ireland, the British empires—were also important. Even in the early 1970s, when so many states had acquired independence all over the world, Europe had over 30 independent states (Gibraltar alone remaining formally a self-governing colony), or nearly a quarter of the world total.

Contemporary trends. The political map of contemporary Europe appears fragmented when compared with that of 1815, or even that of 1913. The two great wars waged in Europe, each followed by territorial resettlements, saw the great empires replaced by many small national states, some of which, like Czechoslovakia and Yugoslavia, were nationally composite. The Russian Empire was replaced by the Soviet Union, the territories of which were at first sharply curtailed in the west, a situation remedied during and after World War II. The 1919–39 period saw western Europe separated from the new Soviet state by a zone of weak national states oriented to the West; but the post-World War II developments saw their replacement by a new belt of Socialist nation states. The ensuing cold war era saw great importance attached, by the British leader Winston Churchill and others, to the sharp linear boundary, publicized by Churchill as the iron curtain, dividing countries associated closely with the Soviet Union from those in the West where United States influence became important. Other legacies of World War II were a politically divided Germany, shorn of its former eastern territories, with part of Berlin a Western enclave in the East.

This political fragmentation of Europe, with all the attendant disadvantages from a continental viewpoint, has been qualified by the growth of the concept of peaceful coexistence and by a trend toward the creation of international organizations with a measure of sovereignty. These associations have come about through common political and defensive interest, by the clear advantages of economies of scale in industrial production, and perhaps, too, by a genuine attempt to bury a warlike past beneath more civilized ideals.

The political fragmentation of Europe

The chief of these associations of states are the Western-oriented Organization for Economic Cooperation

and Development (OECD), which also counts the United States, Canada, and Japan as members; the North Atlantic Treaty Organization (NATO) and the Warsaw Pact, both military and defense alliances; the European Free Trade Association (EFTA), a trade grouping; the European Economic Community (EEC), popularly known as the Common Market, and the Council for Mutual Economic Assistance (CMEA, or Comecon), political and economic groupings of western and eastern Europe respectively.

Intracontinental and intercontinental movements. Although Europe's population remains large absolutely, its share of mankind has decreased to approximately one-sixth. This results not only from demographic changes (see below *Demographic patterns*) but also from emigration, since Europe has sent its people in great numbers to the other continents, notably to the Americas, Australasia, and northern Asia. This outflow, originating in the age of maritime discovery, assumed massive proportions when the steamship and railway facilitated passage overseas. British and Irish, French, Spaniards, and Portuguese, in particular, moved in such numbers and with such political effect that their languages are now spoken over extensive areas beyond Europe.

Within the continent there has always been some mobility of population, high during prehistoric times, and well-marked during the centuries that witnessed the decline and fall of the Roman Empire in the West, when many tribal groups—especially of Germans and Slavs—settled in specific regions where they grew into distinctive nations. During and after World War II many Germans from outlying settlements in central and east central Europe returned to western Germany, some as forced migrants. Many East Germans, too, made their way to the West until the building of the Berlin Wall checked this flow. More recently, and where this is permitted, migrants are chiefly workers, seeking temporary work and, often with less success, new homes. The Common Market countries draw workers from southern Europe, as does Switzerland. The United Kingdom has received sizable numbers of Commonwealth citizens, including West Indians, Indians, Pakistanis, Kenyan and Ugandan Indians, as well as Irish and some aliens. Two other conspicuous forms of mobility in Europe are the daily commuting of city workers and the increasing movements of tourists (see below *Demographic patterns: Extracontinental movements*).

Recent migrations

The European contribution. Human enterprise and labour create wealth, and, in this respect, Europe has outstanding assets. Its early start in civilization, with which only Asia compares, the creative abilities of its peoples, who, despite their apparent diversity, share common cultural sources, their traditional fertility, and thus their substantial numbers (some 640,000,000 by the 1970s)—all these have contributed to Europe's success in exploiting the resources provided by a richly variegated environment. These successes relate to agriculture, mining, and forestry, as well as to manufacturing industry, commerce, transport, and finance. They relate also to science and technology, to literature and the other arts. The population of Europe is nevertheless very unevenly distributed, and the same standards of living and degrees of urbanization are not equally achieved in all parts. There are lowland and plateau areas, where, associated with mining, modern industry, and urbanization, the population is very densely settled. In contrast, areas of really scanty settlement lie mainly in northern Europe and, elsewhere, mostly in mountainous and arid zones.

V. Resource development

The resources of Europe, since they result from the efforts applied at an ever-rising technical level, are in some respects inexhaustible. Although the cultivation of soil and mining and quarrying for metallic minerals were initiated in prehistoric times, the winning of some resources began only in relatively recent times, in response to new needs and new technology. The clearing of woodlands for the plow has continued since the early Middle Ages; the cultivation of the steppe lowlands of the Ukraine and the Lower Danube Basin commenced only in the late 18th century. Effective drainage, in which the Dutch excelled, especially during and following the 17th century, made use of former marshlands. The large-scale mining of coal and iron ore date from the Industrial Revolution. Some industries—many concerned with products of the chemical industry and the refining of aluminium—belong to the 20th century, which also brought electricity as a form of energy and, for transport, the petrol-fuelled engine for use on land, sea, and in the air. Recent international groupings facilitate large-scale economic units, and the value to Europe of both a variety of Socialist planning and the post-World War II United States financial aid channelled through the Marshall Plan should also be noted.

The concept of stage, too, helps an understanding of Europe's economic development, for the application to industry and agriculture of modern technology and scientific research has reached different parts of the continent successively. Great Britain, as the home of the Industrial Revolution, stimulated economic change in western, central, and northern Europe. The Soviet Union and its present east European associates were mostly late starters, and the pace and scale of their industrialization quickened markedly after 1945 on Socialist bases. The countries of southern Europe, notably northern Italy, also advanced economically following World War II, although much progress remains to be made. Europe is thus an outstandingly well-developed part of the world, and, although economic development is uneven regionally, no clear limits appear to impede further progress.

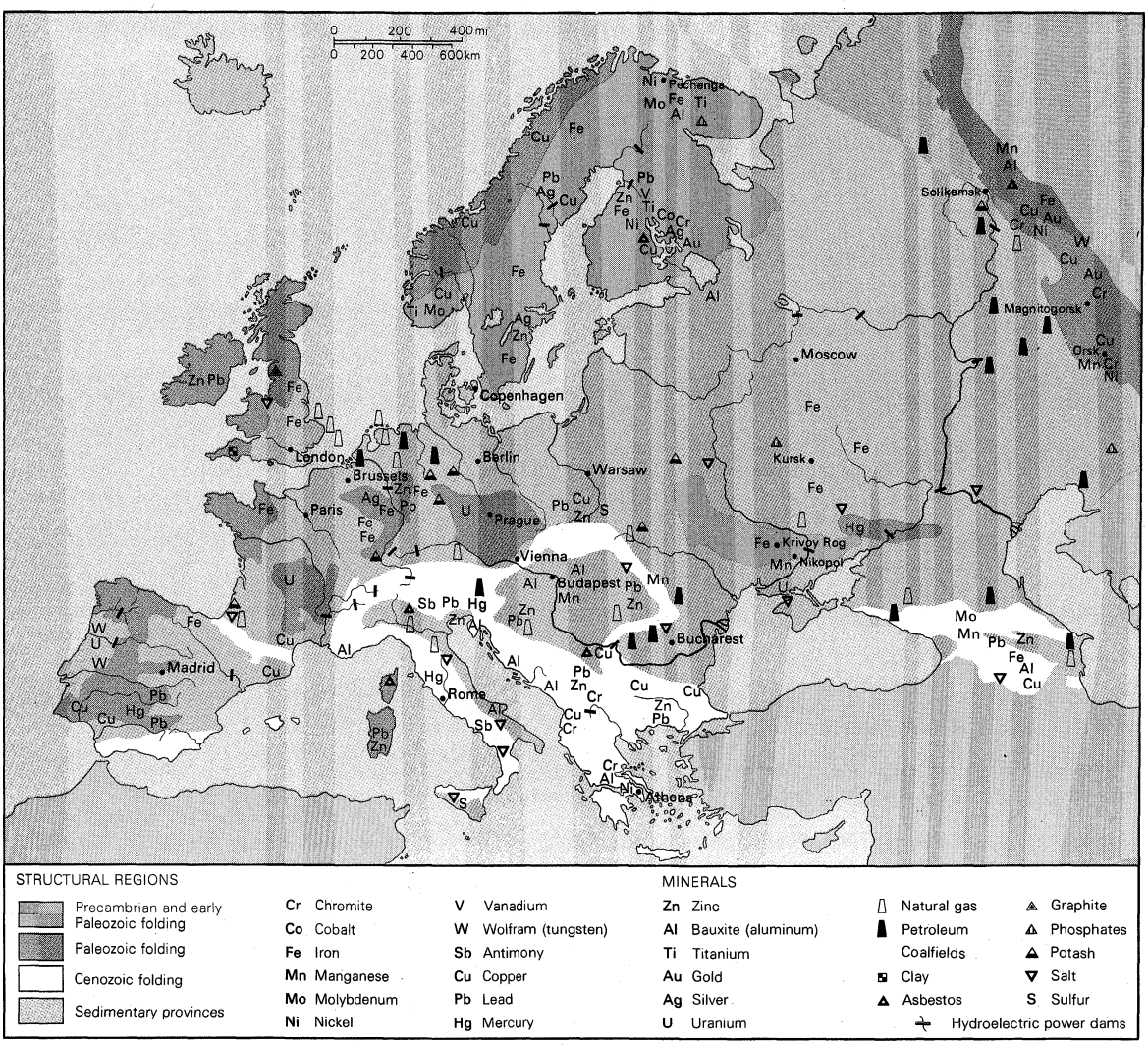
Economic growth stages

EUROPEAN INDUSTRY

Mining. Mining provides employment in all countries, although for smaller numbers as mechanization is applied. High-grade iron ores are mined at Krivoy Rog and near Kursk, Magnitogorsk, and in Arctic Sweden and Norway; these are supplemented by the low-grade minette ores of Lorraine and Luxembourg, low-grade (quarried) Jurassic ores of England, and low-grade Spanish ores. Europe, including the Soviet Union, accounts for nearly a half of the world's coal production and over three-quarters of its lignite, although coal output is shrinking as other forms of energy, especially oil, natural gas, and hydroelectricity become increasingly economic. The chief producers are Great Britain, West Germany (the Ruhr and Saar), Poland (Upper Silesia), and the Soviet Union, where the Donets Basin yields a third of the national output. East Germany is the world's chief source of lignite, mined also in West Germany, Czechoslovakia, and the central European Soviet Union. Many mineral deposits are of only local interest but, as a whole, Europe produces about a quarter of the world's bauxite, a fifth of its copper, and a third of its lead and zinc. Minerals of more than domestic importance are natural gas in The Netherlands; bauxite in France, the Soviet Union and Hungary; petroleum, apatites, and manganese from the Volga-Urals, Kola, and the Ukraine respectively; and china clay from England.

Heavy industry and engineering. The change from charcoal to coke as fuel in blast furnaces led to the localization of Europe's iron and steel industries on its coalfields to economize transport costs, although imported iron ore, cheap American coal, electric furnaces, and technological efficiency have loosened this tie. Thus Northumberland and Durham in England, Rhine-Westphalia, Upper Silesia, and the Donets Basin have their coalfield furnaces and mills, while others are grouped near sources of the ore, as at Krivoy Rog and in Lorraine, or such convenient estuary or port sites as Port Talbot (south Wales), Genoa, and Dunkirk. Europe produces a half of the world's steel, with the Common Market and the Soviet Union each accounting for about two-fifths of the output, but only two-fifths of its iron ore. Steel-using industries making heavy machine tools, and mining, smelting, construction, and electrical equipment favour coalfield locations, while shipbuilding, motor-vehicle, and aircraft construction show a wider distribution, including new sites.

Coalfield-based industry



Basic structural regions and principal mineral and hydroelectric sites of Europe.

Chemical industries. Covering many new and old products, chemical industries have expanded greatly since 1945, partly in relation to hydroelectricity generation and partly as a result of the market-oriented use of refinery by-products. Many heavy chemicals are produced on the coalfields, notably in the Ruhr, where by-products of coke ovens and metallurgical plants are available. Other chemical industries utilize Europe's deposits of salt, potash, phosphates, and sulfur, and the industry has been revolutionized by the increasing production of synthetic rubber, plastics, synthetic fibres, detergents, insecticides, and fertilizers.

Manufacturing, timber, and fisheries. A wide range of light consumer industries is found throughout Europe, but some countries acquire reputations for speciality goods, as in the case of English and Dutch bicycles, Swedish and Finnish glass, Parisian perfumes and fashion goods, and Swiss cheese and chocolate. The United Kingdom's once leading textile industry now concentrates on high-quality goods, including many synthetic fibres, of which, together with the German Federal Republic, France, and Italy, it is a large producer.

Specialty goods

The timber and fisheries extractive industries, now mechanized, are of considerable scale: the Soviet Union, Sweden, and Finland are leading producers of soft and hardwood, and exporters of timber, wood pulp, and newsprint. Fishing and whaling are large industries for Norway, Britain, France, Iceland, and the Soviet Union: catches yield not only human food but materials for many subsidiary industries.

Handicrafts and other industries. Of small importance in a continent where the economies of mass production involve standardized production, handicrafts nevertheless survive to serve a wide market, including that of tourists, who seek speciality goods. Knitwear and Harris tweed are produced by crofters in the Scottish islands; traditional costumes are made and displayed in many eastern European countries; and custom-made ("bespoke") tailoring for men, like dressmaking for women, survives as supplements to ready-made clothing; artistic pottery making is another active craft.

Some other European industries fall but uneasily into the preceding categories. Printing and publishing, especially in English, French, German, and Russian are substantial industries with worldwide effects, notably in the educational field. Europe is a large producer of pharmaceutical drugs and produces such world-famous beverages as the wines of the west and south, the northern beers, and not least, whisky, the status drink from Scotland. Technological and scientific researches have produced nuclear-fuelled plants generating electricity, while the outstanding growth industry of tourism, supplementing business, professional, and student travel, brings employment and foreign exchange to many, especially in the Mediterranean countries, where sunshine, beaches, scenery, and historical monuments provide exploitable capital since workers have become relatively affluent and paid holidays are the rule. Europe, with 86 percent of international tourism, is the tourist "Mecca" of the world.

Growth of tourism

POWER

The message of the Industrial Revolution was that mechanical energy harnessed to machines could so supplement human muscle and animal power as to produce revolutionary changes in the scale and pace of factory production. Contemporary Europe, covering only 8 per-

cent of the inhabited earth, and with 18 percent of its population, uses 43 percent of the world's energy. World energy consumption by the 1970s was equivalent to 0.4 tons of coal per head of population, for Europe, 2.4 tons, ranging from 5.1 tons (Sweden and the United Kingdom), 5.6 tons (Czechoslovakia and the German Democratic Republic), to 1.1 tons (Spain) and lower in Portugal, Yugoslavia, and Greece.

Coal and hydroelectric power. Coal, used to drive steam engines and, as coke, in the smelting of metals, long held the predominant position. Although the United Kingdom depends on coal to the extent of 59 percent of its energy needs, only 37 percent of the energy consumed by the other Common Market countries is derived from coal and lignite. Coal continues to provide energy to coalfield-based industries and is still of prime importance for the production of electricity.

Hydroelectricity has been markedly developed where precipitation and landforms provide good opportunities to dam rivers, as in northern and Alpine Europe and the Soviet Union. Finland, for example, derives 80 percent of its power from this source; Spain, France, Bulgaria, and Yugoslavia derive a large fraction. France has developed power-consuming industries, such as aluminum refining, close to the Alpine and Pyrenean generating sites.

Other power sources. In other countries, hydroelectricity contributes very little, and petroleum and natural gas claim an increasing share of the energy consumed. By 1970 petroleum accounted for more than half of the Common Market energy requirements and natural gas for 5 percent or more. Natural gas has replaced coal gas in the Soviet Union, Romania, and The Netherlands, and supplements it in Great Britain. Fuel oil is widely used by diesel locomotives and electricity-generating stations, and for space heating. Geothermal heat—utilizing underground waters heated by volcanic action—is available in Iceland, while Ireland, which also lacks both coal and oil, makes efficient use of abundant peat resources. Nuclear-reactor electricity generation, promoted by the European Atomic Energy Community (Euratom) in the Common Market countries, provides, as in the Soviet Union, a small but growing source of electrical energy, scheduled to reach 15 percent of electricity production in the Euratom area by the mid-1970s.

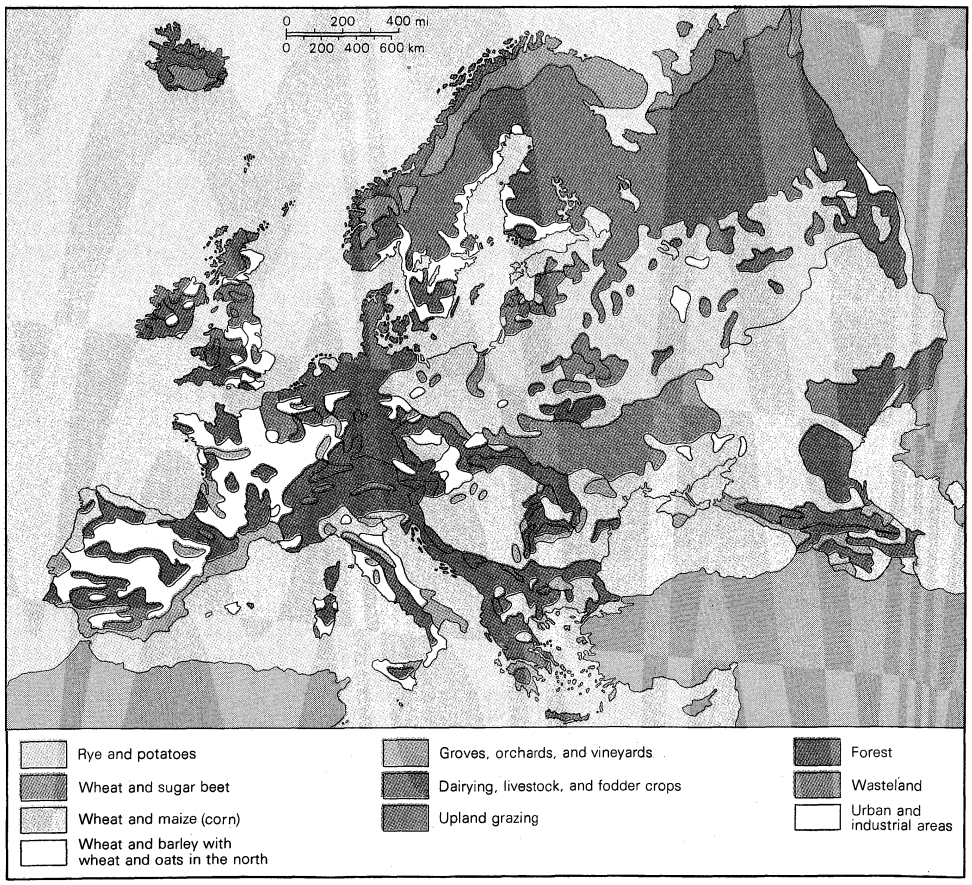
Nuclear-reactor electricity generation

AGRICULTURE AND LAND USE

Distribution. Cultivated land in Europe covers a full quarter of total area, a favourable comparison with both the United States (16 percent) and the Soviet Union (11 percent). Figures for individual countries vary sharply: from over 70 percent in Denmark and 66 percent in The Netherlands; 50 percent in Poland; 40 percent in Romania and Spain; around 30 percent in the United Kingdom, the German Federal Republic, and France; and upward of 25 percent in the European Soviet Union; to about 8 percent in Sweden; below 3 percent in Norway; and virtually nil in Iceland.

Europe's industrialization and urbanization tend to conceal the fact that it is a great producer of cereals, roots, edible oils, fibres, fruit, and livestock and livestock products, accounting for half of the world's rye and potato output, a third of the oats total, and almost a quarter of the wheat output.

Europe's climatic range helped mark out production areas; thus the vine is commercially grown south of about latitude 50° N and the olive is restricted to Mediterranean climatic regions. Maize (corn), grown mainly for silage, is an important crop in the lower Danubian lowlands and southern Soviet Union appearing also in France and Italy. Rice (in north Italy) and citrus fruits (in Spain, Sicily, and Cyprus), depend on irrigation. The northern countries grow few cereals (mainly oats) and concentrate on animal husbandry, especially cattle and dairying. Grain cultivation is found in the lowland belt stretching from eastern Britain to the Urals. Wheat, in rotation, seeks the better soils, oats and rye the poorer soils and moister lands. Mixed farming—*i.e.*, crop and animal farming—and the use of well-tried crop rotations, becomes increasingly the best economic practice.



Agricultural regions of Europe.

Viticulture, although widely distributed, is most important in France, the leading world producer, where it is as remunerative as grain cultivation. As to technical crops, the Soviet Union is the largest producer of flax and hemp, sugar beet (also grown widely elsewhere as a rotation crop in the better soils), sunflower seeds (for edible oil), and tobacco, which figures importantly also in Bulgaria and Macedonian Greece.

Agricultural organization. There are sharp differences in the way that European agriculture is organized and in its regional efficiency. Collective and state farming is the pattern in the Soviet Union and most countries of the Socialist bloc, but cooperative systems, with or without individual land ownership, prevail elsewhere, with the consolidation of over-small holdings progressing steadily in western Europe, paralleled by the breakup of large estates in the east. The capital-intensive agriculture of The Netherlands and Britain, where yields per acre and per man are high, differs from the extensive Soviet system, where, despite the benefits—notably mechanization—brought by collectivization, labour costs are high and yields low. Less than 2 percent of the labour force of the United Kingdom is engaged in agriculture, 13.4 percent in West Germany, 16.5 percent in France in 1967, over 30 percent in Ireland, Spain, and Hungary, over 40 percent in Bulgaria, Greece, and Poland, and 57 percent in Yugoslavia, the higher figures indicating high densities of rural population, a shortage of investment capital, and persistent underemployment.

The relative use of fertilizers—as high as 518 pounds per acre yearly in The Netherlands, and as low as 63 in Poland and 32 in Spain—hints also at the range of crop productivity.

Irrigated areas, lying mainly in southeast Spain, the North Italian Plain, and Mediterranean France, are very small, but disproportionately productive. Long-term prospects for utilizing the irrigation capacity of the Lower Danube and Volga are good.

Livestock farming and dairying associated with pigs and poultry is characteristic of European farms, except in the Mediterranean lands, which are better adapted to sheep and goats. Europe produces well over a third of the world's meat, chiefly beef, pork, and bacon, but this is insufficient to meet rising living standards. Domestic production of wool, hides, and leather is also insufficient, except in Spain and the Soviet Union, which derives much wool (and cotton) from its Asiatic territories. Special features of west European farming include market gardens and glasshouse production of tomatoes, cucumbers, green vegetables, and flowers for urban markets, and production of *primeurs:* table fruit, new potatoes, vegetables, salad crops, and flowers, produced when prices are high, and thanks to the early spring that visits coastal Brittany, Cornwall, and the south of France.

Specialty crops of west Europe

Despite advances in agronomic science, the hazards of harvest shortfalls due to climate have not been eliminated, and intermittent emergency, as well as regular, claims have to be made on grain-surplus areas overseas.

VI. Commerce

Internal and external trade, both by land and by sea, has always been a vigorous part of Europe's economy, and no less so in the 1970s when it faces such strong competitors as the United States and Japan. Trade is made necessary by the regional specialization of production, largely initiated by capitalist enterprise in the past, and now predominantly guided by national or even supranational policy decisions. Economic geology partially explains, for example, the localization of Swedish iron, and Soviet petroleum production, while climate localizes olive-oil and citrus-fruits production. Europe acquired a central position in modern times in the well-settled Northern Hemisphere, which oceanic and air transport systems still exploit. Simultaneously providing large managerial, market, and labour-force attractions, Europe inevitably attracts extra-European traders, with its ever more sophisticated industry producing outstanding exports and its large importation of petroleum, metals, raw materials, and foodstuffs.

Within the continent, the distinction should be drawn between western Europe, which depends greatly on foreign trade, and a Socialist sector that chooses to apply the doctrine of economic self-sufficiency with more inter-regional than international trade. Indeed, because of its territorial extent and the range of its resources, the Socialist sector can do so.

TRANSPORT

All forms of transport are found in Europe. Although they are not equally well-developed everywhere, the traveller can make his way to settled places throughout the continent. The European Soviet Union is, compared with western Europe, where motorways provide fast movement for business and tourism, relatively deficient in engineered motor roads. Motorable roads become ever more widely available; those of Spain and Ireland have particularly improved, and road tunnels now supplement railway tunnels beneath the Alpine passes. Animal transport has minimal importance yet survives locally: the horse-drawn cart may still be seen in east-central Europe; and the ox-drawn plow and the loaded ass, mule, and donkey—sure-footed in rough, hilly country—are still used in parts of southern Europe. And the dog- or reindeer-drawn sled has little competition for short journeys in such winter snow-covered regions as the Soviet North.

The role of animals in transport

Railways. Railways link European ports with their hinterlands and fan out from capitals and major cities to points on the international frontiers where they meet the railway system of their neighbours. In some cases—notably from France to Spain, and the Soviet Union to Poland—this involves a change of gauge. Underground and suburban railways also play an indispensable role for metropolitan commuters. Railways permit passage between the western and eastern European extremities but not quite to the extreme north; they have lost some of their passengers and freight to automobile, coach, and road hauliers, and many uneconomic local lines have been closed. Even so, rail services have notably improved with the use of electrified track or diesel locomotives, faster intercity passenger trains, and container freight trains; and Soviet railways remain all-important.

Waterways and pipelines. Seaports have been modernized and enlarged to deal more efficiently with the increasing size of ships and volume of oceanic trade; even landlocked Switzerland has seagoing ships utilizing Dutch port facilities. The United Kingdom, Norway, and Greece also hold large freighter tonnages for hire.

Inland waterway transport, slow but cheap, is regionally important for the carriage of heavy and bulky commodities. The best waterways—the Rhine below Rheinfelden and the Danube below Belgrade—can carry 1,500-ton barges. The navigable Rhine has the legal status of an international waterway open to all users. Other rivers and canals are usable by smaller vessels. The Soviet Volga, however, is a valuable waterway linking Moscow with Caspian ports and, via the Volga–Don Canal, gives water access to the Donets Basin.

Barge carriers

Giant tankers, up to and beyond 300,000 gross registered tonnage and too deep in draught for most seaports, deliver their cargoes by pipelines that—for petroleum, natural gas, and water—provide the cheapest overland form of transport. They have been built in the United Kingdom for North Sea gas and imported oil; in France, Spain, and Italy for North African oil; and within and beyond the Soviet Union, whose Friendship Pipeline carries crude oil from the Volga-Ural field to east European refineries.

Airways. Air services between principal European cities and to all parts of the world are extensively organized: London, Frankfurt am Main, Paris, and Moscow stand out as airports of first importance. Passengers, mail, and other commodities—such as gold and early spring flowers—of high value in relation to their weight, make use of air transport.

INTERNAL TRADE

Within each European country, a wide variety of goods is moved continually from ports and production centres

to urban markets. Miscellaneous home-produced goods are also traded to consumer centres. Imported goods include fuels, tropical foodstuffs and drink stuffs and raw materials, textile fibres, metals and metallic ores, and a wide range of manufactured goods.

Trading blocs. Active trading within groups of countries that have associated primarily for this purpose, and to rationalize and so increase the profitability of their national economies, is advancing: European Economic Community and Comecon policies have been directed toward economic specialization in increasingly interdependent member countries. West Germany supplies coking coal and chemicals to France, which provides Belgium with iron ore from Lorraine; steel is moved to extranational markets; Dutch natural gas is piped to France, Belgium, and West Germany. Speciality foodstuffs—wines, cheeses, spring vegetables, and fruit—find an enlarged market far beyond their production centres, as do such manufactured items as fashion goods, automobiles, and washing machines. Comecon countries direct their trade largely to each other, with the Soviet Union pre-eminent as producer and market. The U.S.S.R. supplies petroleum, manganese, iron, and chrome ores as well as cotton and other textile fibres; receives machinery, textiles, and Hungarian bauxite.

Comecon and EFTA

The European Free Trade Association (EFTA) has also encouraged trade between its members, who exchange such complementary, rather than similar, products as Swedish and Finnish timber; Danish dairy products and manufactures; Portuguese tungsten, wine, cork, and canned fish; British automobiles, woollen textiles and china; and Swiss watches and food products.

The accession of Norway, Ireland, and the United Kingdom to the Common Market in January 1973 was expected to alter substantially the volume and direction of some trade patterns both within and between EFTA and Common Market blocs. This was expected to be especially true for agricultural commodities and for Commonwealth trade with Britain.

Trade between the West and a Socialist bloc eager for foreign exchange, though much smaller than its pre-World War II equivalent, has increased markedly since the mid-1960s. Soviet petroleum is sold to Italy, Denmark, and the United Kingdom, and Western markets are also used for the sale of gold and diamonds in exchange for ships, machinery, and chemicals. European Socialist countries also supply Soviet automobiles, canned salmon and caviar, and vodka, Polish bacon, Czech glass, and Hungarian and Yugoslav wines.

Intra-Soviet trade. Given its continental scale, the regional trade of the Soviet Union, carried largely by a heavily worked railway system, supplemented by pipelines and an elaborate waterway system, deserves special notice. From the south, grain, meat, vegetable oils, sugar, tobacco, wine, and fruits are moved to central and northern Soviet Union, where consumption exceeds local production; dried fish and salt are carried up the Volga; timber is moved from northern, including Ural, areas to largely unforested southern regions; and Donets Basin coal is shipped by canal to Volgograd.

EXTERNAL TRADE

A major part of the external trade of European countries is with each other since, with regional specialization, dense populations, and relatively high standards of living, they provide strong markets. For the Common Market countries, as well as for those of northern Europe, this trade proportion ranges between 60 and 75 percent, and the British figure, accommodating traditional Commonwealth relations, was still around 45 percent just prior to its admission to the Common Market. About 70 percent of Socialist-bloc foreign trade is intra-Comecon. There is, nevertheless, trading between EEC, EFTA, and Comecon members, and, especially in the United Kingdom and the German Federal Republic, a vigorous two-way trade with the United States. Last, but certainly not least, European trade extends to all parts of the world, including the developing countries, where —in exchange for manufactures—vital supplies of energy, raw materials, metals, ores, and foodstuffs are obtained.

North American trade links

The extracontinental exports of Europe include machine tools, automobiles, aircraft, chemicals (including pharmaceutical drugs), and such consumer items as clothing, textiles, books, expert services, and works of art. Western Europe depends heavily on imported petroleum from the Middle East, Algeria, and Libya, and many raw materials and metals. Europe imports much natural rubber, tea, coffee, cacao, cane sugar, vegetable oilseeds, tobacco, and fruit—fresh, canned, and dried—although its dependence on imported agricultural products, both of food and raw materials, has decreased substantially with greater domestic production and the manufacture of synthetic substitutes for vegetable fibres.

THE COMMERCIAL OUTLOOK

Europe plays the leading role in world commerce, making up about half the total of world exports and imports. The bulk of this trade is carried on by the Western countries, which own nearly half the world's oceangoing tonnage, but there is every indication that the Soviet Union intends to claim a larger share than hitherto. Most of the European countries long held political dependencies overseas where they created captive markets, and this imperialist trading momentum persists. Common Market countries have former colonial territories as associate members, and, similarly, the British Commonwealth nations engage in much trade, now strictly competitive, with Britain, although the latter's new Common Market trading associations will eventually alter many long-standing trade patterns within the Commonwealth.

The international financial crisis

One of the continuing international difficulties that Europe has faced has concerned currency and fluctuating exchange rates, as, in turn, sterling, the French franc, and the United States dollar have come under strain, while the currencies of West Germany, The Netherlands, and Switzerland have grown harder, or relatively more valuable. The survival of sterling as a reserve currency through which a third of the world's population conducts business transactions, is a tribute to its convenience and success, although the process has involved persistent financial crises for the United Kingdom. European financiers play an important world role, as do a variety of such finance-related industries as banking, insurance, and shipping.

VII. Demographic patterns

Europe has always been one of the most populous parts of the world. Although its estimated population numbered only one-third of Asia's in 1650, 1700, and 1800, this nevertheless accounted for one-fifth of mankind. Despite large-scale emigration, this proportion increased to one-quarter by 1900, when Europe's total just exceeded 400,000,000. Such high numbers, achieved by high birthrates and falling death rates, were sustained by expanding economies. As numbers have grown proportionately faster in the Americas, Asia, and Africa, Europe's population, totalling 640,000,000 in 1971, has settled at about one-sixth of the world total.

"Core areas"

In contrast to such recently settled lands as the United States, Canada, and Australia, Europe has many sharp divisions based on language, nationality, political organization, and standards of living. Given the agelong occupation of its soil and minimal mobility for the peasantry—long the bulk of the population—Europe became the home of many linguistic and national "core areas," separated by mountains, forests, and marsh. Its many states, some long established, introduce another divisive element that was augmented by modern nationalistic sentiments. Efforts to associate groups of states for specific defense and trade functions, appear, slowly, to point toward wider unitary associations, with fundamental East–West differences. Europe thus presents two clear-cut, opposing units, open to trade and cultural exchanges, and a number of such relatively neutral states as Ireland, Sweden, Austria, Switzerland, Yugoslavia, Finland, and Spain.

Qualitatively as well as quantitatively, Europe's overall

population is outstanding, with special skills, inventiveness, and initiatives that enable its peoples more than to hold their own in a highly competitive world.

POPULATION DISTRIBUTION

Overall densities. Southern Europe lost its numerical domination as, from medieval times onward, settlement developed vigorously in western and central Europe and as, later still, the steppe lands of the Ukraine and Hungary were settled for crop farming. While northern Europe, from Iceland and the Scottish Highlands to the northern European Soviet Union, is only scantily settled, the population reaches high densities in a more southerly belt, stretching from England across northern France and industrial Germany to the Moscow region, with an outlier in northern Italy. High populations are often associated with coalfields that, in the past more than today, strongly attracted industry, although giant cities like London, Paris, and Leningrad, offering large markets and labour forces, have created regions of high density. Other populous areas are sustained by mining, industry, commerce, and productive agriculture. The Netherlands, with over 1,000 persons per square mile, is the most densely populated country; and Norway and Iceland, the latter with five persons per square mile, the least dense. Population is scantiest in mountain regions, some highlands, and arid parts of Spain and the southeastern Soviet Union.

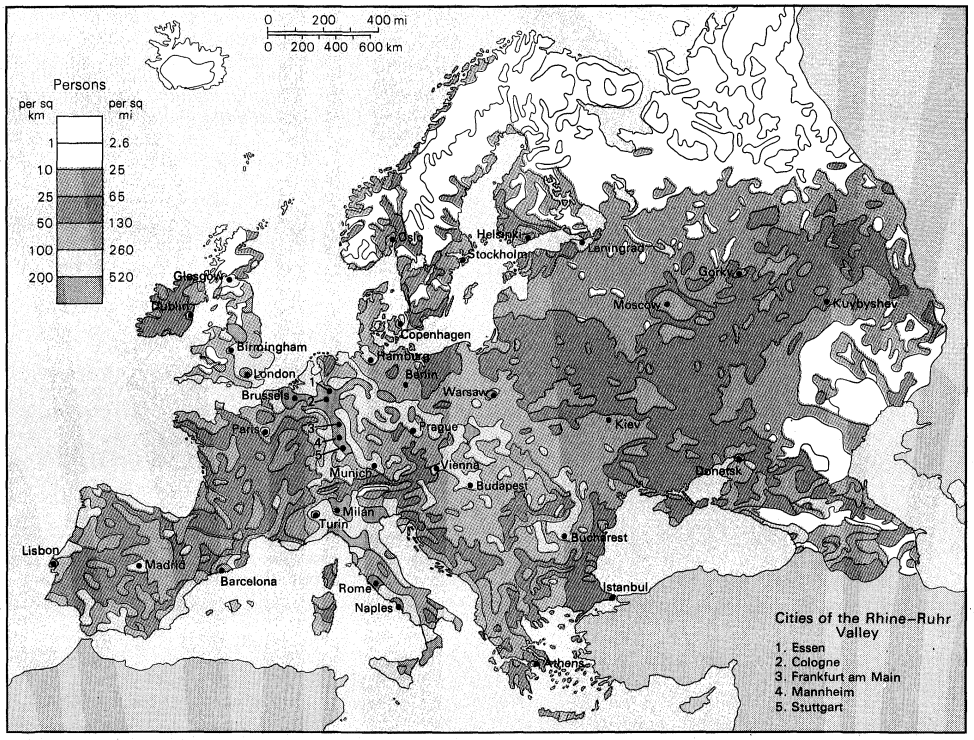
High density areas

Town and country. With easier travel and the lure of developing industrial areas, many culturally-rich, high-altitude areas have suffered severe depopulation. Urbanization, offering varied employment, better social services, and, apparently, a fuller life, has further reduced the rural population, a drift aided by the mechanization of agriculture. City life has, from classical antiquity, nurtured European culture, although tributary rural life was, for centuries, the common lot. The 19th and 20th centuries, however, have seen a revolutionary urbanization that embraces the majority of contemporary Europeans. Some towns are old, containing architectural survivals from their historic past; many more are creatures of the Industrial Revolution. In most of the highly industrialized countries the proportion of urban-dwellers is high: 77 percent in the United Kingdom and The Netherlands and 70 percent in France and Sweden. Others with figures around 50 percent, include Iceland and Norway, Belgium, Poland, and Bulgaria; the Soviet Union's figure is 56 percent and rising steadily. Towns of different scale and varying function continue to grow rapidly, usually in concentric rings outward from the original core. By the early 1970s, Europe contained almost 50 cities of 1,000,000 population, more than a quarter of the world total, and many of the more highly industrialized parts of the continent were marked by giant, sprawling metropolitan areas. One distinct type is represented by the conurbation resulting from outgrowth from London; another, as in the Ruhr, by the fusing together of separate cities. Both stem from an unchecked industrial expansion associated with population growth—including immigration from rural areas. As elsewhere in the world, these giant agglomerations present difficult social and aesthetic problems; however, by concentrating population, they help to preserve the countryside from being too built-up.

Conurbations

Population trends. Western and northern Europe took the lead in the medical and social "death controls" that, since the mid-19th century, have sharply reduced infantile mortality and lengthened life expectancy. Although infantile mortality rates remain relatively high—up to 60 per 1,000—in developing Romania, Portugal, and Yugoslavia—a low of only 13 per 1,000 has been reached in The Netherlands and Sweden. Low crude mortality rates have been achieved virtually everywhere in Europe, with the Soviet rate at only eight per 1,000.

Birthrates fell to low levels in western and northern Europe during the 1930s but recovered substantially after World War II to figures of 15 to 20 per 1,000. In southern Europe, the Iberian countries excepted, and in the Socialist countries other than Poland and Yugoslavia, birthrates now fall within this range; Sweden has the lowest rate—about 13.5—and Romania the highest —23. For the whole continent (the Soviet Union excepted), the annual increase of population is about eight per 1,000, and the comparable figure for the whole U.S.S.R. is little, if at all, higher, for despite its low mortality rate, the birthrate in the U.S.S.R. has fallen to 17 per 1,000 from 25 per 1,000 in 1959. Emigration—notably the Soviet exodus to its Asian lands—retards the European U.S.S.R.'s population growth. The fact that Europe's numbers now edge only slightly upward is one indication of the lesser role it now plays in the world and, given its already high population densities, this may perhaps be regarded as satisfactory.



Population density of Europe.

Birth rates and death rates, as they vary in time and place, necessarily affect the proportion of manpower available to the different European countries for the economy and the armed forces. In most countries, increased longevity and lowered birthrates have generated a rising proportion of retired citizens. Also, the trend toward education over longer periods draws more young people—postwar babies made up a conspicuously large such age-group in the 1960's—from the economy. The labour force thus shrinks somewhat, although it is everywhere (except in Spain) more than 40 percent of the population, and in the Soviet Union, Romania, and Bulgaria it exceeds 50 percent. The economic advantage accruing to such countries results largely from the greater employment of women, especially in agriculture.

EXTRACONTINENTAL MOVEMENTS

Despite heavy mortality resulting from continual wars, Europe has always been in modern times a generous source of emigrants. Since the geographical discoveries of the late 15th century, both "push" and "pull" factors explain an exodus greatly accelerated by modern transportation. The push factors were often sheer poverty, the desire to escape from persecution, or loss of jobs through economic change. The pull factors included new opportunities for better living, often at the expense of original inhabitants elsewhere. All of Europe shared in this huge transfer of population, which effected the settlement and economic development of the Americas, Australia, South Africa, and New Zealand, countries with untapped resources but insufficient manpower. Through their involvement in the horrors of the African slave trade, Europeans also produced forced migrations of non-white peoples that were also to have immense consequences in the Old and New Worlds. Since early in the 19th century, an estimated 60,000,000 persons left Europe itself for overseas, more than half to settle in the United States. Ireland exported most of its population following the potato famine of the 1840s. Northwest Europe—British Isles, Scandinavia, and the Low Countries—contributed the largest share of emigrants, who settled, above all, where English was spoken. Emigrants from central, eastern, and southern Europe moved later, many in the early decades of the 20th century. Affinities of languages, religion, and culture clearly explain migration patterns; South American countries, for example, had more appeal to Spanish, Portuguese, and Italians. It has been estimated that emigration from 1846 to 1932 reduced the growth rate of Europe's population by 3 per 1,000 per annum. The year 1913 marked a peak, with over 1,500,000—one-third Italian and over a quarter British—migrating overseas. Subsequent United States entry restrictions reduced this flood; contemporary European migrants seek new homes mainly in Australia, Canada, South America, and South Africa.

Despite high population densities, some European countries still attract settlers from other continents, mainly because their expanding economies involve labour shortages. Thus France, to increase a manpower depleted by war losses and low 1930s birthrates, has received numerous French-Algerians (as well as other Europeans, including Turks) to supplement its labour force. The United Kingdom, which steadily supplies immigrants to Australia and Canada and specialist workers to the United States, has also attracted immigrants, notably Commonwealth citizens: from 1931–61 immigrants from Ireland, the West Indies, Pakistan, and elsewhere exceeded emigrants by 1,000,000. These immigrants, who largely provide workers for the construction industries, transport, hospitals, and domestic service, include also doctors, scholars, and businessmen. Some, having established themselves, are able to provide homes for their immigrant relatives. Finally, there is a continued and significant migration of Russians and Ukrainians from the European to the Asiatic Soviet Union.

THE OUTLOOK FOR EUROPE

In its landscapes, many nations and tongues, in its political divisions and ideologies, in its cities and historical monuments, Europe presents regional variety and richness; yet it exhibits its own identity, based on a unity of culture, with roots in the common past of Judaism, Christianity, and Islām, ancient Greece, Rome, and Byzantium. Though the citizen of Europe does not exist, the nationals of Europe can feel at home between the Atlantic and the Urals, Stockholm and Rome. The peoples of Europe always showed remarkable fertility, energy, and originality. During their long history, they developed successfully their not ill-favoured geographical plot, created national cultures of great interest, and discovered how to achieve and sustain high living standards. Certainly the Europe of the 1970s has lost the leadership of the world to three great subcontinental powers—the United States, sprung from European settlers; the Soviet Union, itself occupying the greater part of the continent; and China, the modern successor to Europe's ancient contemporary, Cathay—whose resources and technology provide the material bases for world leadership. Europe nevertheless retains the power and spirit to withstand decline and to preserve a high relative status. Europe also shows signs of overcoming its traditional fragmentation: as the European Economic Community enlarges itself, and as the Socialist bloc becomes consolidated, a trend toward closer association points toward the distant goal of unification. In the popular mind, there remain many "Europes" and the unification of the Europe that, as conventionally delimited and as de Gaulle conceived it, stretches from the Atlantic to the Urals, remains a dream for the future. Yet this is an attainable dream since its essential basis—a high common civilization—already exists.

Rise of the superpowers

BIBLIOGRAPHY. GEORGE W. HOFFMAN, *A Geography of Europe Including Asiatic USSR*, 3rd ed. (1969); and JEAN GOTTMANN, *A Geography of Europe*, 4th ed. (1970), provide the best accounts of the whole continent. M.R. SHACKLETON, *Europe: A Regional Geography*, 7th ed. rev. (1969), underlines the geological background; while DENYS HAY, *Europe: The Emergence of an Idea* (1957), discusses the concept "Europe."

Standard texts that deal with the geography of major parts of Europe include: R.J. HARRISON CHURCH *et al.*, *An Advanced Geography of Northern and Western Europe* (1967); F.J. MONKHOUSE, *A Regional Geography of Western Europe*, 3rd ed. rev. (1968); A.C. O'DELL, *The Scandinavian World* (1957); W.R. MEAD, *An Economic Geography of the Scandinavian States and Finland* (1958); A.F.A. MUTTON, *Central Europe*, 2nd ed. (1968); N.J.G. POUNDS, *Eastern Europe* (1969); D.S. WALKER, *The Mediterranean Lands*, 3rd ed. (1965); and J.M. HOUSTON, *The Western Mediterranean World: An Introduction to Its Regional Landscapes* (1964). G.W. HOFFMAN (ed.), *Eastern Europe: Essays in Geographical Problems* (1970), examines the effects of state plans on geographical aspects of national economies. Many studies of Europe restrict themselves to specific aspects of the land and the people. W. GORDON EAST, *An Historical Geography of Europe*, 5th ed. (1966), examines geographical changes since Roman times; while M.M. POSTAN, *An Economic History of Western Europe, 1945–64* (1967), discusses the major social and economic changes since the end of World War II. R. BRINKMANN, *Abriss der Geologie*, 8th ed. (1956; Eng. trans. condensed from vol. 2, *Geologic Evolution of Europe*, 1960), summarizes the geological history of the continent; and W.G. KENDREW in *Climates of the Continents*, 5th ed. (1961), describes the climates of its different parts. C.S. COON, *The Races of Europe* (1939), attempts to distinguish and classify the ethnic types of which the peoples of Europe are composed. J. BEAUJEU-GARNIER, *Géographie de la population*, 2 vol. (1956–58; Eng. trans., *Geography of Population*, 1966), discusses past, recent, and current trends of the population of Europe in relation to those of the other continents. S.H. FRANKLIN, *The European Peasantry: The Final Phase* (1969), relates to countries on both sides of the Iron Curtain; while P. LAMARTINE YATES, *Food, Land and Manpower in Western Europe* (1960), brings out relevant economic and regional aspects of agriculture. I.M. COSGROVE and R.T. JACKSON, *The Geography of Recreation and Leisure* (1972), shed new light on tourism—one of the rapidly growing industries of Europe.

Of special interest are two publications resulting from conferences of geographers sponsored by the Council of Europe: E.C. MARCHANT, *Countries of Europe As Seen by Their Geographers* (1970) and *Geography Teaching and the Revision of Geography Textbooks and Atlases* (1967). In the former book a geographer of each member country has written a brief geographical account of his own country, and in the latter are

included accounts of the five major parts of the continent and an assessment of Europe's unity.

For statistical information, the annual publications of the UNITED NATIONS are invaluable, notably *Statistical Yearbook*, *Demographic Yearbook*, and the *Economic Survey of Europe*. The *Statesman's Yearbook*, published annually, contains useful information about European (and other) countries.

(W.G.E.)

Europe, Ancient

For the purposes of this article, Europe will be defined as the land area west of the Ural Mountains and north of the Black Sea and the Mediterranean, with its adjacent islands. The article consists of a survey of the cultural development in this area from the beginning of the Neolithic Period until the 1st century AD. Ecologically Europe falls into three zones:

1. The circumpolar zone of birch, pine, and tundra, comprising much of Scandinavia and European U.S.S.R., in which hunting and catching were the basic modes of subsistence in prehistoric times.
2. The temperate zone, marked by the predominance at low altitudes of deciduous forest, which was able to support mixed-farming economies.
3. The Mediterranean coastal zone, in which tree culture and viticulture were early combined with stock raising and cereal cultivation, and which was the first area in Europe to support urban, literate communities. The histories of these communities (Rome, the Greek city-states, etc.), about which far more is known than about their nonliterate neighbours, are mainly covered in articles specifically devoted to them. The emphasis in this article is on "barbarian" Europe.

Sources. Apart from occasional and sometimes valuable references to the barbarian world in the texts of Greek and Latin authors, and some more remote clues in Egyptian or Hittite records, knowledge of the peasant phase of European prehistory depends primarily upon archaeology. The different cultures belonging to this phase are often identified by the distinctive style of their artifacts, especially their pottery. But there are many difficulties in seeking to interpret such archaeological data in historical terms. For example, the language a people spoke cannot be determined unless they left inscriptions or others recorded it in literature. Again, there is a fundamental divergence between those who interpret every change in style as evidence of intrusions of new peoples and those who see in them indications of dynamic indigenous development. Frequently, of course, both factors may be involved. In determining chronology radiocarbon dating is of prime value for the Neolithic Period; for the Bronze Age, projections from Aegean dating systems can be used; and for the Early Iron Age, Greco-Roman historical dates put matters on an even firmer basis.

Spread of Neolithic farming communities. Agriculture and stock raising, the economic basis of peasant and ultimately of urban life, first appeared in the Aegean about the end of the 7th millennium BC but did not reach the more remote parts of temperate Europe for another 2,000 or in places even 3,000 years. Although settled life based on farming had been developed earlier in parts of southwest Asia, Europe played a far from passive role. Horses, for instance, were almost certainly domesticated first in south Russia and dogs in northern Europe; and it may even be that cattle were first tamed in southeast

Asian and indigenous influences

Chronology of Ancient Cultures

	Britain	Western Europe	Scandinavia	Danube, Central Europe	Eastern Europe, Western Russia	Greece, Aegean	Egypt	Mesopotamia	Rome
7th millennium						Neolithic farming communities			
6th millennium					Starčevo-Körös culture			Neolithic civilization	
5th millennium			Ertebølle-Ellerbek and Funnel-Neck Beaker cultures	Danubian culture	Boian, Cucuteni-Tripolye				
4th millennium	Megalithic cult	Megalithic cult					1st dynasty		
3rd millennium		Bronze Age civilization		Bell-Beaker culture, Bronze Age civilization	Fatyanovo culture	Bronze Age civilization (Minoan)	Old Kingdom	Sumerian civilization	
2nd millennium		Bell-Beaker and "battle-ax" culture				Early Helladic culture (first half of millennium)		Amorite conquest	
2000–1800 BC	Bronze Age civilization			Únětice, Straubing, Alderberg			Middle Kingdom	Babylonians	
1800–1100 BC		Urnfield culture	Urnfield culture	Urnfield culture	Kistapostag, Perjámos, Tomaszów	Dorian invasion	New Kingdom	Assyrians	
1100–600 BC				Hallstatt culture			period of colonization	Neo-Babylonian period	Etruscans early republican period
600–400 BC		La Tène civilization		La Tène civilization		classical period			
400–300 BC		Celtic Early and Waldalgesheim styles			Celtic Early and Waldalgesheim styles	Macedonian conquest			
300–200 BC		Celtic Plastic style, Celtic Sword style, Roman control of the Mediterranean	continuation of La Tène culture		Celtic Plastic style, Celtic Sword style	Roman ascendancy			beginnings of Roman expansion
200–100 BC									expansion into western Europe
100 BC–AD 1	continuation of Sword style								dictatorship, formation of empire
AD 1–100	Roman conquest								

Europe. Again, while it is true that many elements of material culture, notably pot making and the equipment used in harvesting, were invented earlier in southwest Asia, there was certainly no question of the spread of a single, uniform culture over the European continent. The development of Neolithic farming communities, which lay at the very basis of European civilization, was as complex as it was prolonged, and the diverse pattern in the archaeological record reflects considerabe variety in the culture of the indigenous Mesolithic population.

First phase. Thanks in large measure to the application of radiocarbon dating, it is now possible to trace the progressive enlargement of the territories supporting a peasant way of life. Evidence for this appears first in mainland Greece and on Crete, where communities of mixed farmers appeared about 8,000 years ago, men whose most visible memorials are the mounds, or tells, formed, as in southwest Asia, by the disintegration of settlements of mudbuilt houses rebuilt generation by generation on permanently occupied village sites. The artifacts from such settlements—the handmade pots, clay figurines of women, stamps or seals, reaping knives, querns, and polished stone axes—resemble in their general form and character those from Syria, Iraq, and Iran. Yet the pots from Greece have their own idiosyncrasies of style. The vessels with simple, geometric, painted designs that accompany plain wares from this period in Thessaly or western Macedonia differ from comparable Asiatic wares and contrast even more strongly with the deeply incised wares from early levels at Cnossus in Crete, whose affinities lie rather in Anatolia.

Second and third phases. By the middle of the 6th millennium permanent tell settlements had begun to grow up in the southern Balkans. Pottery and ancillary objects in the Starčevo-Körös style occur with minor local variations over a territory extending from the Adriatic coast of Yugoslavia, across to Bulgaria and Romania, and as far as the Prut River in present-day U.S.S.R. The main characteristics of this extensive culture are common to those earlier established in Greece, and it is significant that some of its distinctive features, such as the prevalence of surface roughening or rustication and the occasional use of naturalistic plastic ornamentation, appear on pottery from Greek Macedonia. The middle of the 5th millennium witnessed the expansion of farming economy over a territory centred on middle Europe, but extending to the Rhineland, the Low Countries and northern France in the west, to Hannover and Silesia in the north, and in the east outflanking the Carpathians as far as the western Ukraine. This culture, called Danubian, used pottery ornamented by pairs of parallel lines arranged in spiral or meander patterns. The Danubian peasants lived in communities very similar to large modern villages in the same territories. Their rectangular farmhouses, commonly up to 30 or even 40 metres long, were built on massive timber frames and apparently included storage space and possibly stalls. Their settlements were not marked by tells as those of the Starčevo-Körös and Aegean culture areas. This was due in part to the fact that it was no longer practicable in more temperate latitudes to build in sundried brick or *pisé* but in part also to the fact that the first peasants to occupy the loess of middle Europe appear to have practiced a form of shifting agriculture that involved frequent changes of settlement. In addition to practicing the same form of economy, living in the same sorts of houses, and making the same kind of pottery, the Danubians everywhere favoured polished stone adz blades for working timber, buried their dead in crouched positions in inhumation cemeteries and, at least in the later stages of their culture, were careful to provide them with ornaments made from the *Spondylus gaederopus* mussel.

Danubian culture

Fourth phase. The fourth phase in the expansion of farming economy in Europe can be studied in three main areas. In the east a new peasant culture, the Boian, grew up in Romania beyond the limits of the Starčevo-Körös area; it used distinctive forms of pottery decorated by surface rippling, painting, and excised or incised linear designs. To the north of this culture, and owing something to its inspiration as well as to that of the Danubian, another one, called after Cucuteni and Tripolye, developed between the Seret and the Bug rivers with an extension to the Dnieper. The pottery of this culture was fired in well-controlled ovens to a red or orange colour after having been decorated in curvilinear designs painted onto or grooved into the surface. Its makers occupied rectangular houses grouped into substantial villages. In some of these the houses were arranged radially. Kolomiishchina, near Kiev, comprised a circular setting of about 39 houses with others in the middle; another at Vladimirovka, also near Kiev, had at least 162 houses in five concentric rings.

On the North European Plain during the 4th millennium, it appears that the hunter-fisher population was influenced by the Danubian peasants to the immediate south. The immediate outcome was the Ertebølle-Ellerbek culture based on a mixed hunter-farmer economy and marked in the archaeological record by pointed based pottery and specific forms of flint and antler artifacts. The end of the 4th millennium saw the emergence of the Funnel-Neck Beaker culture, a farming economy extending from the Netherlands across Poland and as far north as southern Norway and middle Sweden.

The Funnel-Neck Beaker culture

The Megalithic cult. At approximately the same time, a distinct cultural province developed in western Europe over a territory which included much of Switzerland, France, the British Isles, and Iberia. Excavation of the waterlogged deposits of the Swiss lakeside settlements belonging to the Cortaillod culture have yielded particularly full information about subsistence and crafts, such as weaving and embroidery, generally absent from the archaeological record. More significant is the region's contribution to the spiritual life of the time, both as regards funerary ritual and public ceremonial activities. The most numerous visible monuments dating from the 3rd millennium and the centuries either side of it are chamber tombs, sometimes cut out of the living rock, more usually built from megalithic blocks packed around with smaller stones, and held in place by overlying mounds. The construction of such tombs, designed as burial places for successive generations of families or clans, involved a physical and social effort comparable to that implied by the building of churches in early historic times and implies well disciplined religious emotion. Some insight into the nature of the Megalithic cult is provided by the symbolic designs pecked, carved, or painted on stones incorporated into megalithic structures. The spread of this cult can hardly be accounted for in simple terms of diffusion, still less of wholesale ethnic movements. The variety of tomb forms and the varying form and incidence of symbolic art point to intense local and regional development and the transmission of ideas over a rather long time. It is significant that by and large the grave goods relate to the indigenous cultures.

In addition to funerary structures, these people erected single, standing stones, or menhirs, alignments and other elongated structures, and circular monuments, sometimes on a massive scale. The most impressive alignments are those at Carnac, Brittany, comprising nearly 3,000 menhirs in parallel rows over a distance of between three and four miles. Smaller alignments occur in southwest England. The British Isles have many circular settings of menhirs or heavy timbers dating from this period; the best known example is Stonehenge.

Menhirs

The closing centuries of the 3rd and the opening ones of the 2nd millennium were marked by the widespread adoption over much of north, central, and eastern Europe of perforated stone battle-axes. This was a time when farming economy spread in European Russia far beyond the forest steppe (to which the Tripolye peasants had been confined) over the southern Taiga up to the Oka-Volga basin, the homeland of the Fatyanovo culture. In northern Europe the Funnel-Neck Beaker culture gave place to a variety of local battle-ax-using groups all associated with single-grave burial, but in other respects showing many differences. Here, again, it seems that

mixed farming continued to provide the basis of subsistence, and there is evidence that a light plow was already in active use as well as disk-wheeled wagons which likewise were drawn by yoked oxen. The construction of collective tombs was abandoned about the same time in western Europe. The rapid and widespread occurrence over large parts of central and western Europe of pots of bell-shaped beaker form, decorated in horizontal zones by finely toothed stamps, suggests the movement of small mobile groups; and there is evidence that these actively sought out copper and gold.

The Bell-Beaker culture

The circumpolar culture. North of the farming zone, extensive parts of circumpolar Europe continued to support communities that lived by hunting, fishing, and gathering. Although there was some overlapping and contact, the culture of the circumpolar people was quite distinct from that of the peasant peoples of the rest of Europe. Even the art of potting, which they took over from their neighbours, was used to make distinctive ovoid pots reminiscent of basket forms; and the deep pits, which together with stamped comb imprints served to decorate these vessels, themselves stemmed, like much else in the culture, directly from Mesolithic antecedents. The widespread use of polishing was another archaic element originally developed in relation to bone but now applied to a variety of knives, projectile heads, and hollow-ground adzes of slate and stone. As hunters the people were accustomed to movement. During the extensive period of snow cover they used heavy sledges, presumably drawn by dogs, for shifting camp and skis for hunting and personal movement. Skin-covered boats represented on rock engravings were probably used for hunting seals and porpoises and for fishing. Dependence on hunting helps to account for the keen observation shown in the representation of animals that played a leading role in their graphic art. Engravings, sometimes life-size, of game animals and carvings of animal heads on a variety of artifacts, from perforated stone axes to the handles of wooden ladles, reveal a mentality totally opposed to that of the peasants to the south, whose art was mainly abstract and repetitive. Burial rites were also distinctive, taking the form of interment in cemeteries, the body fully clothed and in an extended position.

The Bronze Age (c. 2500–c. 650 BC). While the peasants who pioneered farming economy in remote territories of the temperate zone were restricted to stone, metallurgy was already being practiced in parts of southeastern Europe and Iberia. The smelting of metallic copper from its ores was in itself a remarkable feat in the history of technology; it marked the first of what was to prove a long series of operations by which men were able to conjure up new materials, and it is worth emphasizing that to obtain ore, smelt it, and convert it into artifacts implied a more advanced degree of specialization of labour than was needed for a technology based on working flint or stone. It is no accident that in the Old World metallurgy should have been one of the features persistently found in the context of the earliest literate civilizations. Indeed, it has commonly been assumed that the appearance of metallurgy in prehistoric barbarian societies can only be explained in terms of their contacts with early centres of civilization. In the case of southern Iberia, eastern Mediterranean and specifically Cycladic (south Aegean) influence has been detected at copper-working settlements dating from the mid-3rd millennium; the defenses of Los Millares, for instance, display bastions identical in pattern with those of Chalandriani on Syros (Greece). Here it looks as though there were "factories" established for the prime purpose of securing supplies of copper for Aegean markets. The initial impetus for opening up the copper resources of Transylvania may also have come from the demands of Aegean and possibly even Asiatic markets. It is no less true that the exploitation of copper ores in southeastern and central Europe underwent its own development. Whereas in Iberia the initial impetus was soon lost, the opening up of ores in Transylvania led to the development of cuprous industries, first in Bohemia, central Germany, and the east Alps, and ultimately as far away as Ireland, where copper deposits were first sought out by the Bell-Beaker prospectors at about the end of the 2nd millennium. These copper resources catered wholly or in large measure to barbarian markets. Thus shafthole axes and ax-adzes of Transylvanian copper are found predominantly in Transylvania and the Danubian basin both above and below the Iron Gate, though a few reached the Elbe basin; spiral armbands and embossed disks reached as far north as Denmark, where they occur at the end of the local Early Neolithic. Again, flat axes of trapezoidal plan of the form made from east Alpine and central German copper found their way as far north as Scania (Scandinavia).

The beginnings of European metallurgy

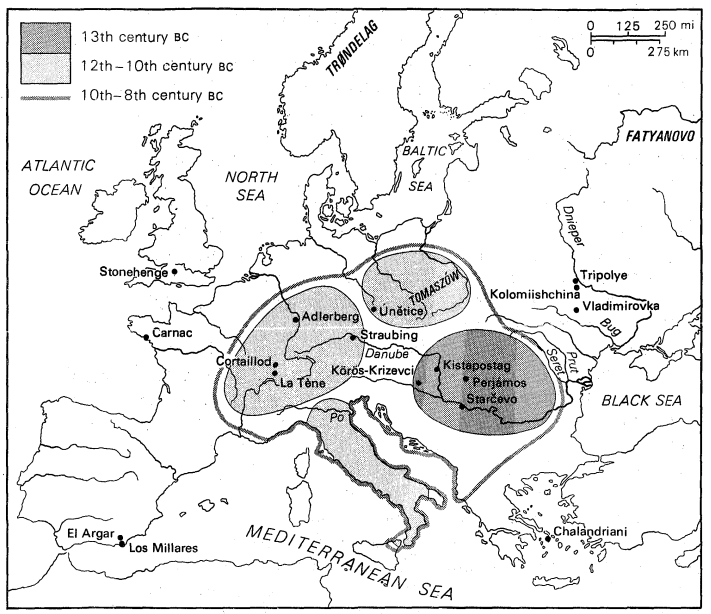
Central Europe was also the original focus of the local use of bronze alloys that not only resulted in tougher tools, but, by making it easier to use valve molds, made it possible to produce more complex forms. The new industries—those of Únětice on the Upper Elbe, of Tomaszów in Poland, of Kistapostag and Perjámos in Hungary, of Straubing on the Upper Danube, and of Adlerberg on the Middle Rhine—produced a variety of bronze items, including daggers, neck ornaments, and pins for securing garments. The occurrence of similar ornaments at Byblos in Lebanon and Ugarit in Syria dating from *c.* 2000–1800 BC suggests continued contact between middle Europe and areas of higher culture to the southeast, but this does not mean that one must attribute the advanced bronze industries of middle Europe around the middle of the 2nd millennium to stimulus from this direction. The Únětice and allied metal industries rested on a middle European tradition already many hundreds of years old. Again, regional bronze-working traditions were developing at this time in many parts of Europe as far apart as Spain, where metallurgy revived in the El Argar culture, and Britain and Ireland, which gave birth to a distinctive Hiberno-British style.

The earliest bronze industries

The use of tin as an alloy laid a great emphasis on organized systems of exchange, since this metal was one with a more restricted distribution than copper. These involved the movement not only of ingots but of finished artifacts, as may be seen in the importation of Hiberno-British and middle European forms of ax to Denmark during the final phase of its Stone Age, a period significantly marked above all by flint daggers based ultimately on metal ones too expensive to import in adequate numbers. It is interesting to note how the Aegean and by this time the specifically Mycenaean world became enmeshed in this wide-ranging system of exchange, due in all probability to their need for central European or Cornish tin. The occurrence of segmented faience beads, of a kind invented in Egypt but probably made also in the Aegean, in the territory of the Únětice and allied cultures, and also notably in Wessex, a territory that probably controlled supplies of Cornish tin, is suggestive of Mediterranean influence in the north. But the importance of the Mycenaeans in the European Bronze Age should not be overemphasized. The indirect effects of the play of the Mycenaean market may have been important —for instance, the Mycenaean demand for amber probably played its part in the rise of the Nordic Bronze Age by making it possible for the Danes to import central European bronzes and ingots—but metalsmiths in several parts of barbarian Europe were producing bronzes that compare favourably with anything produced in the Aegean at the time; and faience beads of local shapes were produced even in northern Britain.

The loss of Mycenaean markets was no longer enough to affect the prosperity of central Europe. Copper was mined on an increasing scale and, being cheaper, was used for tools, armour, and containers, as well as for weapons and ornaments. The later stages of the Bronze Age also were marked by a greater emphasis on cremation and the burial of ashes in urns. Urnfield burial, together with many innovations in industry and armament, was adopted over extensive tracts of Europe from Iberia to eastern Europe and from Italy to south Scandinavia. How much this was due to the normal process

Urnfield cultures



The expansion of Urnfield cultures in Europe.
Adapted from S. Piggott, *Ancient Europe* (1965); Aldine Publishing Company

of diffusion by which innovations spread over extensive territories in the course of time, and how much to ethnic movement, is still obscure. Yet it is hard not to believe that raids by armed marauders did not play a part more especially in relation to the rich territories of the Mycenaean world. Already in the 15th century BC heavy slashing swords, with flanged grips to secure the handle, were being made in east central Europe, and such swords were among the innovations that spread widely with the adoption of urnfield burial. The sudden appearance of these in mainland Greece, Crete, and Cyprus has been identified by some scholars with intrusion of the Dorians mentioned by Thucydides as entering the Peloponnesus at a time when archaeology speaks of the destruction of Mycenaean civilization. Egyptian records tell of raiders from the north as early as the 13th century BC and, in the ensuing centuries, the east Mediterranean region as a whole underwent a time of troubles, a period that witnessed the decline not only of the Mycenaean but also of the Hittite power. Other peoples were on the move at this time and may well have been more important, but it looks as though pressure may have come in part from the north.

The Iron Age in barbarian Europe (c. 650 BC–1st century AD). The main reason why the working of iron was to prove so important is that the ores were widespread and their smelting made it possible for the first time to produce cheap implements in a quantity sufficient to transform technology. Meteoric iron had been utilized from antiquity, but only rarely. Traditionally iron ores were first worked in northern Anatolia, and the Hittites owed some of their military power to their lead in using iron weapons. Conversely, the breakdown of Hittite power early in the 12th century BC allowed ironworking to spread. The first Europeans to practice it were the Greeks; and Greek colonists during the 8th and 7th centuries carried the craft with them both to the Black Sea, on the northern shores of which they established close relations with the Scyths (Scythians), and to much of the European shore of the Mediterranean. Ironworking was brought to east Sicily and Italy as far as Neapolis (Naples) by Greeks, and farther north was adopted by the Etruscans. In the west Mediterranean the Greeks founded Massilia, predecessor of Marseille, around 600 BC and by *c.* 550 BC had established the colony of Emporium on the Catalan coast. Other parts of Mediterranean Europe, notably Sardinia and southern Iberia, were colonized by Phoenicians, who likewise brought ironworking with them.

North of the Alps, ironworking first appeared around the middle of the 7th century BC in the Hallstatt culture centred in Austria, Bohemia, and Bavaria. Thanks largely to a change in burial rites from cremation in urnfields to inhumation in wooden chambers under barrows, a much fuller picture is available at least of the leading members of Hallstatt society. Its warrior chieftains were helmeted and armoured in bronze with breastplates and greaves and were armed with long swords of iron or bronze. In war they were mounted on horses, and in death they were borne to the grave on richly decorated funeral carts mounted on two pairs of spoked wheels. This evidence for armed mobility, with its counterpart of strongly defended hill forts and marsh fortresses, suggests that armed raids, accompanied in some cases by folk movements, may well have played some part in the rapid diffusion of Hallstatt fashions over the territories of the old Urnfield culture, down the Rhine to the Low countries, across France to central Iberia in the west, and in central Europe over the extensive territories from southern Poland to far down the coast of the Adriatic. The Hallstatt culture

The La Tène culture. The later stages of the Iron Age in barbarian Europe were marked above all by the style that takes its name from the Swiss site of La Tène and that first arose between the Upper Danube and the Marne during the 5th century BC. It is hardly possible to explain how the geometric heritage of Hallstatt was transmitted into the brilliant new art of La Tène; nevertheless, certain influences can be recognized. The fantastic treatment of animals and parts of animals could well have been inspired by contacts with the Scyths to the east. Even more important was the impact of Greek art, which influenced both naturalistic representation and the use of designs such as the palmette and tendril. Contact with the Italo-Greek world is indicated by a trail of bronze vessels on the main routes through the Alps; and this can only have been intensified by the Celtic conquest of the Po basin in the 4th century. One outcome was the adoption by the Celts of the two-wheeled horse-drawn chariot.

Celtic art comprised monumental stone figures of men or their heads; coinage; and decorative metalwork related to the warrior class and their women. The last included such things as jewelry, mirrors, armament, and equipment for horses and chariots. Many styles have been recognized. Whereas the Early style still showed strong Greek and Oriental features and was relatively more naturalistic and symmetrical than later Celtic art, the Waldalgesheim style that first appeared around the middle of the 4th century was flowing, nonrepresentational, and asymmetrical. The Plastic style that came in at the beginning of the 3rd century preserved links with both the earlier styles, but laid a significant emphasis on three-dimensional effects. In opposition to this, the Sword style that began at the end of the century relied on flat linear designs. Echoes of continental styles can be detected in Britain, but objects produced there are nevertheless quite distinctive; and since Britain remained outside the Roman Empire until the middle of the 1st century AD, the linear style underwent a prolonged development, particularly well displayed in the engraved bronze mirrors and scabbards. The Celts

The expansion of the Roman Empire over almost the whole of the Celtic-speaking world, apart from northernmost Britain and Ireland, necessarily involved a drastic contraction in the sphere of La Tène culture. The only parts of barbarian Europe to flourish unaffected either by the standardized production of material goods or by the substitution of alien bureaucrats for native heroic chiefs were the extensive tracts to the north and east inhabited by the Teutonic and Slav peoples. These territories, over which ironworking had been practiced for some centuries—in Denmark since *c.* 400 BC—were not of course isolated from what went on in the empire. Mercenaries returning from service must have brought ideas home, and Roman products were imported as far north as Trøndelag (Norway) and Gotland. Yet the importance of the Teutonic and Slav peoples is precisely Romanization and barbarian Europe

that they retained their native vigour and had something positive to contribute to the building of medieval Europe.

BIBLIOGRAPHY. J. BOHM and S.J. DE LAET (eds.), *L'Europe à la fin de l'âge de pierre* (1961), numerous articles by specialists covering the later Neolithic settlement of Europe, mainly on central Europe and the Balkans; V.G. CHILDE, *The Dawn of European Civilization* (1957), the last (6th) edition of Gordon Childe's classic work, still the best survey of European prehistory from the Mesolithic to the Early Bronze Age; G. CLARK, *World Prehistory: A New Outline*, ch. 6–7 (1969), a description of the settled communities of prehistoric Europe before and after *c.* 1500 BC, and with S. PIGGOTT, *Prehistoric Societies*, ch. 10–13 (1970), on the peasant societies of prehistoric Europe; J.G.D. CLARK, *Prehistoric Europe: The Economic Basis* (1952), on the subsistence economies and technology of prehistoric Europeans from the Neolithic to Iron Age; M. GIMBUTAS, *Bronze Age Cultures in Central and Eastern Europe* (1965), especially valuable as a summary of sources originally published in Russian and other east European languages; S. PIGGOTT, *Ancient Europe* (1965), aspects of the cultural history of prehistoric Europe from Neolithic times down to the Celtic period; N.K. SANDARS, *Prehistoric Art in Europe* (1968), especially valuable for Neolithic Balkans and for La Tène art.

(J.G.D.C.)

European Atlantic Coast, Folk Cultures of the

The European Atlantic coast, defined as a cultural area rather than in strictly geographical terms, stretches from the far north of Norway to the west coast of the Iberian Peninsula and includes the Baltic coast areas of Scandinavia, as well as the British Isles. Characterizing this Atlantic fringe of Europe is both the fact and the sense of regionalism, a peculiar fusion of culture and geography backed by an ancient and stubborn ethnicity. The peoples of the area have persistently remained culturally conservative, although they have frequently been forced to switch political allegiance.

PEOPLES AND LANGUAGES

Although they are superficially linked by a common European tradition, these northern and western peripheral areas are the home of diverse peoples of different ethnic stocks and languages. Although the term race has dubious utility for marking off distinctive biological characteristics, most of these coastal peoples have usually been classified, like other Europeans, under the general and geographic racial group Mediterranean, the oldest of the primary Caucasoid races and the one that dominated Europe, North Africa, and the Near East at least from Neolithic times. In the past, scholars also used additional subdivisional terms, such as Nordic, Germanic (or Teutonic), and Alpine (Celtic). But ethnic mixing through migration and intermarriage proceeded on a grand scale for millennia, and regional types blended with the predominant Mediterranean, especially after the extension of ancient Roman influence throughout Europe. Nevertheless, many of these coastal folk cultures successfully resisted complete domination by Rome and thus preserved much of their ethnic and cultural identity. Population geneticists classify peoples in terms of blood-group frequencies; anthropologists use the term "local race." The O-type blood group, for example, is particularly frequent among the Basques, the Irish, the Scots, and the Icelandic peoples. Sometimes, these local races have segments known as isolates, with extremely high endogamy, or inbreeding. Such groups may be called microraces. The Lapps of southern Scandinavia provide a good example of a microrace: low of stature and sallow-skinned, the Lapps's physical characteristics are in marked contrast to their immediate neighbours, the tall, light-skinned, blond Nordic stereotype (*e.g.*, Scandinavians or Celtic-speaking Scots), with gray or blue eyes.

Persistence of local races based on breeding populations

Languages spoken in the Atlantic coastal fringe area are nearly all subfamilies of the Indo-European group—Celtic, Germanic, Baltic, and Romance.

The only modern Celtic language of the Continent, Breton, was taken to Brittany by 5th-century-AD Cornish colonists and refugees from Britain; its closest existing relative is Welsh. There are many descendants of Celtic peoples in Europe who no longer speak Celtic tongues. Continental Celtic or Gaulish languages were once spoken widely in France, in the Iberian Peninsula, throughout central Europe, and into Galatia (Anatolia), but they succumbed well before the Middle Ages to Germanic, Slavic, and Romanic languages. Celtic is divided into the Brythonic (Cymric) branch (Welsh, Breton, and Cornish, the latter extinct since 1777 despite spasmodic attempts to revive it) and the Goidelic (Gaelic) branch (Irish, Scottish, and Manx, the latter virtually extinct).

The Germanic languages of the coastal areas include the Icelandic, Faeroese (in the Faeroe Isles), Norwegian, Danish, and Swedish tongues, in the north, and Dutch, Flemish (in Belgium and northern France), Frisian (closely related to English and spoken in the Dutch province of Friesland), English, and Low German, in the west. The Baltic subfamily includes Latvian, Lithuanian, and Old Prussian (extinct since the 18th century). The Romance (Italic) subfamily includes French (under which is included Walloon, spoken in south Belgium and neighbouring French regions), Provençal, Castilian Spanish, Portuguese, and Galician (spoken in northwest Spain).

Lappish, Estonian, and Finnish properly belong to a separate language family, Finno-Ugric. Basque, which is spoken on both sides of the Pyrenees—in France and in Spain—also defies classification and is not usually included in the Indo-European language family. Also found are vernaculars, or pidgins, which are not true languages since they are derivations from two or more parent tongues. An example is Russenorsk (from Russian and Norwegian), used by Norwegian fishermen with Russian traders.

TRADITIONAL CULTURAL PATTERNS

The diverse peoples occupying this area may be delimited and described in terms of "culture clusters"—regional complexes of cultural traditions that are geographically localized and of long duration.

A main culture cluster is located primarily around the North Sea and also around the Baltic Sea—regions that have frequently attracted peoples from the north European floodplains. This region includes Scandinavia, the Low Countries, and the British Isles. Iceland, settled by Norsemen since the 9th century, is generally included in any survey of the area, whereas Greenland, with its resident Eskimo population and relatively recent Danish colony, is not. At the northern edge, delimited by the Barents Sea, European and Circumpolar peoples (Ostyaks, Samoyeds, Lapps) interpenetrate.

Western France, a peripheral zone of composite character where various cultures have blended, forms a second culture cluster. A zone transitional to the Mediterranean and Alpine areas, it is aptly termed the isthmus of Europe. Linguists distinguish the north of France from the southwestern region, which is associated with the Catalonian fringe.

The Pyrenees form the northern limit of a third culture cluster located in the north and west of the Iberian Peninsula and including the Basques, Portuguese, Asturians, and the Galicians. The Basques are uniquely different from everybody, and Galicia, in northwest Spain, is much closer culturally to northern Portugal than to Castilian Madrid.

Patterns of settlement and economic organization. *Peasants and farmers.* The rural settlement pattern of the Atlantic fringe area has always been notable for its single-family farms in dispersed neighbourhoods or hamlets (*tun* in Norway, *clachan* in Scotland, *bally* in Ireland, and *minifundio* in Galicia). Hayfields and grazing, grain, and kitchen gardens in meadowlands are the predominant forms of land use. Male crofters, or small farmers, tend the cattle, with no pooling of herds and flocks.

Settlement patterns are influenced by a variety of factors, such as ethnic tradition, economic necessity, security, shifting agriculture, specialized cultivation, and fishing. Concentrated settlement is found more among Germanic peoples, and dispersed settlements are common among the Celts; but contrary examples are found. Be-

fore the construction of dikes, settlement sites may have been elevated and agglomerated, becoming dispersed later when tides and floods were controlled. Dependency on the sea and the need sometimes to protect oneself from it are evident in the agglomeration in Norwegian fishing villages and in the dike villages of Holland. No doubt, too, fear of the Norse sea rover drove plainsmen in the coastal areas to the more mountainous regions. Dispersion seems to increase as the land rises and the surface becomes more rugged and broken. There is certainly not the close relationship between water resources and concentration of settlement that occurs in the Mediterranean.

Main agricultural areas

Mixed farming or commercial livestock farming and cereal crop farming extend in a broad belt across the north of Spain, south and middle France, and well into eastern Europe. Subsistence crop and stock farming with little or no surplus and a dependence on rye, maize, potatoes, and barley is restricted, by and large, to the marginal highland areas in Scotland, Wales, and parts of Norway. Commercial dairy farming, which is profitable only in proximity to a large urban market, is found in most of the English Channel and North Sea coastal area —Brittany, Normandy, the Low Countries, northeastern Germany, Denmark, south Sweden, and England.

Holland—with its mixed-farming villages closely related to the topography and nature of the soil, strip villages with farms along a road, peat colonies, and market towns —is characterized by wide agricultural diversity. In the Spanish Basque provinces there is pastoralism in the higher mountains, peasant small holdings in the hills of the northern regions, large agricultural estates in the flat, arid plains of the southern areas, and a fishing and merchant-marine economy on the seacoast. On the ragged coastlines of Scandinavia, the British Isles, Brittany, northern Spain, and Portugal, there is a local dependence on small-scale agriculture supplemented with fishing and pastoralism and perhaps typified by summer pasturing of livestock in the hills.

These forms of economy and the cultural patterns that go with them establish cultural boundaries that are not mutually exclusive or strictly delimited by area; rather, they shade by degrees into each other. Within them there are sub-areal specializations such as that of the wine-growing area of the eastern Paris basin—one of the most ancient districts of specialized horticulture—and the cider-apple belt of Brittany, Normandy, and Devon (England) farther north.

With few exceptions (Belgium, Holland, the Channel Islands, and the Spanish Basque coast, for example), there is a remarkably low population density. Norway, with only 4 percent of its area under cultivation, has the lowest average density (31 persons per square mile in 1970) of any continental European country. Other Scandinavian countries also have low densities because of the high proportion of tundra, waste, lakes, and forests. Where climates and soils are more benign and productive, as in the south and southeast, the people settle thickly. By contrast, some provinces in south Holland have as many as 2,500 persons per square mile (1,000 persons per square kilometre).

Landholdings are small and generally fragmented. Farms in western or southern Ireland may be up to 50 acres (20 hectares) but are usually less than 30 acres, especially in the more remote regions. In the northeast province of Trás-os-Montes, Portugal, or the coastal province of Beira Litoral or scattered and isolated among the mountainous and heath-covered districts of Galicia, Spain, the majority of small farmers cultivate less than 2.5 acres each. Production is low and necessarily wasteful of human energy when only an elementary plow or hoe is used and when much time must be spent going from one small fragmented parcel of land to another.

Land tenure

Land-tenure systems vary. The family-community system, in which the property belongs to the family and the family head chooses his successor, is common in Ireland and to the whole Pyrenean region, though this is customary rather than national law. So, too, is share-tenancy (métayage in Brittany; *aparcería* in northern Spain), by which rent is paid by a share of crops, livestock, or cash. Particular forms of tenancy shade into near-proprietorship and have been the subject of bitter divisions, as, for example, in Galicia. Conflicts generated by the landholding system in Galicia were accompanied by large-scale emigration, debt servitude, and local political bossism (*caciquismo*) until well into the 20th century.

Cooperative work groups are not uncommon, whether, in the past, for road maintenance in Denmark by corvées, for peat cutting in Ireland, for slash-and-burn preparation of the ground for planting (*roza ambulante*) in Galicia, or, in contemporary Ireland, for building a haystack (*meitheal*). Communal or collectivist ownership of resources among agriculturists, shepherds, and fishermen were once found in a great many small towns or villages in northern Spain and the Pyrenees.

The peasant, conceptualized as an ideal type, is locally bound to his subsistence economy, whereas the farmer is ideally conceptualized as a producer for commercial agriculture and the larger world. While both peasants and farmers produce a variety of crops and livestock, it is peasant culture that gives a local stamp to these goods. Since the processing is done locally, products vary in flavour or design according to the locality. Hence, distinctive cheeses, wines, confections, clothes, and handicrafts distinguish one traditional area from another. Some of these, caught up by national or international markets, have become famous beyond the locality. Examples are Mechlin lacework from Mechelen (Malines), Belgium; the knitted jersey from the Channel Island of Jersey; or champagne from the former province of Champagne in northeast France; however, only in remote communities has the local craftsman, or artisan, resisted the competition of the urban factory. More often than not, he is a landless man with little but a garden patch. Indispensable though his services may be, they do not give him equality with his peasant-proprietor neighbours.

Pastoralists. Nor is the social standing of the shepherd and pastoralist high when this is his full-time occupation. True nomadism, in which whole families or communities accompany their herds in seasonal movements between grazing areas, exists only among a declining number of Lapps, among whom herding is a collective undertaking. Their Swedish, Norwegian, and Finnish neighbours tend to look down on them as inferior and uncivilized. Transhumance, in which only those needed to herd the livestock move, is common in Scandinavia and northern Spain, especially in the rainy mountainous areas, such as Galicia and Oviedo. There—known and regarded as a distinctive social, if not ethnic, group—the *vaqueiros de alzada* move from their winter quarters on the seacoast to their summer pastures in the high Cordillera Cantábrica and the Pyrenees. Also semi-nomadic are small bands of Gypsies, who can still be found in most European countries in occupations related to horses and smithing; the tinkers in Ireland, peripatetic salesmen and minor craftsmen, are notoriously distrusted by the settled populace.

Fishermen. Sea fishing has been an integral part of the coastal folk way of life since Neolithic times. Local patois in Normandy and the Channel Islands both abound with old Norse terms related to ships and the sea. Nearly 20 percent of all Norwegians take to the sea in one form or another of merchant craft, their fleets of small boats plying inshore fishing grounds for sardines, in contrast to the deep-sea, distant-water trawlers of the British fishermen seeking cod. Many coastal Lapps are settled now as fisherfolk. The only Gypsies, who traditionally abhor the sea, known to have become sailors were 40 families living among the French Basques in the area of Saint-Jean-de-Luz. Longboat regattas are the only vestiges of a whaling culture in northern coastal towns of Spain, where fishing as a local industry has been closely regulated for centuries. Until dissolved by royal order in 1865, for example, the Cantabrian coastal brotherhood (*cofradía* or *gremio*) of San Martín de Hijosdalgo, in Laredo, Santander, controlled and apportioned fishing grounds, especially for bream, by lot. Fishing can be a cooperative enterprise requiring large numbers of men (and women, for net

mending), and elaborate rules still exist for regulating disputes. Especially in smaller villages but even in the larger ports, local traditional lore still sets formulas for sharecropping of the catch, which is usually auctioned at the wharfside. Yokes of oxen may even now be seen dragging enormous pocket seines to the beach at Finisterre and other ports in the Galician fjords.

Social organization. Atlantic coastal folk did not usually fall under the sway of feudal lords or great princes as serfs or villeins, for the power of feudal lords did not reach effectively into the mountains, forests, or fenny marshlands where the peasantry more freely practiced their autonomous forms of social control and organization based on a local rather than a "national" stratification of society. Absentee landlord systems in these marginal areas permitted tenants to maintain a cultural isolation; at most, a local political bossism affected them in times of crisis, rather than with calendrical regularity.

Kinship patterns

Local endogamy reinforced a bilateral (patrilineal and matrilineal) family system (possibly patrilineal in old Ireland) and ensured that most members of a small community were related in some way. The small, or stem, family from which branches broke away was the norm. A group of kindred was the repository for land property ownership and the unit of cooperative action, especially in the case of factional dispute with another locality. But control of land and capital was invested in the male family head, not in any lord of a manor. Sons often remained throughout their youth and into middle age as unmarried labourers on the family farm, until the family head chose their successors. Late and uncertain marriage was a condition of this arrangement, which, in both Ireland and Norway, is known as a marriage-succession system.

Isolated coastal folk generally had few links with their more stratified neighbours in larger villages or towns. Nonpeasant itinerants, such as entertainers, salesmen, bards, and the like, were the most important agents for the transmission to isolated coastal villages of beliefs, proverbs, ballads, and news.

Religion. Local traditions of cult practice and belief have persisted despite the national or international impact of organized religion, for local gods proved difficult to supplant. Scandinavian mythology is full of culture heroes immortalized in sagas or heroic poetry (*Eddas*) and exalted to divine rank. This polytheism, or veneration of many deities, gave a local quality to worship, for the gods were regional champions. Associated beliefs and practices relating holy wells or medicinal springs, megaliths, sacred trees or groves, sacrifices, divination by oracle (consulting the waters), burials, and pilgrimages were common, as were fairy and witch cults, the laying of curses, and spiritual punishment for polluting the wells.

Local religious traditions were to some extent linked with Christianity through the cult of early Christian saints. Before the Reformation, pilgrimages to holy wells (and devotional exercises there) were undertaken by the highest in the land, and, despite insistent attack since the Reformation, such local cults have persisted into the 20th century in one form or another. Pilgrims streamed to Santiago de Compostela, in northwestern Spain, throughout the Middle Ages after the 9th-century discovery there of the St. James' tomb. Fátima in Portugal and Lourdes in France are more modern pilgrimage sites. Watering places, or spas, are simply examples of a laicized tradition. The tossing of coins or pins into a "wishing well" may be compared to the prehistoric votive offerings of bronze at medicinal springs (in Denmark) and is an example of a superstitious practice supplanting a religious one. The patron-saint complex is still very much alive among Catholic folk, and pastors and priests in Protestant and Catholic areas alike are still called upon to serve a spiritual need in local agricultural and pastoral activities, as, for example, in the blessing of crops or animals on selected festival dates.

MODERNIZATION

The move from a medieval localism to a centralism personified by a temporal sovereign substituted the lord–follower relationship, by which Atlantic coastal folk were rarely affected, for a king–subject relationship, which affected them increasingly. The sense of nationalism that developed in Europe from the 16th century on was based on a common language, common religion, and patriotic military service, which tended to submerge the cultural distinctiveness of the coastal areas. But, if, at times, the isolation of folk cultures was broken by this new cultural system, the limited contact that ensued did not result in acculturation or complete integration. Political wars between European nations and overseas expansion, in fact, drew attention away from these marginal areas.

Regionalism, in the sense of mere physical isolation, was not completely broken down in many folk communities until the 19th century. Begun by an economic process and increased communications, it is now being completed by tourism. "Park-keeper peasants" is an ironic term applied to those living in areas that governments are developing for tourism rather than for industrial or agricultural growth. In general, the important changes can be described in terms of an opening up of local economies to national markets. This involves a shift from either mixed farming to market gardening or from subsistence farming to commercial agriculture and other forms of enterprise.

The linking of Brittany by a new railway to Paris in 1880 greatly accelerated cash cropping and the growth of canneries. It also led to larger fishing fleets; today, in all major ports along the Atlantic coast, corporations rather than individuals provide vessels, equipment, hire fishermen at fixed wages, and generally govern the industry. One Lapp herdsman on a snowmobile and with a walkie-talkie radio can now do more than ten herdsmen did in bygone years. Those who are thus freed from traditional tasks are not all re-employed in the mining, railway, or hydroelectric projects that are invading previously pastoral retreats.

The mobility—physical and social—of coastal folk has dramatically increased, and there has been much emigration abroad, especially to the United States, and, internally, to the cities, a phenomenon that has not ceased.

Politicized cultural movements of the later 20th century

Political representatives of many European folk minorities see their people as victims both of dominance and neglect by central governments. There have been repeated demands from these minorities for sovereignty over their own political affairs, as in Wales, or militant movements in favour of separate nationhood, as among Basques, Bretons, and Galicians; and, under the banner of the Celtic cross, Bretons and Welshmen have combined in a Pan-Celtic movement to fight to maintain their cultural heritage. Separatist-movement leaders are generally urban based but have recourse to a rural folk symbolism already propagated by intellectuals in cultural movements. A doggedly persistent cultural consciousness has thus been given new political expression in recent years.

BIBLIOGRAPHY. ROBERT J. THEODORATUS, *Europe: A Selected Ethnographic Bibliography* (1969), is the latest general bibliographic source. ROBERT T. ANDERSON, *Traditional Europe: A Study in Anthropology and History* (1971), deals with Europe in general up to 1800; his *Modern Europe* (1972), synthesizes the period from 1800 to the present day, delineating trends of culture change. JOHN GEIPEL, *The Europeans* (1969), covers origins, interrelations, and languages and explodes racial myths. WILLIAM Z. RIPLEY, *The Races of Europe* (1899), persisted as the classic authority for racial classification until superseded by population geneticists, such as W.C. BOYD, *Genetics and the Races of Man* (1950). STUART PIGGOTT, *Ancient Europe from the Beginnings of Agriculture to Classical Antiquity* (1965), covers the interrelations of literate innovating cultures and nonliterate conservative cultures. V. GORDON CHILDE, *Prehistoric Migrations in Europe* (1950), presents the diffusionist view of culture change but makes minimal use of the migrationist hypothesis. R.W. EHRICH, "On the Persistences and Recurrences of Culture Areas and Culture Boundaries During the Course of European Prehistory, Protohistory, and History," in *Bericht über den V. Internationalen Kongress für Vor- und Frühgeschichte, Hamburg 1958*, item 77, pp. 253–257 (1961), provides a dynamic view of the culture area approach that is discussed

as part of a more detailed subareal coverage by MICHAEL KENNY and CONRAD ARENSBERG *et al.*, "Europe and Its Cultures," *Anthrop. Q.*, vol. 36, no. 3 (1963). The evolution and interrelation of different types of economy and settlement may be found in HERBERT J. FLEURE, "Social Organisation and Environment," in PHILIP L. WAGNER and MARVIN W. MIKESELL (eds.), *Readings in Cultural Geography*, pp. 491–505 (1962), and his and others' articles in *Jl. R. Anthrop. Inst.* (1916, 1920, 1936, 1958); and DERWENT S. WHITTLESEY, "Major Agricultural Regions of the Earth," *Ann. Ass. Am. Geogr.*, 26: 199–240 (1936). Useful, more localized, sources are GERALD BRENAN, *The Spanish Labyrinth* (1960); T. LYNN SMITH, "The Social Relationships of Man to the Land in Portugal," *Sociologia*, 25:319–343 (1963); CONRAD M. ARENSBERG, *The Irish Countryman* (1937, reprinted 1968); RONALD FRANKENBERG, *Communities in Britain* (1966); ROBERT T. and BARBARA G. ANDERSON, *The Vanishing Village* (1964), on Denmark; GORDON WRIGHT, *Rural Revolution in France* (1964); and WILLIAM A. DOUGLASS, *Death in Murélaga* (1969), on a Basque village. Detailed discussions of local and regional phenomena are found in FRANCIS JONES, *The Holy Wells of Wales* (1954); and J.A. MACCULLOCH, *The Celtic and Scandinavian Religions* (1948). The blurring of the rural–urban distinction and the growth of a predominantly middle-class society is examined in ROBERT T. ANDERSON (*op. cit.*); see also S.H. FRANKLIN, *The European Peasantry: The Final Phase* (1969); and for contrast JOHN J. HONIGMANN (ed.), "Modernization and Tradition in Central European Rural Cultures," *Anthropologica*, vol. 12, no. 1 (1970). The political cost of integration is lucidly presented by ORIOL PI-SUNYER (ed.), *The Limits of Integration: Ethnicity and Nationalism in Modern Europe* (1971).

(M.Ky.)

European Culture Since 1800

In the cultural history of the 19th and 20th centuries, the first notable fact is the continuous presence of two great sources of energy: one, the French Revolution with its sequel, the Napoleonic saga; the other, the Romantic movement, whose origins lay in the human tendencies that the preceding Enlightenment had tended to ignore. Thanks to the momentum of these two forces, what happens in European culture during the years following 1789 is an unfolding rather than a series of clashes and reactions. The continuity and kinship between the numerous schools of art and thought exist despite evidences of opposition, and they are particularly evident in the realm of artistic techniques. Nor does this connectedness contradict the actual conflicts—often violent—between the moods of the successive generations and between the aims they verbalized while producing the works now classified under the various labels of Romanticism, Realism, Pre-Raphaelitism, Symbolism, Naturalism, Impressionism, Neoclassicism, and Modernism.

This article is divided into the following sections:

Romanticism, reaction, and revolution (early 19th century)
- The legacy of the French Revolution
- General character of the Romantic movement
- Romanticism in literature and the arts
- Social and political thought
- Philosophy
- Religion and its alternatives

The middle and late 19th century
- Realism and *Realpolitik* in midcentury
- Realism in the arts and philosophy
- Late-century decadence and doubt

Alienation and despair (20th century)
- Attempts at improvement
- The prewar period
- The impact of war
- Culture since 1920

ROMANTICISM, REACTION, AND REVOLUTION (EARLY 19TH CENTURY)

The legacy of the French Revolution. To make the story start in the year of the French Revolution is at once convenient and accurate, even though nothing in history "starts" at a precise moment. For although the Revolution itself had its beginnings in ideas and conditions preceding that date, it is clear that the events of 1789 brought together and crystallized a multitude of hopes, fears, and desires into something visible, potent, and irreversible. To say that in 1789 reform becomes revolt is to record a positive change, a genuine starting point. One who lived through the change, Liancourt, was even sharper in his vision when (as the story goes) he answered Louis XVI, who had asked whether the tumult outside was a revolt: "No, sire, it is a revolution." In cultural history as in political, significance is properly said to reside in events; that is, in the acts of certain men or the appearance of certain works that not only embody the feelings of the hour but also prevent other acts or works from having importance or effect. To take an example from English poetry, after Wordsworth and Coleridge's *Lyrical Ballads*, which appeared in 1798 and remained a long while obscure, the writing of heroic couplets in the 18th-century manner could be but a pastime—not poetry, as it had once been. A good many such "punctuation marks" can be noted in the text of cultural activity and in its context of political and social life.

To say, then, that the cultural history of the later modern age—1789 to the present—begins with the French Revolution is to discuss that Revolution's ideas rather than the details of its onward march during its first 10 years. These ideas are: the recognition of individual rights, the sovereignty of the people, and the universal applicability of this pair of propositions. In politics the powerful combination of all three brings about a permanent state of affairs: "the revolution" as defined here has not yet stopped. It continues to move the minds of men, in the West and beyond. The revolution is "dynamic" because it does not simply change rulers or codes of law but it also arouses a demand and a hope in every man and every people. When the daily paper tells of another new nation born by breaking away, violently or not, from some other group, the Revolutionary doctrine of the sovereignty of the people may be observed still at work after two centuries.

The ideas of the French Revolution

Cultural nationalism. The counterpart of this political idea in the 19th century is cultural nationalism. The phrase denotes the belief that each nation in Europe had from its earliest formation developed a culture of its own, with features as unique as its language, even though its language and culture might have near relatives over the frontier. Europe was thus seen as a bouquet of diverse flowers harmoniously bunched, rather than as a uniform civilization stretching from Paris to St. Petersburg, from London to Rome, and from Berlin to Lisbon—wherever "polite society" could be found, a society acknowledging the same artistic ideals and speaking French. In still other words, the Revolutionary idea of the people as the source of power ended the idea of a cosmopolitan Europe.

The "uniform" conception presupposed a class or elite transcending boundaries; the "diverse" implied a nation of citizens attached to their native soil and having an inborn and exclusive understanding of all that had been produced on it. In each nation it is the people as a whole, not just the educated class, that is deemed the creator and repository of culture; and that culture is not a conscious product fashioned by the court artists of the moment: it is the slow growth of centuries. Thus can be seen one of the motives of the post-Revolutionary passion for historical studies—the change to cultural populism as the replacement of a single horizontal, Europe-wide, and "sophisticated" culture by a series of vertical, national cultures, popular not only in their generality but also in their simplicity.

This new outlook, though propagated by the Revolution, began as one of those subdued feelings mentioned earlier, as undercurrents beneath Enlightenment doctrine. In England and Germany especially, a taste developed for folk literature—the border ballads, the legends and love songs of the people, their dialects and superstitions. Educated gentlemen collected and published these materials; poets and storytellers imitated them. Horace Walpole in *The Castle of Otranto*, Macpherson in *Ossian*, Chatterton in his forgeries of early verse, and Goethe in his lyrics exploited this new vein of picturesque sentiment. A scholar such as Herder or a poet-dramatist such as Schiller drew lessons of moral, psychological, and philosophical import from the wisdom found in the subculture of *das Volk*. The folk or people was not as yet very clearly defined, but the Revolution would shortly take care of this omission.

Beginnings of cultural nationalism

In France, where the Revolution occurred, the situation was somewhat different. There were no collectors of border ballads or exploiters of Gothic superstitions. France by 1789 had been for more than a century the cultural dictator of Europe, and it is clear that in England and Germany the search for native sources of art was stimulated by the desire to break the tyranny of the French language and literature. The rediscovery of Shakespeare, for example, was in part a move in the liberation from French classical tragedy.

Simplicity and truth. But it was also the expression of a genuine desire for simplicity and truth and for the release of feelings that the confidence of the Enlightenment in the power of reason had tended to suppress. Two 18th-century figures tapped this fount of emotion, Richardson and Rousseau. The novels of Richardson, in which innocent girls are portrayed as withstanding the artful seductions of titled gentlemen, might be said to foreshadow in symbolic form the struggle between high cosmopolitan culture and the new popular simplicity. These novels were best sellers in France, and Rousseau's *Nouvelle Héloïse* followed in their wake, as did the bourgeois dramas of Diderot, Beaumarchais's satirical comedies about the plebeian Figaro, and the peasant narratives of Restif de la Bretonne, to mention only the most striking exemplars of the new simplicity.

At the very centre of sophistication the simple life became a fad, the French court (including Marie-Antoinette) dressing up and playing at the rustic existence of milkmaids and shepherds. However silly the symptoms, the underlying passion was real. It was the periodic urge of complex civilizations to strip off the social mask and recover the happiness imagined as still dwelling among the humble. What was held up to admiration was honesty and sincerity, the strong and pure feelings of people unspoiled by city and court life. Literature therefore came to express an acute sensitivity to scenes of undeserved misfortune, of heroic self-sacrifice, of virtue unexpectedly rewarded—a sensitivity marked by tearfulness, actual or "literary."

Back-to-nature movement

This surge of self-consciousness about sophisticated culture has often been confused with an idealization of primitive man and attributed to Rousseau. But contrary to common opinion, the so-called back-to-nature movement does not at all echo the noble-savage doctrine of the 17th century. Rousseau's attack on "civilization," which evoked such a powerful response in the latent feelings of his contemporaries, goes with a characterization of the savage as stupid, coarse, and amoral. In Rousseau and his abettors, what is preached is the simple life. What nature and the natural really are remains to be found by trial and error—the fit methods and forms of religion, marriage, child-rearing, hygiene, and daily work.

Populism. It is easy to see in these beliefs and sentiments (which often passed into sentimentality) additional materials for the populism that the Revolution fostered. Revolution, to begin with, is also an urge to simplify. The Revolutionary style was necessarily populist—Marat's newspaper was called "The Friend of the People." The visible signs that a revolution had occurred included the wearing of natural hair instead of wigs and of common trousers instead of silk breeches, as well as the use of the title of *citoyen* instead of *Monsieur* or any other term of rank. Now, equality coupled with sincerity and simplicity logically leads to fraternity, just as honest feeling coupled with devotion to the people leads to puritanism: a good and true citizen behaves like a moral man. He is, under the Revolutionary principles, a responsible unit in the nation, a conscious particle of the will of the sovereign people, and as such his most compelling obligation is love of country—patriotism.

With this last word the circle of ideas making up the cultural ambient of the French Revolution might seem to be complete. But in the effort to trace back and interweave the strands of feeling and opinion that make up populism, one must not overlook the first political axiom of revolutionary thought, which is the recognition of individual rights, often said to be natural rights as well. Their essence is a subject for political theory. Here their cultural role is of interest, and it can be stated in a very obvious way: individual rights generate individualism. The "ism" rightly suggests an attitude or doctrine—sometimes a passionate faith—asserting that every human being is an object of interest in himself, an end in himself. What is more, the truly valuable part of each individual is his uniqueness, which is entitled to thrive and develop free of oppression. That is why the state guarantees the citizen rights as against itself and other citizens. Again, this power accrues to him for himself because he is inherently important—not because he is son or father, peasant or overlord, member of a clan or a guild.

Doctrine of individualism

These ideas shift the emphasis of several thousand years of social beliefs and let loose innumerable implications. Individualism lowers the value of tradition and puts a premium on novelty. True, the individual soul had long been held unique and precious by Christian theology, but Christian society had not extended the doctrine to every man's mundane comings and goings. Nor were his practical rights and powers attached to him as a man but, rather, to his status within the social scheme. Now the human being as such was being officially considered self-contained and self-propelling; it was a new regime and its name was liberty.

Nature of the changes. The contents and implications of these potent words—liberty, equality, and fraternity, individualism and populism, simplicity and naturalness—aid in the understanding of the cultural situation of Europe at the dawn of the era under review. But these continuing ideas necessarily modified each other and in different times and countries were subject to still other influences.

For example, the active phase of the Revolution in France—say, 1789 to 1804—was influenced by the classical education of most of its public men. They had been brought up on Roman history and the tales of Plutarch's republican heroes, so that when catapulted into a republic of their own making, the symbols and myths of Rome were often their most natural means of expression. The eloquence of the successive national assemblies is full of Roman allusions. Later, when General Bonaparte let it be seen that he meant to rule France, he was denounced in the Chamber as a Caesar; when he succeeded, he took care to make himself first consul, flanked by two other consuls, to show that Caesar was in abeyance.

Changes in the fine arts

In the fine arts this Roman symbolism facilitated a thorough change of taste and technique. The former "grand style" of painting had been derived from royal and aristocratic elegance, and its allusions to the ancient classical past were gentle and distant, architectural and mythological. Now, under the leadership of the painter David, the great dramatic scenes of ancient history were portrayed in sharp, uncompromising outlines that struck the beholder as the utmost realism of the day.

In David's "Death of Socrates" and "Oath of the Horatii" civic and military courage are the respective subjects; in his pencil sketches of the victims of the Terror as they were led to execution, reportorial realism dominates; and, in his designs for the setting of huge popular festivals, David, in collaboration with the musicians Méhul and Grétry, provided the first examples of an art in scale with the new populism: the taste of the court and the city was transcended by the needs of the nation.

But it must be added that except for a few canvases and a few tunes (including the "Marseillaise") the quality of French Revolutionary art was not on a par with its aspirations. Literature in particular showed the limitations under which revolutionary artists must work: political doctrine takes precedence over truth, and the broad effects required to move the masses encourage banality. There is no French poetry in this period except the odes of Chénier, whom the Revolution promptly guillotined, as it did France's greatest scientist, Lavoisier. The French stage was flourishing but not with plays that can still be read. The Revolutionary playwrights only increased the dose of sentiment and melodrama that had characterized the stage at the close of the old regime. The aim was to hold up priests and kings to execration and to portray examples of superhuman courage and virtue.

Modern opera goers who know the plot of Beethoven's *Fidelio* can judge from that sample what the French theatre of the Revolutionary years thrived on. Others can imagine for themselves Molière's *Misanthrope* rewritten so as to make Alceste a pure patriot and hero, undermined by the intrigues of the vile courtier Philinte.

It may seem odd that once the Revolution was under way there should be so much indignation and protest against courtiers, priests, and kings and such fulsome homage paid to virtue and patriotism. What accounts for it is the difficulty of transforming culture overnight. People have to be persuaded out of old habits—and must keep on persuading themselves. Even politically, the Revolution proceeded by phases and experienced regressions. Manners and customs themselves did not change uniformly, as one can see from portraits of Robespierre at the height of his power wearing a short wig and knee breeches, republican and Rousseauist though he was.

Bonaparte's coup d'etat

Napoleon's influence. After Bonaparte's coup d'etat tension eased as the high ideals dropped to a more workaday level, just as the puritanism was replaced by moral license. And as a by-product of Bonaparte's expeditions to Egypt in 1798–99, a new stylistic influence was added to the Roman: sphinxes and the lion-claw motif appeared in furniture, and the Near East began to attract attention. The Roman idea itself shifted from republic to empire as the successful general and consul Bonaparte made himself into Napoleon.

The Emperor had an extraordinary capacity for attending to all things, and he was concerned that his regime should be distinguished in the arts. He accordingly gave them a sustained patronage such as a Revolutionary party rent by internal struggles could not provide. But Napoleon had tastes of his own, and he had to control public opinion besides. In literature (he had been a poet and scribbler in his youth), he relished the Celtic legends of *Ossian* and encouraged his official composer Lesueur in the composition of the opera *Ossian ou les Bardes.* In painting, he favoured the surviving David and the younger men Gros and Géricault, both "realists" concerned with perpetuating the colour and drama of imperial life, including the life of the battlefield. But to depict contemporary issues on the stage (except perhaps in the ballet, which was flourishing) did not prove possible, for the stage must present genuine moral conflict if it is to produce great works, and moral issues are not discussable under a political censorship.

The paradox of the Napoleonic period is that its most lasting cultural contributions were side effects and not the result of imperial intentions. Two of these contributions were books. One, Chateaubriand's *Génie du christianisme* (1802; *The Genius of Christianity*), was a long tract designed to make the author's peace with the ruler and revigorate Catholic faith. The other, Madame de Staël's *De l'Allemagne*, was a description of the new culture in Germany. Napoleon prohibited the circulation of the book in France, but its message easily survived. The two other sources of future light were men. One was the group of philosophers known as Idéologues, who were scientific materialists and concerned with abnormal psychology. The other was Napoleon Bonaparte himself, or rather the figure of Napoleon as seen by his age after Waterloo.

General character of the Romantic movement. The mention of Waterloo (1815) suggests the need to make clear a number of chronological discrepancies. It has been possible so far to discuss the general shift in the temper of European life without naming fixed points. It sufficed to say "before or after 1789" or "from 1789 to the Napoleonic Empire." But from now on the generations of culture makers and the dates of some of their works must be duly situated, without on that account losing sight of unities and similarities in the onward march of artistic and intellectual movements. If, for example, one takes the Romantic poets, one finds that Goethe came to maturity in the 1770s, when the English Romantics were just beginning to be born. Their French, Italian, Russian, Polish, and Spanish counterparts were, in turn, born about the year 1800, when the English were already in mid-career. The same irregularity in the onset of Romanticism is found in the other arts, and it is complicated (at least superficially) by the names given to various movements and persons in the different countries of Europe. Thus, in Germany the term *Romantismus* is applied to only a small group of writers, and Goethe and Schiller are called classic. In Poland and in Russia, classic is likewise the label for the great writers whose characteristics in fact align them with the Romantics elsewhere.

Chronological discrepancies in the Romantic movement

All these accidents of birth and nomenclature can be taken in stride by remembering the patterns found in each country or decade and the reasons for their appearance at that time and place. Within the slightly more than half century between 1789 and 1848, the phenomenon of Romanticism occurred and developed its first phase. Those who made it may have come early or late, belonged to this or that nationality, proved to be originators or synthesizers of existing elements—all such considerations appertain to individual biography or the history of a particular art or nation. What matters in the evolution of European culture considered as a whole is the orchestration of all the voices as they come in to swell the ensemble.

The main purport of the Romantic movement is said to be a revolt against 18th-century Rationalism and a resulting variety of new attitudes and activities: a turning in upon the self, a love of nature, the rediscovery of the Middle Ages, the cult of art, a taste for the exotic, a return to religion, a fresh sense of history, a yearning for the infinite, a maudlin sentimentality, an overvaluing of emotion as such, a liberal outlook in politics, a conservative outlook, a reactionary outlook, a Socialist-utopian outlook, and several other "characteristic features."

It is clear that not all of these can be equally true, characteristic, or important, since some contradict the others. At the same time it was inevitable that so sweeping a cultural revolution as Romanticism should contain incompatible elements. For instance, the political opinions enumerated above did in fact win the allegiance of different groups among the Romantic artists and thinkers for a longer or shorter time. But—to take note of other supposed definitions—not all Romanticists returned to religion: Goethe and Berlioz were pantheists; Byron and Heine, atheists; and Victor Hugo, a sort of Swedenborgian. As for sentimentality, its occurrence was rather a hangover from the 18th century than a new fashion of feeling, for the Romantic cult of art and of strong emotion goes dead against the weak sentimental mood. Similarly, the taste for history, for the Middle Ages, and for the exotic shows a strong curiosity about the particulars of what is real though ignored by previous conventions. All critics, however, are agreed upon one Romantic trait: individualism. And it is here that the figure of Napoleon plays its cultural role.

Incompatible elements of Romanticism

Napoleon was regarded as the model of a new man, even during his life, because his career was manifestly the product of his own thought and will working against the greatest imaginable resistance. He typified the individual challenging the world and subduing it by his genius. A movement that numbered as many artists and geniuses as did Romanticism was bound to find in Napoleon the individual par excellence or, as might be said in modern jargon, a supremely autonomous personality. This perception explains why nearly all the great names of the first half of the 19th century are found on the roster of those who praised Napoleon—from Beethoven and Byron to Hazlitt and Stendhal and Manzoni. Some who were politically his enemies—Sir Walter Scott, for example—nonetheless respected and pondered over the miracle of his achievements. No comparable attention has been paid to the dictators of the 20th century, a fact sufficiently explained by the real difference between Napoleon and them. Stendhal, who had taken part in the Russian Campaign of 1812, analyzed that difference ahead of time: Napoleon had intellectual power—and not merely political and military. In whatever he did, he

showed originality of conception, a stupendous grasp of detail in execution, and the utmost speed in acting out his vision. This sequence, translated to other realms, was the very pattern of the creative imagination. It also seemed the vindication of individualism as a philosophy of life: open the world to the individual and the world will witness marvels unimagined before.

The Romantics' attitude toward themselves

These remarks about Napoleon should convey a sense of the Romantics' attitude toward themselves and their situation. It is true that culturally they stood in opposition to their immediate forebears. But all generations do the same, and yet it is only once in a while that the conflict produces great things. Romanticism had the paradoxical advantage of making its way during and after the 25 years of Revolutionary and Napoleonic wars. For the tumult of battle and political overturns did its share to clear the ground for artistic innovation. When habits and expectations are repeatedly upset and frustrated in the broad public realm, the general mind opens up to novelty offered in other realms. That is one avenue of cultural, stylistic, and emotional change. When Stendhal was expounding Romanticism to the French in 1822, he argued that to go on writing in the Neoclassic vein was "to provide literary pleasure for one's grandfather." His remark was readily understood—at least by his young readers. Mighty events had dug a chasm between past and present, making plain the remoteness of the 18th century.

And yet a paradox remains. When the Romantic innovator produced his first characteristic work—say Wordsworth with the *Lyrical Ballads* of 1798—he had to wait a good while for a hearing though his sense of the death of Neoclassicism was clear; it arose from a poet's quick perception of the decay of forms. Already in 1783 Blake had written of contemporary English verse that "The sound is forced, the notes are few." But these two poets' estimate of the state of poetry was not widely shared till the passage of time—*i.e.*, of men, and events—had worn down old routines. It is generally said that Blake and Wordsworth and most other Romantic artists were ahead of their time; it would be more accurate to say "ahead of their audience."

This phenomenon, characteristic of the modern age, further complicates the question of when Romanticism emerged or was established in different countries. Even so, it is possible to assign, with sufficient reason, a particular decade when in a given nation the movement was in being and active in the production of important works, though not necessarily accepted by a large public. In England and Germany that decade was the 1790s: Blake, Wordsworth, and Coleridge; Goethe (with the first fragment of *Faust*), Schiller, Herder, Jean Paul (Richter), Beethoven, Tieck, Wackenroder, Hölderlin, Schelling, Schleiermacher, together with the beginning of a German translation of Shakespeare by Tieck and Schlegel, mark the advent of the new age.

In Italy, France, and Russia, the decisive years opened in 1820. They are signallized in Russia by the abundant poetic output of Pushkin, in Italy by the work of Manzoni and Leopardi and by the surrounding discussions of literary theory, and in France by the poems of Lamartine, Vigny, Victor Hugo, and Mme Desbordes-Valmore. The paintings of Delacroix, the first compositions of Berlioz, and Balzac's *Chouans* show that a new spirit was at work. Finally, in the 1830s, Poland—through its poet and novelist Mickiewicz—and Spain—through the works of Rivas, Espronceda, José de Larra, and Zorrilla—joined the rest of Europe in its richest artistic flowering since the Renaissance: the leading nations can boast one or more Romantic artists of the first magnitude.

The fundamental Romantic purpose

Romanticism in literature and the arts. The fundamental Romantic purpose was to grasp and render the many kinds of experience that classicism had neglected or had stylized. Romanticism was the first upsurge of realism—exploratory and imaginative as to subject matter and inventive as to forms and techniques. The exploration of reality surveyed both the external world of peoples and places and the internal world of man. The Scottish and medieval novels of Sir Walter Scott, beginning with *Waverley* in 1814, illustrate the range of the new curiosity, for Scotland was a "wild" place, outside the centres of civilization, and the Middle Ages were similarly "barbarous" and distant in time. When Byron went to the Near East and used it as the setting of his adventure poems, he was creating the same sort of interest, this time in a region as real and "central" (to itself) as Paris and London. In both these writers, factual detail is essential to the new sort of effect: the scenery is observably true, and so is the history, given through local colour. As Byron said when criticized: "I don't care two lumps of sugar for my poetry, but my costume is correct." Blake, 20 years earlier, had taken a stand against Sir Joshua Reynold's academic doctrine that the highest form of painting depicted the broadest general truth. Said Blake: "To particularize is the only merit."

Particulars, moreover, are all equally proper for the artist; the use he makes of them is what matters. When Wordsworth and Coleridge sought to revivify English poetry, they hit upon two divergent kinds of subject: Coleridge took superstition and the folk tale and wrote "The Rime of the Ancient Mariner" in the form of an old ballad; Wordsworth took the modern street ballad—a kind of rhymed newspaper—and produced his versified incidents of common life in common speech. In France, where the division of the vocabulary into "noble" and "common" (*i.e.*, unfit for poetry) had been made and recorded in dictionaries, the Romantics led by Hugo used the prohibited words whenever they saw fit. Hugo's verse drama *Hernani* (1830) created a scandal in the audience when the heroine was heard to speak of her handkerchief and when a character did not use a roundabout phrase about the march of the hours to say: "It is midnight."

The rediscovery of Shakespeare

The importance of such details can hardly be exaggerated and can perhaps be best understood by recalling what the rediscovery of Shakespeare meant to the Romantics. His rise from grudging esteem, even in England, to European idolatry by 1830 had a significance beyond the one already mentioned of serving to put down French classical tragedy and, with it, French cultural tyranny. The German scholar, critic, and playwright Lessing was among the first to use Shakespeare for that purpose, but the arguments in his theatre reviews, called *Hamburgische Dramaturgie*, sprang from critical genius and not mere national resentment. Shakespeare spelled freedom from narrow conventions—the set verse form in couplets, the lofty language and long declamations, the adherence to verse throughout, the exclusion of low characters, comic effects, and violent action—or, in a word, from the atmosphere of the court and the aristocratic etiquette.

What Shakespeare typified (and not in poetry or plays alone—witness his influence on Berlioz) was the right of the creative artist to invent his own forms, loosen the joints of grammar and metric (or the canons of any art), follow the promptings of his spirit (tragic or gay, vulgar or mysterious, but in any case venturesome), and see where this emancipation from artificial rules led the muse. There was danger in freedom, as always; the conventions ensure safety. But the aim of the Romantic genius was not to play safe or even to succeed; it was to explore and invent, multiply modes of feeling and truth, and thereby revigorate a dead or dying culture. The motto was not common sense but courage. This resolve explains why the men who came to idolize Shakespeare also rediscovered Rabelais and Villon and revalued Spinoza, the lone dissenter who had worshipped a God pervading the cosmos; Benvenuto Cellini, the fearless artist at grips with the principalities and powers; and "Rameau's Nephew," the ambiguous hero of Diderot's posthumous dialogue, a strange figure disturbingly in touch with the dark forces of the creative unconscious.

Drama. With so much feeling astir and so many novel ideas being agitated, it might seem logical to expect a flourishing school of Romantic drama. But only a few isolated works, more interesting than irreplaceable, compose the dramatic output of the Romanticists—Shelley's *Cenci*, Byron's *Manfred*, and Kleist's brilliant pieces in several genres. Ironically, Shakespeare's new role as

emancipator had a curiously paralyzing effect on the theatre down to the middle of the century and beyond. In England, poet after poet tried his hand at poetic drama, only to fail from stage fright at the thought of Shakespeare. On the Continent, various misconceptions about him and old habits of classical tragedy prevented a new drama from coming to life. Victor Hugo's plays contained brilliant verse, and their form influenced grand opera (Wagner's no less than Verdi's), but the fact remained: the dramatic quality could be found everywhere in Romanticist art except on the stage.

Reflection on this point suggests that, quite apart from Shakespeare, the very concern of the Romantics with exploring the inner and outer worlds precluded the writing of great plays. These perhaps require that one world or the other be taken as settled, at least at the start of the dramatic action. Be that as it may, the double reality in flux was the stimulus of all Romanticist endeavour.

Painting. This generality holds for the painters as well; their "reality," too, was by no means "given," so that the notation of fresh detail and the study of new means to transmute the visible into art occupied all those who came after David. Goya led the way in Spain by depicting the vulgarity of court figures and the horrors of the Peninsular War. In England, Constable painted country scenes with a vividness at first unacceptable to connoisseurs. He had to argue with his patron, Sir George Beaumont, about the actual colour of grass. To prove that it was not of the conventional brownish tint used by academicians, he seized a violin, ran out of the room with it, and laid it on the lawn, forcing the unaccustomed eye to perceive the difference between chlorophyll and old varnish. At the same time, Géricault astonished the Parisians by painting, in harrowing detail, "The Raft of the Medusa," not an antique and noble subject but a recent event: the survivors of a shipwreck adrift and starving on a raft.

The new "reality" in painting

The young Delacroix was emboldened by the example and, inspired also by the work of his English friend Bonington, began to paint contemporary scenes of vivid realism—*e.g.*, the Turkish massacre of the Greek peasants at Chios. Later, Delacroix was to visit Morocco (exoticism again) and to discover there the secret of coloured shadows and other pre-Impressionist techniques. His English counterpart, J.M.W. Turner, was pursuing the same goal of realistic truth, though along a different path that nonetheless also led to Impressionism—and beyond. When asked one day why he had pasted a scrap of black paper on a portion of his canvas, he replied that ordinary pigment was not black enough. And he added: "If I could find something even blacker, I would use that."

Sculpture and architecture. No similar transformations of the visual occurred in sculpture or architecture. Canova and Thorvaldsen continued to produce figures and busts on Neoclassical lines; and only Barye, the great sculptor of animals, and Rude, the creator of the Marseillaise panel on the Arc de Triomphe, showed any signs of the new passions. As for architecture, it may have been the love of history that prevented distinctive work. Pugin and Viollet-le-Duc did grasp the principles of what a new style should be, the former's love of Gothic reinstating the merit of framework construction and the latter's breadth of vision as a restorer leading him to predict that iron construction would one day pass from mere utility to high art.

It was actually in railway construction that the seeds of a new architecture were sown. Tunnels and bridges and terminals were needed as early as the mid-1830s, and unassuming engineers such as Brunel and Robert Stephenson set to work to design them. All they had for solving the new and awkward problems of topography, speed, and cost were the ideas they drew from machinery and the vulgar materials they knew how to handle. The results were often remarkable, and they remained to inspire the makers of 20th-century steel and concrete architecture.

Music. It may seem as if the art of music by its nature would not lend itself to the exploration and expression of reality characteristic of Romanticism, but that is not so. True, music does not tell stories or paint pictures, but it stirs feelings and evokes moods, through both of which various kinds of reality can be rendered or expressed. It was in the Rationalist 18th century that musicians rather mechanically attempted to reproduce stories and subjects in sound. These literal renderings naturally failed, and the Romanticists profited from the error. Their discovery of new realms of experience proved communicable in the first place because they were in touch with the spirit of renovation, particularly through poetry. What Goethe meant to Beethoven and Berlioz and what German folk tales and contemporary lyricists meant to Weber, Schumann, and Schubert are familiar to all who are acquainted with the music of these men.

Changes in the relationship between music and literature

There is, of course, no way to demonstrate that Beethoven's *Egmont* music—or, indeed, its overture alone—corresponds to Goethe's drama and thereby enlarges the hearer's consciousness of it; but it cannot be an accident or an aberration that the greatest composers of the period employed the resources of their art for the creation of works expressly related to such lyrical and dramatic subjects. Similarly, the love of nature stirred Beethoven, Weber, and Berlioz, and here too the correspondence is felt and persuades the fit listener that his own experience is being expanded. The words of the creators themselves record this new comprehensiveness. Beethoven referred to his activity of mingled contemplation and composition as *dichten*, making a poem; and Berlioz tells in his *Memoires* of the impetus given to his genius by the revelation not only of Beethoven's and Weber's music but also of Goethe's and Shakespeare's works and—not least—of the spectacle of nature. Nor did the public that ultimately understood their works gainsay their claims.

It must be added that the Romantic musicians—including Chopin, Mendelssohn, Glinka, and Liszt—had at their disposal greatly improved instruments. The beginning of the 19th century produced the modern piano, of greater range and dynamics than heretofore, and made all wind instruments more exact and powerful by the use of keys and valves. The modern full orchestra was the result. Berlioz, whose classic treatise on instrumentation and orchestration helped to give it definitive form, was also the first to exploit its resources to the full, in the *Symphonie fantastique* of 1830. That work, besides its technical significance just mentioned, can also be regarded as uniting the characteristics of Romanticism in music: it is both lyrical and dramatic, and although it makes use of a "story," that use is not to describe the scenes but to connect them; its slow movement is a "nature poem" in the Beethovenian manner; the second, fourth, and fifth movements include "realistic" detail of the most vivid kind; and the opening one is an introspective reverie.

Self-analysis. In this Romantic investigation of the self, some critics have seen little more than excessive ego or, in modern terms, a tiresome Narcissism. No doubt certain Romantic works arouse boredom or disgust with hairsplitting analysis. But the boredom is often due to the fact that after a hundred years the discoveries have staled. When fresh, they came as a revelation; in the works of the great poets and novelists, in Hazlitt's essays and Jean Paul's fictions, and the irony of Byron's letters or Heine's journalism, the truth has not grown dim or platitudinous.

Desirability of self-analysis

It was in any case desirable that this extensive analysis of the self should be attempted then, for only an age in which individualism was both theoretical and passionate could see the logic of the undertaking and act upon it. The logic was this: given the autonomous and unique individual, a search by himself into his moods, motives, fears, and loves must bring forth data otherwise unobtainable. Add these results together, and you have a repertory of clues to the inner life of mankind as a whole. For the uniqueness of each individual is bounded by traits he shares with his fellows, and this common element enables the psychologist to connect and organize the reports of the self-searchers. It is on this hypothesis, incidentally, that the demand for originality in art has

continued unabated since the Romanticists. Forget the model, for there is no such thing; avoid conformity; discover your true self, the buried child; be authentic and sincere—these precepts, which still govern art and criticism, are the legacy of Romantic individualism.

Introspection naturally implies an inner life worth looking into, and most Romantic artists brought forth extraordinary findings. They form the groundwork of modern thought. One cannot easily imagine Freud or Joyce, much less the degree of self-consciousness shared by Western men today, without the deliverances of Blake, Wordsworth, Keats, Leopardi, Stendhal, Benjamin Constant, Sainte-Beuve, Heine, and innumerable other writers of the early 19th century. And towering above them as the creator of the prototype of Romantic introspection is Goethe with his Faust.

Faust was the figure in which a whole age recognized its mind and soul; and the adjective Faustian, as Spengler's use of it makes clear, still describes tendencies at work in culture today. The principal one, already mentioned, self-consciousness—the identity crisis—remains. The belief, moreover, that movement, activity, is better than repose and that striving is better than achieving is clearly the great postulate of contemporary civilization. Faust himself ends by giving his life to practical works in behalf of his fellow man. But he sets himself on that path only after a slow and deep analysis of his divided soul, which has been ruled together or in turn by despair, lust, superstition and the forces of the unconscious, the love of innocence, the conviction of sin and crime, the horrors of hypocrisy and conventional life, the temptations of wealth and power, the disgust with pedantry and established religion, the yearning for infinite knowledge, and through knowledge, peace. Faust, in short, traverses the whole cosmos, made up of the inner and outer worlds, to find in the act of self-dedication to humanity the justification of his existence.

Social and political thought. The Romantics who studied society through the novel or discoursed about it in essays and pamphlets were also devoted to the cause of humanity, but they arrived at politically different conclusions from Goethe's and from one another's. Scott and Disraeli were forerunners of Tory democracy as Burke was of liberal conservatism. Dickens, a passionate humanitarian, stirred the masses with his examples of the law's stupid cruelty, but he proposed no agency of betterment, content to despise Parliament, the law courts, and the complacency of the wealthy. Balzac wrote his "social zoology," *La Comédie humaine*, to show that society becomes a bloody jungle without the church and the monarchy to restrain human passions.

The cause of humanity

Stendhal noted the same reality but was more concerned with the free play of individual genius; he resigned himself to the social struggle, provided not too many stupid men ran the inevitably heavy-handed regimes. Freedom might be found by the happy few through the loopholes of a mixed government such as England's, whereas in the ostensibly free United States there was no protection against social pressure and no likelihood of genius, in art or in politics.

The great authority on American democracy was Tocqueville, whose astonishing survey in two volumes is still packed with useful lessons. Tocqueville confirmed Stendhal's low estimate of freedom of thought in America, but he foresaw in the United States the first example of a type of democracy that would surely overtake the Western world. He found in such a future many good things and many defects; he predicted a day when slavery would threaten disaster to America; he foretold what kind of poetry a democracy would produce and delineated the art of Walt Whitman; he apprehended the complication of laws and the declining quality of justice; but he was reconciled to what must be.

Post-Revolutionary thinking. What lay behind all 19th-century writings on politics and society was the shadow of the French Revolution. In the 1790s the Revolution had aroused Burke to write his famous *Reflections* and Joseph de Maistre his *Considérations sur la France*. They differed on many points, but what both saw, like their successors, was that revolution was self-perpetuating. There is no way to stop it, because liberty and equality can be endlessly claimed by group after group that feels deprived or degraded. And the idea that these principles are universally applicable removes any braking power that national tradition or circumstance might afford.

Proof that the revolution marched on, slow or fast, could be read (as it still can be) in every issue of the daily paper since 1789. In the early 19th century the greatest pressure came from the liberals, whether students, bankers, manufacturers, or workmen enlisted in their cause. They wanted written constitutions, an extension of the suffrage, civil rights, and from time to time wars of national liberation or aggrandizement in the name of cultural and linguistic unity. For example, all the intellect of western Europe sided with Greece in the 1820s when it began its war of emancipation from Turkey. Byron himself died at Missolonghi while helping the Greeks. Poets wrote odes, and painters painted scenes of war. Between liberalism and national liberation the line could not be clearly drawn. In Italy, Germany, Poland, Russia, Spain, Portugal, and South America, revolt in the name of liberty was endemic until the middle of the century. Only England escaped, but it was by a narrow margin, after threats and violent incidents expressing the same animus as elsewhere.

Meanwhile, the first disturbances resulting from machine industry—sabotage, strikes, and conspiracies (for trade unions were generally held illegal)—reinforced the Revolutionary momentum, not only in fact but also in theory. As early as 1810 the business cycle, the doctrine of the exploitation of the worker, and the degradation of life in industrial societies had been noted and discussed. By 1825 the writings of the Comte de Saint-Simon, which proposed a reorganization of society to cure these evils, had won adherents; by 1830 the Saint-Simonians were an acknowledged party with sympathizers abroad, and by 1832 the words Socialism and Socialist were in use.

Industrialization

The Saint-Simonian doctrine relied on a benevolent dictatorship of industrialists and scientists to remove the inequities of the free-for-all liberal system. Other reformers, such as the practical Robert Owen, who organized successful communities in Scotland and the United States, depended on a strong leader using ad hoc methods. Still others, such as Leroux and Cabet, were Communists of divergent kinds seeking to carry out elaborate theories of the perfect state. Proudhon denounced the state, as such, and all private property. As a philosophical anarchist, he wished to substitute free association and contract for all legal compulsions. In England, the school of Bentham and Mill—Utilitarians or Philosophical Radicals—attacked existing institutions in the name of the greatest good for the greatest number, and by their arguments they succeeded in reforming the top-heavy legal system. Without doctrine but moved by a similar sense of wrong, Thomas Carlyle fought the Utilitarians for their materialistic expediency and himself sought light on the common problem by pondering the lessons of the French Revolution and publishing in 1837 what is still the greatest account of its catastrophic course. Later, Carlyle gave in *Past and Present* a suggestive picture of what he deemed a true community: quasi-medieval, based on the Faustian joy of work, and relying for its cohesion on its leader's genius and strength of soul.

Numerous other schemes of reform or revolt could be mentioned. One irony about them is that the name that has clung to most of them is utopian Socialism, which suggests unrealizable visions, whereas the historical fact is that a great many were tried out in practice, and some lasted for a considerable time. As in Carlyle's book, the force of character of one man (Owen was a striking example) usually proved to be the efficient cause of success. Throughout this social theorizing, whatever the means or ends proposed, two assumptions hold: one is that men have a duty to change European society, to purge it of its evils; the other is that men *can* change society—they need only come together and decide what form the

change shall take. These axioms by themselves, without the memory of 1789, were enough to keep alive in European culture the hope and the threat of continuing revolution.

The principle of evolution. Yet it should not be imagined that revolution by force inspired every thinking European. Even if liberals and reactionaries were still ready to take to the barricades to achieve their ends, the conservatives were not, except in self-defense. The conservative philosophy, stemming from Burke and reinforced by modern historical studies, maintained the contrary principle of evolution. Evolution indeed swayed as many 19th-century minds as its rival, and it was sometimes the same minds.

Evolution was the belief that lasting and beneficial change comes about by slow and small degrees. It is often imperceptible and therefore congenial to human habits. It breaks no heads and spills no blood; it is natural, organic. In truth, the idea of evolution is patterned on biology—the slow growth and decay of living things. More than that, evolution in the zoological sense of "descent with modification" had been a recognized speculation among men of science since 1750, when Buffon included it in his *Histoire naturelle*. Lamarck had elaborated the idea at the turn of the 18th century, while Erasmus Darwin, the grandfather of Charles, had by 1796 worked out for himself a compendious theory of similar import. In 1830–33 the geologist Lyell, setting forth the corresponding notion that changes in the Earth take place through the operation of constant and not cataclysmic causes, devoted a chapter to Lamarckian biology—to the evolution of species by imperceptible steps.

Revival of interest in history

As if these teachings were not enough to implant a form of thought, the revival of interest in history made easy and obvious the transition from the world of nature to that of man. It seemed logical to think of both as evolutions and even to liken the state to an organism. Certainly the student of institutions finds them steadily and profoundly altered by minute incidents and variations. Compared to these causes, the violent breaks made by war and revolution seem more superficial and less permanent.

The evolutionary scheme permitted several other beliefs while affording a number of arguments and conveniences. Anyone who had inherited from the previous era a faith in progress could now attach it to this motive power, evolution. Anyone who wished to classify nations or institutions by rank could place them as he thought proper on an evolutionary scale. Anyone who resisted change or wished to speed it up could be admonished with some evolutionary yardstick. Finally, anyone who intended to write a work of history or propaganda found his organizing principle ready-made for him. In the first half of the 19th century, every subject of interest, from costume to the criminal law, was presented in innumerable studies as proceeding majestically at an evolutionary pace.

Another way of stating the influence of this great idea is to say that the mind of Europe had experienced the "biological revolution." Whereas in the 17th century the Newtonian revolution had imposed the model of mechanics and mathematics, what impressed itself on the 19th century was the living organism—change and variety as against fixity and regularity. The logic of preferring "biology" to "mechanics" in an age of individualism, of realism about concrete particulars, and of passionate imagination and introspection need only be stated to be evident.

Science. This is not to say that the science of physics stood still during the Romanticist period. It was the time when the conservation of energy was established and the mechanical equivalent of heat demonstrated. There also prevailed the "physical" pseudo-science of phrenology, which professed to relate individual attributes to bumps and hollows in the skull, and which led to the physical anthropology that defined three, 10, 20, and 100 different races of man by the end of the century. Still, the 19th was more emphatically the century that furnished the theory of the cell (Schleiden and Schwann, 1838–39), which led ultimately to the notion of microscopic creatures responsible for putrefaction and disease and, later still, to cytology and genetics.

It is noteworthy, too, that the 19th century saw the establishment of chemistry on the Daltonian hypothesis of the atom, but it was coloured by the "biological" notion of elective affinities to explain compounds. Goethe, who was an early evolutionist and the scientific expositor of the metamorphosis of plants, called his last novel of human love *Elective Affinities.*

Poetic ideals versus science

On the surface the poetic mind of the age seemed hostile to both science and technology. Wordsworth looks like an enemy of science when he says: "We murder to dissect" and deprecates the man who is willing to "peep and botanize upon his mother's grave." But reflection shows that the animus here is not so much against science as for the science of life. And to make the effort of recapturing the temper of the times is to be at once struck by the precarious status of science itself. Thought in the 18th century had ended in materialism and skepticism. Some writers, such as d'Holbach, had reduced all phenomena to the interaction of hard and unfeeling particles; others, such as David Hume, had "proved" that man can know nothing beyond his impressions and therefore can have no certainty about the truth of cause and effect, on which scientific statements depend. The Romanticist generations could neither agree that life was a concourse of unfeeling atoms nor trust the physicists' assertions based on a law of causation that the most acute thinkers discredited.

Such were the iron constraints within which the famous "crises of the soul" and conversions to religions new or old took place in the '20s and '30s of the last century. Carlyle, Mill, Lamennais, and many others described them in autobiographical works. The choice seemed to be a blind and meaningless universe or a meaning restricted to the motions of matter, equally blind. Even if the latter scheme "explained," it was vulnerable to David Hume's irrefutable doubts.

Philosophy. What enabled 19th-century culture to pursue the scientific quest and regain confidence in spiritual truth was the work of the German Idealistic philosophers, beginning with Immanuel Kant.

Kant. Kant took up Hume's challenge and showed that although we may never know "things as they are," we can know truthfully and reliably the data of experience. The reason for this certitude is that the mind imposes its categories of time and space and causation on the flowing stream and gives it shape. Science, therefore, is not a guess, nor is human knowledge a dream. Both are solid and verifiable. Indeed, certainty, according to Kant, extends as far as morals and aesthetics. The essence of morals is the commandment not to perform any act that one would not want to become a precedent for all human action and always to consider an individual as an end in himself, not as the instrument of another's purpose. The fusion in Kant of ideas stemming from Rousseau and the Enlightenment with ideas fitting the needs of the coming century (Kant died in 1804) made him the fountainhead of European philosophy for 50 years.

Kant's disciples. His disciples—Fichte, Hegel, Schopenhauer—twisted or amplified his teachings. Coleridge in England and Victor Cousin in France adapted to home use what seemed fitting. The school as a whole was known as German Idealism because it relied on the distinction between the thinking subject and the perceived object; "idea" and "thing" were unlike, but idea (or the mind) played a role in shaping the reality of things, from which derived all stability and regularity in the universe.

German Idealism

Stability was desirable as a guarantor of natural science, but to any observer of human history it seemed contradicted by events, especially those since the French Revolution. Hegel, coming after Kant and witnessing Napoleon's victory at Jena in 1806, was therefore impelled to modify the Kantian logic. To account for the great torrent of human history, he imagined a logic of movement, by which opposing forces are in perpetual battle. Neither side wins, but the upshot of their struggle is an

amalgam of their rival intentions. Hegel called the pros and the cons and their survivors thesis, antithesis, and synthesis. Human affairs are ever in dialectic (dialoguing) progression. At times a "world-historical figure" (Luther, Napoleon) embodies the aspirations of the masses and gives them effect through war, revolution, or religious reformation. But throughout the succession of events, what is taking place is the unfolding of Spirit or Idea taking on itself the concrete forms of the real. Hegel's was another version of evolution and progress, for he foretold the extension of liberty to all men as the fulfillment of history. It is interesting to note that until 1848 or '50 Hegel was generally considered a dangerous revolutionary, a believer in an irresistible progress that mankind must earn by blood and battle. Karl Marx, as a younger Hegelian, was to carry out Hegel's unspoken promise on a different base.

Other branches of the all-powerful German philosophy deserve attention but can be spoken of only as they relate to high Romantic themes. Fichte's modification of Kant made the ego the "creator" of the world, an extreme extension or generalization of individualism. At the other extreme, but more in tune with contemporary science and art, Schelling made nature the source of all energy, from which individual consciousness takes off to become the observer of the universe. Nature is a work of art and man is, so to say, its critic, and because human consciousness results from an act of self-limitation, it perceives moral duty and feels the need to worship.

Religion and its alternatives. That need made itself felt ecumenically throughout Europe from the beginning of the 19th century. It had indeed been prepared by the writings of Rousseau as early as 1762 and the pastoral work of John and Charles Wesley much earlier. The atheism and materialism of the late 18th century was in truth far more responsible for the religious revival of the 1800s than the brief secularism of the French Revolution itself, so that when the Catholic writings of Chateaubriand and Lamennais in France, the neo-Catholic Tractarian movement in England, and the writings of Schleiermacher and his followers in Germany began to take effect, their success was due to the same conditions that made Romanticist art, German idealism, and all the "biological" analogies succeed: the great thirst caused by dry abstractions in the Age of Reason needed quenching. Religious fervour, artistic passion, and "gothic" systems of philosophy filled a void created by the previous simple and mechanical formulas.

The political consequences of the religious revivals

The religious revivals, Catholic or Protestant, also aimed at political ends. Their participants feared the continuation in the 19th century of secularism and wholly material plans. In every country the liberals proposed to set up in the name of tolerance ("indifference," said the Christian believers) governments that would serve exclusively practical (indeed commercial) interests. Church and state were to be separated, education was to be secular, which would really mean anti-religious. National traditions would be broken, forgotten, and youth would grow into "economic man," Benthamite Utilitarian man, with no intuition of unseen realities, no sensitivity to art or nature, no humility, and no inbred morals or sanction for their dictates.

Scientific Positivism. The alternatives to the richness of a life lived within traditional religion were two: scientific Positivism and the cult of art. The name Positivism is the creation of Auguste Comte, a French thinker of mathematical cast of mind, who in 1824 began to supply a philosophy of the natural sciences opposed to all metaphysics. Science, according to Comte, delivers unshakable truth by limiting itself to the statement of relations among phenomena. It does not explain but describes—and that is all mankind needs to know. From the physical sciences rise the social and mental sciences in regular gradation (Comte coined the word sociology), and from these man will learn, in time, how to live in society.

Having elaborated this austere system, Comte discovered the softer emotions through a woman's love, and he amended his scheme to provide a "religion of humanity" with the worship of secular saints, under a political arrangement that the sympathetic John Stuart Mill nonetheless described as "the government of a beleaguered town." Comte did not attract many orthodox disciples, but the influence of his Positivism was very great down to recent times. Not alone in Europe but also in South America it formed a certain type of mind that survives to this day among scientists and engineers.

The cult of art. The second "religious" alternative, the cult of art, has had even greater potency, being at the present time the main outlet for spirituality among Western intellectuals. In the Romantic period this fervour was allied with the love of nature and the idolatrous admiration of the man of genius, beginning with Napoleon. A writer as sober as Scott, a thinker as cogent as Hegel, and an artist as skeptical as Berlioz could all say that to them art and its masters were a religion; and they were not alone. At the death of Goethe in 1832, Heine inveighed against the great man's followers who made art the only reality. In the second and third Romantic generations, born around 1820, the religion of art grew still more pronounced and took on an antisocial tone that became more and more emphatic as time passed. "Art for art's sake" ended by signifying, among other things, "art the judge of society and the state." This doctrine was expounded in full detail by the Romantic poet Gautier as early as 1835 in the preface to his entertaining and sexually daring novel, *Mademoiselle de Maupin.* In those pages the familiar argument against bourgeois philistinism, against practical utility, against the prevailing dullness, ugliness, and wrongness of daily life was set forth with much wit and that spirited defiance of all established things which is associated rather with the 1890s and the present day than with Romantic passion. Its occurrence then is but another proof that Romanticism was the comprehensive culture from which later styles, thoughts, and isms have sprung—down to our own time.

"Art for art's sake"

THE MIDDLE AND LATE 19TH CENTURY

During the half century when Romanticism was deploying its talents and ideas, the political minds inside or outside Romanticist culture were engaged in the effort to settle—each party or group or theory in its own way—the legacy of 1789. There were at least half a dozen great issues claiming attention and arousing passion. One was the fulfillment of the Revolutionary promise to give all Europe political liberty—the vote for all, a free press, a parliament, and a written constitution. Between 1815 and 1848 many outbreaks occurred for this cause. Steadily successful in France and England, they were put down in central and eastern Europe under the repressive system of Metternich.

A second issue was the maintenance of the territorial arrangements of the treaties made at the Congress of Vienna in 1815. Metternich's spies and generals also worked to keep this part of the post-Napoleonic world intact; that is, the boundaries that often linked (or separated) national groups in order to buttress dynastic interests. Except in Belgium, the surge of national, as distinct from liberal, aspirations throughout Europe was unsuccessful in the 1830s. But defeats only strengthened resolve, particularly in Germany and Italy, where the repeated invasions by the French during the Revolutionary period had led to reforms and stimulated alike royal and popular ambitions. In these two regions, liberalism and nationalism merged into one unceasing agitation that involved not merely the politically militant but the intellectual elite. Poets and musicians, students and lawyers joined with journalists, artisans, and good bourgeois in open or secret societies working for independence: they were all patriots and all more or less imbued with a Romanticist regard for the people as the originator of the living culture, which the nation was to enshrine and protect.

Merger of liberalism and nationalism

To be sure, this patriotic union of hearts did not mean agreement on the details of future political states, and the same disunion existed to the west, in England and France, where liberals, only half satisfied by the compromises of 1830 and 1832, felt the push of new radical demands: of the Socialists, Communists, and anarchists. Reinforcing these pressures was the unrest caused by in-

dustrialization—the workingman's claims on society, expressed in strikes, trade unions, or (in England) Chartist groups. This cluster of parties agitated for a change that went well beyond what the advanced liberals themselves had not yet won. Add to these movements those that purposed to hold still or to restore former systems of monarchy, religion, or aristocracy, and it is not hard to understand why the great revolutionary furnace of 1848–52 was a catastrophe for European culture. The four years of war, exile, deportation, betrayals, coups d'etat, and summary executions shattered not only lives and regimes but also the heart and will of the survivors. The hoped-for evolution of culture was broken and, with all other hopes and imaginings, rendered ridiculous. The search began for new ways to achieve, on the one side, stability and, on the opposite, the final desperate revolution that would usher in the good society.

For although they seemed decisive, the battles of '48 and after did not, in fact, test the worth of any one idea. Nationalism won and lost in different parts of Europe. Liberalism gained in Italy and Switzerland, but was set back in Germany and France. English Chartism seemed to collapse, yet its demands began to be carried out. The Socialist experiment in France (Louis Blanc's national workshops) also seemed discredited; yet the ensuing regime of Napoleon III made attempts, however clumsy, to deal with poverty by welfare methods. Repression had shifted men and territories—Metternich was gone from Austria—but Germany and France were not on that account flourishing under liberty, equality, and fraternity. There was peace, but war was imminent; and subversive groups continued to plot and publish and frighten the bourgeois, while industry and urbanization contributed their gains at the cost of the now familiar miseries and sordor.

Mood of despondency

In these circumstances the mind of Europe suffered an eclipse, followed by a protracted mood of despondency. Many established or emerging artists and thinkers had been killed or torn from their homes or deprived of their livelihood. Richard Wagner fleeing Dresden, where he conducted the opera; Chopin and Berlioz at loose ends in London, because in Paris music other than opera was moribund; Verdi going back to Milan with high patriotic hopes and returning to Paris in a few months, utterly disillusioned; and Victor Hugo in exile in Belgium and later in Guernsey—all typify the vicissitudes in which men of reputation found themselves in mid-career. For the young and unknown, such as the poet Baudelaire or the English painters who formed the Pre-Raphaelite Brotherhood, it was no time to invite the public to admire boldness and accept innovation. Critics and public alike were all nerves and hostility to subversion. To read Flaubert's masterpiece, *L'Éducation sentimentale*, is to understand the atmosphere in which the first phase of Romanticism ended and its ramified sequels came into being.

Realism and Realpolitik in midcentury. The dominant feeling was that high hopes had perished in gunfire, and this realization bred the thought that hope itself was an error. Any new effort must therefore stay close to the possible, the "real." Realism with a capital *R* and *Realpolitik* together sink their roots in a distrust of man's imagination. This grim caution born of harsh experience coincided with a sense of fatigue that made Romanticist work seem like the foolishness of youth.

The appropriate cultural note must no longer be the infinite or heroic or colourful but rather their opposites. If the commonly accepted term Realism for this reaction of the 1850s is used, it must be with these presuppositions in mind. For the Romantic passion for the particular and exact was a realism, too; it was what Dr. Johnson much earlier had called "vehement real life." The Realism of the disillusioned '50s dropped the vehement, the passionate and, in order to run no risk of further illusion, limited what it called real to what could be readily seen and felt: the commonplace, the normal, the workaday, and often the sordid.

Limitation of "real"

In the same spirit *Realpolitik* rejected principles. The word did not mean "real" in the English sense; in German it connotes "things"—hence a politics of adaptation to existing facts, pursuing plain objects, admitting no obligation to ideals. In this light we can understand the unexpected epithet "scientific" that Marx and his followers bestowed on their brand of Socialism. It was a science not merely because it was presumably based on the laws of history but even more because in its view the advent of the Socialist state was to result from the interaction of things (classes, means of production, and economic necessity) and not, as in earlier socialism, from the will (that is, the imaginative efforts of thinking men). The "objective" appearance given to the new politics of things, Socialist or other, generated that tough, no-nonsense atmosphere, which men then wanted as a source of reassurance in all their dealings.

Scientific Materialism. This search for certainty went with a swinging back of the pendulum in science itself from the vitalism of the previous period to the Materialism of the midcentury. German philosophers derided Idealism and taught the equivalence of consciousness and chemistry: "without phosphorus, no thinking." The machine once more became the great model of thought and analogy—and nowhere more vividly and persuasively than in biology, where Darwin's advocacy of natural selection won the day because it provided a mechanical means for the march of evolution. The struggle for life (Spencer's phrase of 1850, adopted by Darwin in his subtitle) obviously had the requisite "toughness" to convince and, like *Realpolitik*, it followed no principle—whoever survived survived. That Darwin to the very last included other factors in his theory of evolution—Lamarckian "use and disuse" as well as direct environmental forces—carried no weight with a generation bent upon machine certainty. These secondary explanations were ignored, in the usual way of cultural single-mindedness, and for 30 years after the publication of the *Origin of Species* in 1859, an orthodoxy of universal mechanism reigned over all departments of thought.

It prevented the recognition of Mendel's work on genetics; it put religious, philosophical, and ethical thought on the defensive—only what was "positive" (*i.e.*, material) held a presumption of being real and true. The same reasoning produced a school of social Darwinists who saw war between nations and economic struggle among men as beneficent competition leading to the survival of "favoured races"—another phrase from Darwin's subtitle. And by a final twist of logic, Materialism reinforced the moral gloom of the period by casting doubt on both the permanence and the validity of all that was being redefined as "really real." For on the one side, the second law of thermodynamics guaranteed the cooling of the Sun and the pulverization of the cosmos into cold and motionless bits of matter; and, on the other, orthodox "machine-ism" brought its leading prophets, Huxley and Tyndall, to consider men and animals as automatons moved as helplessly as atoms and planets. Consciousness is an epiphenomenon—in plain word, an illusion—precisely as in Karl Marx consciousness and culture are illusions floating above the reality of economic relations.

The "really real"

Victorian morality. To be sure, not everybody in Europe believed or cared about these affirmations. And although ideas long debated do in the end filter down to the least intellectual layers of the population, the time and place of triumph for a philosophy are limited by this cultural lag—a fortunate delay, without which whole societies might collapse soon after the publication of a single book. What kept mid-19th-century civilization whole was a subdued faith in the reality of all the things Realism and materialistic science denied: religious belief, civic and social habits, the dogma of moral responsibility, and the hope that consciousness and will did exist.

The sum of these invisible forces is conveniently known as the Victorian ethos or morality, a formula applicable to the Continent as well as Britain and one whose meaning antedates not only the midcentury revolutions but also the accession of Queen Victoria in 1837. Like Romanticism, this powerful moralism had its roots in the late 18th century—in Wesleyan Methodism, the Evangelical movement, in Rousseau, Schiller, and Kant. Its earnestness was of popular origin; it was anti-aristocratic in man-

ners, and it sought the good and the true in a simple, direct, unhesitating way. Perceiving with warm feeling that all men are brothers under God, the moral man saw that slavery was wrong; and having so concluded, he proceeded to have it abolished by act of Parliament (Britain, 1833).

Such fervent convictions when widely shared exert tremendous power, and it was this concentration of belief and emotion that made Victorian morality for so long impregnable. As Chesterton said of the Victorian painter Watts:

> He has the one great certainty which marks off all the great Victorians from those who have come after them: he may not be certain that he is successful, or certain that he is great, or certain that he is good, or certain that he is capable: but he is certain that he is right.

The sense of rightness generated a sense of power, which the Victorians applied to the monumental task of keeping order in a post-revolutionary society.

Partly by taking thought and partly by instinct, they perceived that the drive to revolution and the sexual urge were somehow linked. Therefore they repressed sexuality; that is, repressed it in themselves and their literature, while containing it within specified limits in society. Further, they knew that the successful working of the vast industrial machine required a strict, inhuman discipline.

Idolatry of respectability

The idolatry of respectability was the answer to natural waywardness. To pay one's bills, wear dark clothes, stifle individual fancy, go to church regularly, and turn aggression upon oneself in the form of worry about salvation became the approved common modes of pursuing the pilgrimage of life.

It could not be expected that everybody would or could conform. From its beginning to the end, the Victorian age numbered a galaxy of dissenters and critics who scorned the conformity, called the religion a sham, and viewed respectability as mere hypocrisy. But the front held, and the massed forces behind it were at their strongest after the multiplied assaults of 1848.

Nothing gives a better idea of the astonishing moral structure called Victorianism than the development of the London Metropolitan Police, begun under Sir Robert Peel in 1829. A lawyer and a former captain who had fought in the Peninsular War were the first joint commissioners and creators of the force. At first they had to weed out the drunks and the bullies who had been the main types of recruit in earlier attempts at policing cities. At first, too, the people both ridiculed and fought with the new police. Gradually, the "peelers" came to be trusted; they remained unarmed regardless of circumstances; they learned to handle rioters without shedding blood; and in the putting down of crime they finally enlisted the public on their side. For something less than a century this unique relationship lasted, in which "law-abiding" and "police" were terms of respect—correlative terms, since the peelers (later "bobbies") could not have become what they were without the self-discipline and moral cohesion of the "respectable."

The upheavals of the midcentury, cultural as well as political, subjected Victorianism to a severe test, for after wars and civil disorders laxity is natural, and ensuing despair induces a reckless fatalism. There was cause indeed for apprehension. When the Great Exhibition of 1851 was being planned on a scale hitherto unattempted, many expressed the fear that to allow tens of thousands from all over Europe to come together under the Crystal Palace in Hyde Park was to invite massive riots. Ministers and heads of state would surely be assassinated. In the event, no protracted assembly of common people and their leaders was ever so quiet and orderly. The moral machinery worked as efficiently as that which was on display under the glass dome.

The advance of democracy. Yet, while a stringent moralism held in check endemic subversion and anarchy, Darwinism and the machine analogy stimulated endless forms of self-consciousness. If man could fashion and continually improve these engines, perhaps he could also engineer an improved society. Because evolution was at last "proved," thanks to Darwin, perhaps it also gave warrant for social and political progress by gradual steps. Spencer's all-inclusive philosophy, likened then to Aristotle's, foresaw an inevitable movement from the simple and undifferentiated to the complex and specialized—as in modern life. Clearly, whether automatons or not, men kept thinking and having purposes; and among evolutionists and scientific socialists alike, thought and purpose included the hastening by voluntary action of what was sure to come by force of natural laws. These and other desires acting in the light of Realism and taking shape in the increasing organization of the toiling masses brought Europe to accept democracy as inevitable.

Evolution as a warrant for social and political progress

The word democracy is used here in a cultural sense. It does not imply a set of political institutions so much as the signs and the agencies that herald the coming populist state of our day: for example, the extension of the franchise, in parliamentary or plebiscite form; the secret ballot; the legalization of trade unions; the rise of a Catholic social movement; the passage of education acts providing free, public, and compulsory schooling; the formulation of Tory democracy as a cure for the evils of free-for-all economic liberalism; the beginnings of welfare legislation (in France under Napoleon III, in Germany under Bismarck); the secularization of life by state action, by the prestige of science, and also by the liberal movements within the churches themselves; and finally, after a decade or so of public education, the great extension and popularization of the press. At the passage of the Reform Act of 1867 in Britain, which gave the vote to urban workingmen, Robert Lowe had said, "Now we must educate our masters." In a parliamentary system the means to that education cannot be the schools alone. The adult "common man" must continually be informed and appealed to for his satisfaction as well as for coherent policy in government. The instrument for this purpose was the new journalism. The quarterlies of the early 19th century gave way to the monthlies in the '60s and they in turn to the weeklies, while the daily papers, costing now but a penny and simplifying all they touched, began to reach the millions.

Realism in the arts and philosophy. In the period of so-called Realism, the arts and philosophy as usual gave both form and substance to the prevailing fears and desires. The mood of soberness and objectivity was alone acceptable, and what art presented to the public confirmed the reasonableness of the mood.

Literature. This interaction accounts for such things as the marked change of tone in Dickens' novels that occurs between *David Copperfield* (1850) and *Bleak House* (1853). The temper expressed in most concentrated form the very next year in *Hard Times* now dominates Dickens' mind and works to the end: life is a dreary sort of underworld; happy endings are artificial conventions not to be believed.

The same mood explains why Flaubert's *Madame Bovary*, which ranks today as the realistic novel par excellence, and is on all counts grim enough in its rendering of boredom and vulgar misery, was judged "too artistic" by some contemporary critics, not close enough to the most common of realities, that of common speech.

Yet the evolution of Flaubert's work remains instructive for an understanding of Realism as a literary creed. Flaubert had begun by writing a highly coloured, imaginative story on *La Tentation de Saint Antoine*, which the author's friends advised him to burn, tone down, or rewrite. Flaubert put it aside and began the novel that became *Madame Bovary*. Its setting was the provincial world around him, not the Egyptian desert; the characters were of the most ordinary type, not an improbable Christian ascetic haunted by visions. But even in the working out of his plain tale Flaubert had to subdue his lyrical Romantic genius to the discipline he had adopted. The description of a rainstorm, for instance, had to be done over and over again so that it would not stand out and be "interesting" by virtue of the observer's mind. It had to be made ordinary and the observer kept outside, just as in science. *Madame Bovary*, begun as a magazine serial, was soon censored by the editor and then prosecuted as immoral by the state. For Flaubert's Realism had gone so far as to

The significance of *Madame Bovary*

portray in no flattering colours the dreary lives and motives of average provincials of both sexes, and the picture violated the rules of the indispensable moralism. What is more, the fate of Flaubert's unhappy heroine symbolized what had happened to the more daring and poetic-glorious time before 1848: as Flaubert said, Emma Bovary was himself.

His novel is thus simultaneously a model and a critique of the new genre—a critique, too, of the state of Europe that produced it. Many other writers between 1850 and 1890 pursued matter-of-factness without this ulterior effect and rendered the details of middling life with such impassiveness and fidelity that to this day many use "realistic" as a synonym for dreary or sordid and regard "the novel" as a reliable historical source. On the precise definition of Realism, George Gissing gave, through a character in one of his own novels, a brilliant commentary: the character is at work on a novel which shall be so true to the dullness of daily life that no one will be able to read it.

Painting and sculpture. The term Realism applies no less to the plastic arts than to literature, but in painting and sculpture it proved difficult to give form overnight to the change of attitude just noticed in literature and political life. The transition between the passionate poetry and drama of Géricault and Delacroix and the Realism of Courbet and Manet was gradual. It came by way of the "open-air" school of Barbizon, whose landscapes seemed arid (at least to the classically trained academic painters of the day) and pointless in the sense that they depicted the commonplace. Still, when the full shock of realism inflicted by the works of Courbet and Manet occurred, it was severe: here were coarseness and violence in manner and subject. Courbet's backgrounds are thick and his people drab; Manet's nude "Olympia" is no goddess nor even a beautiful woman; she is a prostitute, and her name seems like a piece of irony. The portrait of his parents is a painful representation of simple poverty unrelieved by any glow of spirit or intelligence—yet the work itself is beautiful: such was, throughout, the aim and achievement of Realism.

The gradual transition to Realism in painting

In England, by an historical accident, pictorial realism was embodied in subjects that seem far removed from the commonplace. The school that took up the challenge against academic painting and modified the vision of Constable and Turner called itself Pre-Raphaelite. Its members were Holman Hunt, John Millais, and Dante Gabriel Rossetti, and the name they took for their "brotherhood" expressed their resolve to paint like the masters who came before the imitators of Raphael. It is necessary to put it in this clumsy way in order to make clear that Raphael himself was not being condemned, but only his academic followers who introduced "unreality."

To be a Pre-Raphaelite was to see the world with a sharp eye and an undistorting mind and to render it with intense application to solidity of form, bright colour, and natural pose and grouping. All this was to be understood from the motto *Death to Slosh!* In order to make the new virtues vividly clear and also because the Pre-Raphaelites were reared on great literature, their subjects tended to draw upon legend, or Dante, or the New Testament. It was the conception and treatment that constituted the innovation. Everybody could see it, because it went against the habit of "pretty-pretty" illustration. In fact the nominal subject dropped out of sight in the startled response to form and colour. Paradoxically, then, the commonplace subjects of the French realists and the legendary ones of the English Pre-Raphaelites were alike insignificant when compared with the effort to re-create by art the texture and "feel" of actuality—and nothing more. Such was precisely the goal Flaubert pursued and reached in *Madame Bovary*.

Popular art. It hardly needs to be added that this conscious purpose of high art could interest but a relatively small portion of the public and that, for the growing mass of readers of fiction and viewers of art, other kinds of satisfaction were necessary. The ordinary three-volume novel from the lending library and the continued serial in the magazine or newspaper supplied the demand by aping, adapting, and diluting not one but half a dozen literary tendencies, old and new. The number of novels produced in all languages in the 19th century has never been estimated, but it surely must be of the order of astronomical magnitudes. And the whole output was realistic in the sense that it professed to impart the real truth about life. It was contemporary in setting and speech, took the form of a history, and taught its readers how other people lived. The pictorial counterpart was the "chromo," the cheap colour lithograph that illustrated either fiction or news stories in forms which, however false they must seem to a critical eye, again gave the illusion of commonplace reality.

Growth of the novel in the 19th century

Music. At first sight, it would seem as if music were a medium in its nature resistant to Realism. But that is to reckon without the obvious use that music has always made of sounds directly associated with life—church bells, hunting horns, military bands, and the like. In an age when Realism was at a premium, the opera would be the form where these and other associations easily found their place. So it was in midcentury Europe, where Meyerbeer and others provided the effects to suit the fussily "real" staging of all plays, musical or not. Clocks, tables, animals, waterfalls, and especially costume could be relied on to be genuine up to the limit of the possible: live bullets for real deaths were shied away from, and real lightning was out of reach.

But a genius who is often mistakenly grouped with the Romantics, Richard Wagner, supplied this ultimate deficiency—and by musical means. As critics have pointed out, Wagner's system of leitmotives, or musical tags that denoted an object, a person, or an idea, was consciously or unconsciously an accommodation of Realist intent to operatic understanding. This is true not simply because the musical notes "wave" up and down as Isolde waves her scarf at Tristan—a trivial enough device of a sort found in many composers; it is also true in the deeper sense, which constitutes Wagner's unique genius, namely that he was able to compose great music that was steadily and precisely denotative of items in the story by repeating and interweaving their assigned musical tags.

Late-century decadence and doubt. The combination of Realism in art and thought with its cheap derivatives intended for a mass audience gave rise to a characteristic temper, which we find reflected in writers as different as Baudelaire and Matthew Arnold, Henry Adams and Flaubert, Ruskin and Nietzsche. They and many others were exasperated by the banality and smugness of surrounding humanity. It seemed as if with the onset of Positivism and science, *Realpolitik* and Darwinism, realistic art and popular culture, all noble thought and true emotion had been suffocated. The only things that stood out from banality and smugness were their own appalling extremes—vulgarity and arrogance—against which all the weapons of the mind seemed powerless.

Such intellectuals and artists were hopelessly outnumbered not only in the literal sense but also in the means of influencing culture. A newspaper that reached half a million readers with its cliches, its serial story, and its garish illustrations "educated" the people in a fashion that actively prevented any understanding of any part of true culture. The barrier was far more insurmountable than mere ignorance or illiteracy; and it was cutting off not just the populace but also—to use Arnold's terms—the barbarian upper class and the Philistine middle class. Similarly, Nietzsche anatomized the culture Philistine; that is, the person whose mind was being filled by the mediocre thoughts of writers and artists high in repute and generally believed to be leaders of civilization. In other words, the prudent, self-limiting impulse of Realism after 1848 had generated the middlebrow, while the evolution of industrial democracy had generated the mass man. By the late 1880s the gap between this compact army with its honoured officers and common soldiers and the hostile, half-visible avant-garde was a permanent feature of cultural evolution.

The diminished influence of the intellectuals

The individual expressions of disgust with Realism naturally took many forms, but they can be grouped

into half a dozen kinds, not all on the same plane and only some of which have come to bear distinctive names. Some artists retreated from the ugly world into a species of Neoclassicism. Such were the French poets known as Parnassiens. Strict form, antique subjects, and the pose of impassivity constitute their hallmark. In painting, the work of Puvis de Chavannes stands in parallel.

In music, the explicit revolt against Wagner and Liszt, of which Brahms was made the torchbearer, offers similarities, particularly in the desire to learn and employ the "purer" forms of an earlier time. Likewise, the shift in tone and temper of the later poems of Tennyson, Arnold, or Gautier; the resurgence of Thomist orthodoxy in Catholic thought; the haughty detachment in the plays of Becque and those of Ibsen's middle period, all suggest a search for stability, for a fixed point from which to judge and condemn contemporary "progress."

It was also possible, after withdrawing by conscious effort from any struggle with vulgar actuality, to make art itself an independent and inhabitable world and to escape alienation in beauty. This cult of beauty goes by the names of Symbolism and Impressionism. The Impressionists were able to take as subjects some of the sights that most depressed their fellow man and by recomposing them in brilliant, shimmering colour to create a refreshing world of new sensation. Subject once again mercifully disappeared. As Monet said: "The principal subject in a painting is light."

Symbolism and Impressionism

The Symbolists in literature had a more difficult task than the painters, because their medium, words, must be shared with all those who speak the language for ordinary purposes. To disinfect grammar and vocabulary for poetry and "art prose" required severe measures. All set phrases had to be broken up, unusual words revived or common ones used in archaic or etymological senses; syntax had to be bent to permit fresh juxtapositions from which new meanings might emerge; above all, the familiar rhetoric and rhythms had to be avoided, until the literary work, poetry or prose, created the desired "new world." It is a world difficult to access but worth exploring, all its tangible parts being the symbols of a radiant reality beyond—in short, the antithesis of a newspaper editorial.

In music there was no need of any indirect device to establish the mood of Impressionism. It was already to be found here and there in the great Romantics, and when the new generation began to compose on themes drawn from literature, the hints and opportunities needed only a delicate genius to develop them into a style. Debussy was that genius, soon followed by Ravel, Delius, Hugo Wolf, and others. Alike, yet independently of one another, they replaced eloquence, melodic clarity, and harmonic consecutiveness by capricious melodic contour and pointillist chord progressions to produce the shimmer and mystery of musical Impressionism.

To those who dedicated their lives to Symbolist literature and criticism the name of aesthetes is often given, for it was at this time, from 1870 to the end of the century, that questions of aesthetics became the intense concern of artists, critics, and a portion of the public. The old Romanticist phrase "art for art's sake" was revived; many persons not themselves artists confessed to a "cult of art"; much time was devoted—necessarily—to deciphering, commenting, and discriminating, for much of the new art was obscure and arcane and took time to master and enjoy. Other observers called these manifestations decadent—some with contempt, others with pride. For a time, a group of French artists took the name to themselves as a badge of honour.

Aestheticism

But aestheticism was by no means as languid and fatalistic as it tried to appear. At its centre stands the active figure of Oscar Wilde, whom it is easy, because of his notoriety on many counts, to dismiss as colourful but ephemeral. Yet he is far more than a symptomatic figure; he is a representative one, and a good part of his work, namely his entire critical output, some of his fiction, one or two poems, the autobiographical *De Profundis*, and the greatest farce in the language, *The Importance of Being Earnest*, deserves to last. What Wilde accomplished through these works was the liberation of English literature from ancestral (and not merely Victorian) preconceptions. He reconnected England with the Continent artistically by phrasing with finality their different assumptions. He showed that art could be morally responsible only by discarding moralism. In a word, he played again in 1890 the role Gautier had played in France in 1835 with his art for art's sake diatribe in *Mademoiselle de Maupin*. Whoever, starting with Wilde or Gautier, wishes to follow the historical sequence and recapture the atmosphere in which this activity went on, will find no better source than the journal of the Goncourt brothers, the inventors of "art prose" as well as the recorders of contemporary lives, characters, and gossip.

But the reader of their voluminous pages will also find there references to the movement called Naturalism, which does not merely parallel but also intermingles with Symbolism and Impressionism. The Goncourts themselves wrote a number of Naturalistic novels; their friend Zola was the theorist and greatest master of the genre; another novelist, J.-K. Huysmans, passed from Naturalism to Symbolism, as did several other writers. In the poets Rimbaud and Verlaine, as later in the Irish Yeats, the elements of the two tendencies alternated or mixed.

Naturalism

The name Naturalism suggests the philosophy of science, and the connection is genuine. Zola thought that in his great series of novels, *Les Rougon-Macquart*, he was studying the "natural and social history" of a family during the time of Napoleon III. The claim was bolstered by the method Zola used of gathering data like a scientist—every material fact could be proved by reference to actuality or statistics. Naturalism would thus appear to be an intensification of Realism, as indeed it was—more "research." But it differed markedly in spirit. Realism professed to be depiction of the commonplace in a mood of stoicism or indifference—a photographic plate from a camera held almost at random in front of unselected mediocrity; it was, as Flaubert was the first to say, a refusal to share previous Romanticist hopes and interests. Naturalism, on the contrary, readmitted purpose and selectivity. Each novel was a "study" designed to show up and denounce the dismal truths of social existence, for which purpose the worst are the best. Zola's novels throb with a passionate love of life, a life which he showed as tortured and twisted by character and condition. In the end he defined his scientific or "experimental" novel as "nature seen through a temperament."

In the plastic arts, a plausible counterpart of Naturalism is the work of those known as Postimpressionists, notably Cézanne and van Gogh in painting, Rodin and Maillol in sculpture. Their various styles and aims had a common result in restoring solidity and "weight" to the visual object after the fluidity and lightness of Impressionism.

Musical naturalism was, by contrast, an attempt at dramatic literalness. Richard Strauss boasted that he could render a soup spoon in music if he wanted to; but he could not and he did not. The noises of his *Sinfonia Domestica* are standard orchestral sounds fitted with a preliminary explanation, like the libretto or synopsis of a Wagnerian or other opera. When the sheep bleat in Strauss's *Don Quixote*, the clarinets play notes that are decorative on their own account and do not in the least suggest wool. It is rather the thickness of Strauss's orchestration and chromatic harmony that connect him with naturalist doctrine—the headlong embrace with matter. And so it is also in the operas of Bruneau or Charpentier or in the *verismo* of Puccini and the late Italian school generally. Music remains atmospheric; never, except in Wagner's system, denotative.

This definition of Naturalism, coupled with the aesthetic, or "art for art's sake," impetus in Symbolism and with the Impressionists' transmutation of concreteness into light, justifies the name of Neo-romanticism that has been given to the cultural temper with which the 19th century ended. After the glum self-repression of the middle period, it was an outburst of vehement self-assertion, whether directed inward or outward. "Art for art's sake" and Naturalism are indeed but twin branches of one doctrine: art for life's sake.

The final phase of Romanticism

To redefine in this way the cult and the products of art in this period is not to say that the participants in this final phase of Romanticism were necessarily joyful. Love of life and of beauty can coexist with angry despair. The contrast between what man is capable of and what he actually does may prove unendurable; and not everybody can find in art the solace of his woes. Some took to drugs, that "artificial paradise" that Baudelaire had earlier said was indispensable to life in the modern city. Others, with a last confused hope of total transformation, joined those anarchists who thought that a bomb or an assassination would bring down the whole wretched "system."

Still others fled to more primitive parts of the world: Rimbaud to the Near East, Stevenson and Gauguin to the South Seas. Many became Catholic converts—some of them to retreat in monastic orders—the conversion expressing in yet another way a condemnation of "the world." Baudelaire, who died in 1867, just before this period begins, was in fact its true prophet, and his fame coincided with the fulfillment of his word; small as is his output of verse and prose, it affords a complete review of the century. In it is seen the last glow of original Romanticism, the griminess of realistic description, the perversity of Satanism and drug addiction, the impulse to violence and destruction, the intimations of a universal symbolism, the remnants of sincere Christian belief, and the effort to fashion in words a realm for the sole contemplation of beauty.

As these various emotions rioted in the souls of Western men, the revelation of "the Russians" came with Tolstoy and Dostoyevsky. The novel *Fathers and Sons* by Turgenev (1862) had already acquainted Europeans with "moral nihilism," which is the paralysis of the will from an excess of self-conscious mind. But Tolstoy and Dostoyevsky stirred deeper layers by showing that the impossibility of coming to terms with a godless civilization was a general and inevitable state and not a remediable accident. Western facts bore out the conclusion, or seemed to: the hospitals were filled with mental cases, the suicide rate went up with urbanization, everybody was (in the current jargon) "neurasthenic" and distraught. A physician writing in 1892–93 under the pseudonym of Max Nordau entitled his huge best seller *Entartung* (Eng. trans., *Degeneration*, 1895), and after elaborate diagnoses found all the styles and schools of contemporary art one vast confirmation of his social and medical statistics. It was the end of the century, anyway, and the phrase *fin de siècle*, taken over unchanged in every language, exerted a magic spell persuading its users that the numerical close of the 1800s meant also the close and fall of civilization.

ALIENATION AND DESPAIR (20TH CENTURY)

In that same year 1895 George Bernard Shaw said: "France is certainly decadent if she thinks she is." The remark is characteristic of Shaw, but also symptomatic of a new wave of energy. From under the despair and decadence, the scattered retreats and the violent nihilism, the same human strength that produced Symbolist and Naturalist art was trying to reshape the civilization that all found so unsatisfactory.

Attempts at improvement. In England, the Fabians, of whom Shaw was one, were preaching the "inevitableness of gradualism" toward the socialist state. It was they, seconded by the growing strength of the trade unions after a spectacular dock strike of 1889, who paved the way to Labour governments and the British welfare state. Throughout Europe, Socialism was no longer the creed of a lunatic fringe but was the ideal of many among the masses and the intellectuals. The original fight for liberty and democracy in political action had turned into a fight for economic democracy—freedom from want. Laissez faire liberalism had turned inside out, and the liberal imagination at work in the many brands of socialism now demanded state interference to remove the appalling conditions causing all the despair.

The growth of Socialism

Arts and Crafts movement. Among the Socialists belonging to no party, Ruskin and William Morris worked also to effect immediate changes in the quality of their surroundings: they started the so-called Arts and Crafts movement, whose aim was to make objects once again beautiful. Because machine industry produced only the cheap and nasty, they tried by hand to produce the cheap and handsome—good furniture, hangings, and household articles; fast dyes of good colour; well-printed books on good paper; and jewelry and ornaments of all kinds that showed visual talent as well as manual skill. In a word, the movement reinstated the ideal of design and succeeded in forcing it on machine industry itself. Within two decades manufacturers began to hire artists as designers, and by 1910 the 20th-century omnipresence of design, from clothes to print and from gadgetry to packaging, was a *fait accompli*. The visual revolution can be seen easily by looking back with modern eyes to a page of advertising at the turn of the century.

New trends in technology and science. In parallel with the new craftsmanship, the new technology of the 1900s began to give hope of wider improvements. The use and transmission of electric power suggested the possibility of the clean factory, all glass and white tile. Better machines, new materials and alloys, a greatly expanded chemical industry—all supplied more exact, more functional, less hazardous objects of use and consumption, while the application of science to medicine nourished the hope of greatly reducing the physical sores of mankind. Those closest to all these developments were certainly not among the despairers and fugitives from the world. Like all men who struggle successfully with difficulties, they were inspirited by what they knew to be demonstrable progress along their chosen lines.

The same outlook animated workers in the natural and social sciences. It was for both a time of transformation, and genuine novelty exerted its usual invigorating effect. From the '80s onward it had been clear that simple mechanistic explanations based on "dead" matter were inadequate. The Michelson–Morley experiment of 1887 had given the *coup de grace* to the mere push-pull principle by showing that though light consisted of waves, the waves were not in or of anything, such as the ether, which did not exist. Even earlier, James Clerk Maxwell's attempt to work out the facts of electromagnetism on Newtonian principles had failed. And on the philosophic front, the notion of natural "laws" was being radically modified by thinkers such as Poincaré, Boutroux, Ernst Mach, Bergson, and William James. All this prepared the ground for the twin revolutions of relativity and quantum theory on which the 20th-century scientific regime is established.

The breaking away from the machine analogy had its counterpart in the biological sciences. With narrow Darwinian dogmas in abeyance, the genetics of Gregor Mendel were rediscovered, and a new science was born. The fixity of species was again regarded as important (Bateson), while the phenomenon of large mutations (de Vries) caught the public imagination, as the opposite slow, small changes had done 60 years earlier. The mysterious "fitness of the environment" was held deserving of study, and new formulas for evolution—emergent, creative—received serious attention. Vitalism itself once more reasserted its claims, as it seems bound to do in an eternal seesaw with mechanism.

Rediscovery of genetics

The social sciences. Finally, in the social sciences, fresh starts were made on new premises. Anthropology dropped its concern with physique and race and turned to "culture" as the proper unit of scientific study. Similarly in sociology, Durkheim, seconded by Tönnies, Tarde, and Le Bon, concentrated on "the social fact" as an independent and measurable reality equivalent to a physical datum. Psychology, also long under the exclusive sway of physics and physiology, now established at the hands of William James that the irreducible element of its subject matter was the "stream of consciousness"—not a compound of atomized "ideas" or "impressions" or "mind-stuff" but a live force in which image and feeling, subconscious drive and purposive interest, were not separable except abstractly. A last domain of research was mythology, to the existence of which James

George Frazer's *Golden Bough* gave massive witness, thereby exerting proportional influence on literature and criticism.

Re-examination of the universe. The net effect of these innovations in the sciences of man and of nature was liberating. Whatever each specialty or subspecialty meant to its practitioners, the persons who carry in their minds the general culture of an age took the new message to mean that the universe, formerly closed forever, had been reopened, and that on fresh inspection it had proved more alive than dead. These conclusions offered a field for strong individual action. Subjective choice became again respectable as a genuine fact. Truth, God, beauty, and immortality obviously were in need of re-examination.

In philosophy, politics, and criticism this re-examination may be called the pragmatic revolution; in social and moral life, the liquidation of Victorianism. But the Pragmatic Revolution must not be thought of as being only the work of those who, like James, called themselves Pragmatists. Nietzsche, Samuel Butler, Shaw, Bergson, and others constitute the headwaters of the stream of thought that issues in present-day Existentialism. The common features are the turning away from absolutes and unities to pluralisms and the method of testing by consequences. Subjective and objective tests looking to future thought and action—not authority or antecedents—are to decide the true, the good, and the beautiful.

Turning away from absolutes

Such an outlook, of which the refinements are, like the defects, beyond the scope of this article, is the logical and appropriate one for an age of reconstruction. It boils down to trying all things new and holding fast to that which is good. But it presupposes the creation of new things to try, and here it is allied to the liquidation of Victorianism. In morals the work of destruction generally begins by affirming the opposite of the accepted rule. An excellent source book for this attitude is Samuel Butler's *Way of All Flesh*, written in 1885 but not published till 1903. The Victorian Tennyson had said: " 'Tis better to have loved and lost than never to have loved at all." Butler said: " 'Tis better to have loved and lost than never to have lost at all." This inversion of values —don't weep over loss; there are plenty of loves to be had and the more the merrier—is but an indication of method. At first the denial was uttered as humour and paradox: Butler's *Note-books*, Shaw's *Arms and the Man* (the soldier wants chocolate, not ammunition), Wilde's *Importance of Being Earnest*, Jarry's *Ubu roi*, Strindberg's tragicomedies—to cite but a few subverters of the Victorian—all used derision and topsy-turviness to make their point.

But underneath the joke was the new purpose, which soon found open expression in positive utterance and action. In the plays of Hauptmann and Brieux, the novels and anticipations of H.G. Wells, the essays of Tolstoy, Péguy, Georges Sorel, Ellen Key, Havelock Ellis, Unamuno, Ortega y Gasset, or Shaw, the new modes of feeling and the new scale of virtues and vices are set forth with as much earnestness and vigour as the old Victorian kind.

Nor did action wait until all the books were out. From the onset of the overturn, say 1890 onward, the rebellion was a biographical fact. People braved public opinion and got divorced, lived together unmarried, practiced and preached contraception, studied the psychology of sex, and defended homosexuality. Or again, the sons of the rich turned Socialist, became labour leaders, and fomented syndicalist (*i.e.*, direct action) strikes, while the daughters demanded the vote as suffragettes, assaulted policemen, and went to jail for chaining themselves to door handles. Meanwhile, students rioted about international incidents or university affairs; schools were subjected to the devastation of the softer pedagogies; "rational clothing" exhibited itself in spite of derision, like the bicycle and the newfangled automobile; and new cults multiplied like mushrooms—outdoor sports, nudism, Theosophy, esoteric Buddhism, Rosicrucianism, New Thought, the Society for Psychical Research, Christian Science, the Salvation Army, and the "Maximinism" of Stefan George.

Of these, hardly any need explanation here. But a word must be added about Theosophy if only because of its historical importance in developing Yeats's genius and for expressing once again the attraction that the "wisdom of the East" has for Westerners. Not that the doctrine elaborated by Madame Blavatsky rested on any genuine knowledge of Hindu religion and philosophy. That is not its point. The point is rather that Theosophy supplied the need for quietude, mystery, transcendence, and immortality in the wearied souls of Europeans. In Theosophy the doctrine of reincarnation offers satisfaction of immortal longings and inspires to wisdom, the demands of which are periodically revealed by mahatmas, or holy men.

Theosophy

As for the poet Stefan George's worship of his young friend Maximin, who died at 15, it answers a similar impulse to permanent truth but with the additional urge to abolish (rather than escape) "contemporary materialism." George was but one among many European writers who wanted to found a new society in place of the actual one. What has fitly been called the politics of cultural despair fastened on a great many saviours as the new hope—monarchy, "integral nationalism," a new aristocracy (usually tinged with intellect), technocracy (rule by science and the engineers), the proletariat (in syndicalist "cells" or communist collectives), trade and professional guilds federated in a corporate state, or again the mystic unity of "blood" and "race." In all these creeds, at least at their beginnings, the thirst for the ideal is evident; together they formed a new utopianism, of which the later fruits are familiar but quite other than those predicted: Soviet and Chinese Communism, Italian Fascism, German National Socialism. Beyond them, in the present, elements of this second wave of cultist thought and practical utopianism that ushered in our century are easily discerned.

In one country, as the 19th century passed into the 20th, all the energies seeking an ideal found an unexpected outlet. The occasion for battle was the conviction of a French officer for espionage; *i.e.*, the Dreyfus affair. Its cultural suggestiveness is apparent: on one side, the ideal of justice and the regard for the individual as an end in himself; on the other, the social or collective ideal typified by the army and the nation; throughout, the ideal of truth—the facts—pursued, lost, and found again in an embittered struggle that threw up a host of endemic prejudices—about race, about class, for and against intellect—to say nothing of individual egotisms and obsessions that had been charged with the force of pent-up aggression.

The prewar period. The same universal aggressiveness was to have its field day in the coming war of nations, but in the intervening decade (1905–14) occurred the remarkable outburst of a creativeness, which, for the first time since 1789, had its source elsewhere than in Romanticism. The "Cubist decade" (as it has been conveniently called) gave the models and the methods of a new art, just as the natural and social sciences had begun to do for science a little earlier. Cubism in painting defined itself as a new classicism, but it was obviously not Neoclassical. In painting and sculpture, in music and poetry, and in architecture especially, the new qualities were: simplicity, abstraction, and the importance of mass.

The Cubist decade

Since the artists at work were numerous and diverse, in both origins and purpose, it is idle to try to derive their productions from any single set of conditions. Clearly, the Cubist painters found encouragement in the canvases and ideas of their immediate predecessors, the Postimpressionists. But the new architects found it rather in new materials—reinforced concrete and the steel constructions of the engineers. Some of the poets (Futurists, Unanimists, Simultanists) were influenced by sociological theories about the group; others (and novelists too) read the new psychologists and philosophers. The sculptors responded to African and Javanese artifacts. And the musicians perceived at once the exhaustion of a system of

harmony, the impact of the crowd and the primitive, and the opportunities offered by using and making sounds hitherto ignored by composers. In a culture as cosmopolitan and quick-witted as that of Europe before 1914, it is unlikely that any important concept or speculation should not have found its mark somewhere in art, even if the individual artist did not consciously record it.

Consequently, it is not arbitrary to say that the Cubist decade was stimulated and prepared for its characteristic work by everything that was pouring in upon it—motion pictures and the airplane, X-rays and organ transplants, the philosophy of time and space in relativity and Bergsonism, wireless and the Eiffel Tower, the Jamesian "stream of consciousness" and Le Bon's analyses of the crowd, the importations of colonial art and the European discovery of Walt Whitman, the myths in Frazer and the double personalities in the psychopathic clinics, the fever of revolution and of sexuality reacknowledged, the transcendental visions of various orientalisms and the vitalist current that was reanimating the cosmos and authorizing an infinite number of geometries. Multiplicity outside, bottomless depths within, equally validated by science and intuition, furnished unlimited scope to perceptive genius; it was only a question of pondering, choosing, and creating fit forms.

The impact of war. To the many-sided productiveness the war of 1914 put an instantaneous stop. It was a war of a sort Europe had not known since 1815—the nation in arms. And at that earlier time, the absence of large industry had precluded the involvement, physical and mental, of every adult citizen simultaneously throughout Europe. In 1914 Beethoven, Hegel, and Goethe would have been in the trenches.

The death of Renaissance high civilization

The cessation of cultural activities; their replacement everywhere by a propaganda of hate; the rapid decimation of talent and genius in the murderous warfare of bombardment and infantry assault; the gradual demoralization through four years of less and less intelligible war aims; and after the Armistice, the long sequel of horrors—starvation, dispersion, disease, and massacre—together shattered the high civilization born of the Renaissance and based on the idea of the national state. Too many able men and women had been killed at the start or at the peak of their creative powers for the continuity of culture. Too many intimate faiths and civil traditions had been ground down for any recovery of self-confidence and public hope to be possible. Skepticism about this conclusion is easily met by reference to the present-day view of "the human condition" as it is reflected in novels and plays.

The period "between wars"—1920 to 1939—gave sufficient indications that forward direction had been lost. It combined a frivolous neo-decadence (if the term may be allowed) with a restless search for starting points. The artists who survived were unable to pursue their individual development from the point of its interruption, whether physical or spiritual. Many felt a kind of shame to be busy with art after the human disaster they had witnessed. This feeling accounts for the tiresome mocking of oneself and one's work that marks the production of the 1920s and '30s. Nothing being important, everything is silly, and it can be offered at all only if one laughs at it first. Hence the appearance of Dada, the meaning of which is that the thing is a harmless and even childish obsession.

More serious minds—Stravinsky's, for example—looked farther back than prewar times for possible models: the 18th-century or the late Middle Ages. Playwrights took up the ready-made of antiquity, like 17th-century "ancients," but turned the plots inside out or upside down, those acrobatics mirroring their times and their souls simultaneously. T.S. Eliot, after expressing his disgust at all he saw and felt, declared himself a royalist, an Anglo-Catholic, and a Neoclassicist. In all the arts, eclecticism and, even more, allusiveness were the preferred methods. The partakers' enjoyment came from recognizing the borrowed plums. To this device the sole alternative was to deny the validity of art itself, as in Surrealism, which relied for its deliverances on automatic writing uncensored by the conscious mind. The ultimate creator was believed to be the dark side of our common nature.

Culture since 1920. It is always true, of course, that as the observer comes nearer to his own times his blindness increases. When he has said his say, it turns out that something or someone remarkable has been overlooked—Blake or Kierkegaard, the aesthetics of Poe and Whitman, or the scientific truths of Mendel or Semmelweis—all delayed by the myopia of the ignorant and the well informed. Therefore, any generalities made about the period since 1920 are and ought to remain suspect for a good while. Still, they are not without value for the present for two reasons: they are suggestive and they allow for a sifting through of a variety of tenable views. In addition, such conclusions inform later comers of how things seemed before time had cleared away the rubble.

The most important truth about culture since 1920 is the so-called cultural lag. The meaning of the term is simple: culture in any sense at all is being made somewhere and it takes five, 10, or 20 years to become known and diffused. The contributions of Freud and Einstein, for example, did not enter the public domain until 1920, though clearly perceivable by 1905. Subtracting five years for war and peace leaves a net lag of a whole decade. Again, there is the North Atlantic cultural lag. Until the first war sent American youth to Europe and conscious expatriation followed, the lag between the continents was 20 years or more. But since 1920, in both Europe and America, a different kind of cultural lag has been noticeable. It is the joint result of swift social evolution and failure to move creatively beyond the Cubist decade.

The "cultural lag"

The social evolution has consisted in bringing new layers of the population in contact with cultural affairs. The mass media have done the work of popularizing—radio, television, magazines have reached the millions and told them in words and pictures "all about art"; while display advertising has quietly adapted the tone and techniques of modern literature and painting to charm and seduce the customer. The schools, including the universities, have likewise "gone modern" and bred a taste for novelty in place of the classics. But with each enlargement of the public exposed to these influences, the revelation has been less and less up-to-date. It is still necessary in Europe and America to exhibit and explain Baudelaire, who died more than a century ago. His continuing relevance is unquestionable, both because he follows upon what the popular mind has previously absorbed and because in a certain sense European culture has taken no great step beyond him.

The summing up can only be that except for the work of the Cubist decade, the 20th century has not fashioned a culture of its own. It has solaced and entertained itself with historical mixtures and retrospects and what the 1960s and 1970s believe to be the newest of the new is repetition half-disguised: anti-art is all in Futurism; madness and the absurd are in Jarry and Unamuno; jazz, serial, and electronic music (and *musique concrète* as well) antedate 1914; "ecology" and consumerism are full-fledged in Ruskin and Shaw; youth movements, feminism, sexual freedom, and the liberated school are today the second viewing of a former script; existentialism merely amplifies; even science fiction and the spy story as emblematic of the possible and the actual are old stories. What is new in the 20th century is the magnification of all the predicaments and the blinding concentration on details.

The enlargement of scale has obvious causes—more people, more "communication" (which clogs its own channels and bewilders wider constituencies), more apprehension resulting from World War II and its prospects—nuclear fire and biological blight. Fear is a cultural force, too; it inspires wholesale rejections and temporary hopes, and when large numbers of people have at least an inkling of what is and what was, the moods of favour and disgust are violent and short. Schools of

thought and their leaders are taken up and dropped with quick boredom and panic rapidly. Persons and things are "in" or "out" with scarcely an argument for or against. But the cultural lag remains for want of definitive assimilation. Although all profess to desire a world of reason—that is, of peace and plenty, of brotherhood and equal rights—the prevailing dogma is that Western civilization, heartless because of rationality, deserves to perish.

The means proposed to put an end to it is a general emancipation and loosening of all bonds. Language (or "free speech") is to be purposefully obscene; art is to be "disposable" (that is, impermanent and stripped of meaning, even aesthetic); manners and dress must be improvised, like schooling—and indeed society had best be "deschooled" lest something fixed be transmitted, which would clearly be oppression. The cultural war of the '60s against the Establishment was in truth not against any caste or clique in power that might deserve the capitalized name; it was against anything established, including a rule or a fence set up with the very consent of those who tore it down.

The drive to self-destruction

Whether or not such an attitude lasts, it is permissible to see it as part of a larger tendency affecting all European culture—the drive to self-destruction. It could be argued that for 80 years or more, the preponderance of effort has been aimed at making an end of the aristocratic, elitist high culture of Renaissance man and his successors. That that culture has democratized itself and brought to the fore all the ideas of liberty, equality, social responsibility, humaneness, world unity, and cosmic brotherhood has not helped to save it. It is, many believe, being destroyed from within in the name of the very ideas it forged in the teeth of opposition and promulgated for the good of its present destroyers. Nor is this destruction the work of healthy barbarian appetite; it is led by the heirs of the great tradition, including the best artists of the time, who sense, share, and express the hatred requisite to annihilate. As Jarry said as early as 1900: "We will bring down everything to ruin—and then we will go on and destroy the ruins."

In such a state of affairs, only one supposition may be possible: European man has thrived for half a millennium on the principle of individualism and has exhausted all its possibilities. A new type of man, actuated by a different principle, is required to cope with the conditions created by his predecessor. If he arises, a new culture—art, science, philosophy, social and moral relations, political state, and personal metabolism—will follow. Meantime, the conscious and unconscious prophets of his coming are busy at their proper work, which is to undo, to pulverize, and to level the ground for the unimaginable creature of tomorrow.

BIBLIOGRAPHY. EUGEN WEBER, *Paths to the Present* (1960), is a full and well-organized collection of documents and extracts covering the period from Romanticism to Existentialism. For the corresponding historical narrative, the *Columbia History of the World*, ch. 66–101 (1972), together with E.J. HOBSBAWM, *The Age of Revolution, 1789–1848* (1962), is useful. HAROLD T. PARKER, *The Cult of Antiquity and the French Revolutionaries* (1937, reprinted 1965), is an excellent introduction; as is CRANE BRINTON, *The Political Ideas of the English Romanticists*, rev. ed. (1962). JACQUES BARZUN, *Classic, Romantic, and Modern*, rev. ed. (1962), fills out the interpretation given in the present article. ROBERT C. BINKLEY, *Realism and Nationalism: 1852–1871* (1935); and C.J.H. HAYES, *A Generation of Materialism: 1871–1900* (1941), amplify the political and economic narrative and also interpret culture. EGON FRIEDELL, *Kulturgeschichte der Neuzeit: Die Krisis der europäischen Seele von der schwarzen Pest bis zum Weltkrieg*, 3 vol. (1927–31; Eng. trans., *A Cultural History of the Modern Age*, 3 vol., 1930–32, reprinted 1952–54), is uneven in pace but full of brilliant insights. G.M. YOUNG, *Victorian England* (1936), paints the portrait of an age. JACQUES BARZUN, *Darwin, Marx, Wagner*, rev. ed. (1958), takes up more than the three figures named and outlines the thought of mid-century Realism and of its critics. HOLBROOK JACKSON, *The Eighteen Nineties* (1913, reprinted 1971), is lively and comprehensive, but should be read in conjunction with ARTHUR SYMONS, *The Symbolist Movement in Literature* (1899, reprinted 1971). For a 19th-century view of the literature the century itself produced, GEORG BRANDES, *Hovedstrømninger i det 19de Aarhundredes Litteratur*, 6 vol. (1872–90; Eng. trans., *Main Currents in Nineteenth Century Literature*, 6 vol., 1901–05), is unsurpassed and may be supplemented by PAUL MILIUKOV, *Outlines of Russian Culture*, pt. 2–3 (1942; orig. pub. in Russian, 1898–1903). For all scientific subjects the best synthesis will be found in WILLIAM DAMPIER, *A History of Science and Its Relations with Philosophy and Religion*, 3rd ed. rev. (1942), but can be usefully supplemented by the brief essay by J.D. BERNAL, *Science and Industry in the Nineteenth Century* (1953); and two books by BANESH HOFFMANN: *The Strange Story of the Quantum*, 2nd ed. (1959), and *Einstein* (1972). JOHN BOWLE, *Politics and Opinion in the Nineteenth Century* (1954), offers a careful survey, the reading of which should be followed by H. STUART HUGHES, *Consciousness and Society: The Reorientation of European Social Thought, 1890–1930* (1958); FRITZ STERN, *The Politics of Cultural Despair* (1961); and JOSE ORTEGA Y GASSET, *La Rebelión de las Masas* (1930; Eng. trans., *The Revolt of the Masses*, 1932). On the earliest influences on Romanticism, see A.M. OSBORN, *Rousseau and Burke* (1940); A.B. COBBAN, *Edmund Burke and the Revolt Against the Eighteenth Century*, 2nd ed. (1960); E.H. WRIGHT, *The Meaning of Rousseau* (1929, reprinted 1963); and the symposium by PIETER GEYL, *Napoleon, voor en tegen in de franse geschiedschrijving* (1946; Eng. trans., *Napoleon, For and Against*, 1949). For studies of writers and their contribution to the modern temper, see ERICH HELLER, *The Disinherited Mind*, 3rd ed. (1971); WALTER SOKEL, *The Writer in Extremis* (1959); and KARL LOWITH, *Von Hegel zu Nietzsche*, 3rd ed. (1953; Eng. trans., *From Hegel to Nietzsche*, 1964). Three overlapping but diversely oriented surveys of the period since the 1880s are G.P. GOOCH, *A History of Our Time, 1885–1914*, 2nd ed. rev. (1946); FELIX GILBERT, *The End of the European Era: 1890 to the Present* (1970); and J. KIM MUNHOLLAND, *Origins of Contemporary Europe: 1890–1914* (1970). For the plastic arts, see BERNARD S. MYERS, *Modern Art in the Making* (1950); WYLIE SYPHER, *Rococo to Cubism in Art and Literature* (1960); and JOHN GOLDING, *Cubism* (1959). On music, the most compact and stimulating single volume is CECIL GRAY, *The History of Music*, rev. ed. (1947), to be followed by his essays on modern composers in *Predicaments; or, Music and the Future* (1936) and *A Survey of Contemporary Music*, 2nd ed. (1929). CONSTANT LAMBERT, *Music Ho!* (1934), adds valuable interpretations as does ANDRE HODEIR, *La Musique depuis Debussy* (1961; Eng. trans., *Since Debussy*, 1961), for the ensuing period. For the philosophic and social ideas underlying the culture of the last hundred years, see A.C. MCGIFFERT, *The Rise of Modern Religious Ideas* (1915); and the two anthologies: WALTER KAUFMANN (ed.), *Existentialism from Dostoevsky to Sartre* (1956); and D.O. WAGNER (ed.), *Social Reformers*, rev. ed. (1941). Interpretations from various points of view are found in GERALD HEARD, *Morals Since 1900* (1950); A.N. WHITEHEAD, *Science and the Modern World* (1926); LLOYD R. MORRIS, *William James: The Message of a Modern Mind* (1950); NIKOLAUS PEVSNER, *Pioneers of Modern Design, from William Morris to Walter Gropius*, rev. ed. (1964); CAROLINE PLAYNE, *Society at War* (1931); ARTHUR MARWICK, *The Deluge: British Society and the First World War* (1965); and COLIN WILSON, *The Age of Defeat* (1959). In addition, two important collections, *The Popular Arts*, ed. by IRVING and H.A. DEER (1967); and *Modern Culture and the Arts*, ed. by J.B. HALL and BARRY ULANOV (1967), contain the views of some 75 writers, painters, musicians, and critics on their several disciplines and the present state thereof.

(J.Ba.)

European Diplomacy and Wars (c. 1500–1914)

This article is intended to provide a broad narrative of the diplomatic and military history of Europe from the beginnings of the European state system at the end of the 15th century to the outbreak of World War I in 1914. The end of the 15th century roughly marks the transition of Europe from a congeries of petty feudal interests under the theoretical rule of the Holy Roman emperor and the pope into a system of competing sovereign nation-states. Henceforth, relations between the nations of Europe must be seen through a prism of constantly shifting national alliances entered into for fear that one nation might become too strong and upset the "balance of power." For nearly four centuries concern for the preservation of the balance of power engendered an almost endless series of European wars. But that concern also exerted a restraining influence that kept the

states of Europe from complete mutual destruction. Indeed, between 1500 and 1914 Europe achieved a mastery over the globe. In 1914 the tension of the balance-of-power system exploded in fratricidal war. The period of European hegemony came to a close, and the European focus of balance-of-power wars and diplomacy gave way to international relations on a world scale.

This article is divided into the following sections:

I. About 1500–1648
The Italian wars, 1494–1516
The struggle between the Habsburgs and the French
Religious wars and dynastic struggles prior to the Thirty Years' War
II. 1648–1814
Europe after the Thirty Years' War
The Age of Louis XIV
Peace and war, 1713–89
The French Revolution and Napoleon
III. 1815–1914
The Age of Metternich and the challenge of 1848
Revision of the settlement of Vienna
The impact of economics on politics
The Bismarckian period, 1870–90
The new imperialism
The German challenge

I. About 1500–1648

THE ITALIAN WARS, 1494–1516

The advent of the balance of power. Between the onset of invasion of Italy in 1494 by King Charles VIII of France and the advent of general peace in 1516, a great political revolution occurred in Europe: the transfer of the balance-of-power system in Italy to all Europe. Since the mid-15th century, a balance of power had been established among the five large Italian states that sought to dominate the peninsula—the Venetian Republic, the duchy of Milan, the Republic of Florence, the Papal States, and the Kingdom of Naples. With the creation of the Venetian League in March 1495 against France, this system of political equilibrium became a European-wide system—one that was to last for four centuries.

The balance-of-power system was first defined by the Venetian Humanist Francesco Barbaro in 1439 and was further discussed by the Florentine historian and statesman Francesco Guicciardini in the early 16th century. Italian metaphors found entrance into the diplomatic language of Europe: "balance of power," "equilibrium of forces," "equal balance of powers," "counterweight of forces"—all metaphors deriving from the use of scales.

The system was based on recurring patterns of political behaviour. Whenever one state threatened to become too powerful, the other states formed a coalition to topple it. Thus, governments worried not only about neighbouring countries with common borders but also about distant states, the progress of which made them potentially dangerous, even in the remote future. States claimed the right to intervene either in disputes between two or more states or within each state to support a given political or religious party against the government of that state. It therefore became necessary to know at every moment the intentions and available strength of each government. Toward that end, the states of Europe began to send permanent ambassadors abroad—a Venetian custom imitated throughout Europe. These ambassadors, together with their hosts of secretaries and agents, bribed statesmen and civil servants, hired informers, supported opposition groups, organized conspiracies; in short, they were privileged spies. To lessen the chance of war, the states tacitly agreed upon a number of rules that, in time, came to be considered international law. To re-establish the balance of power, reversals of alliances became commonplace.

The growth of formal diplomacy

The medieval concept of a united and stratified Christian Europe ruled from on high by the pope and the Holy Roman emperor gave way to the idea of a system of states based on the principle of equal rights. These rights formed "the freedoms of Europe." They did not, however, create a politically united Europe. They simply meant, in law, the independence of each state and, in fact, the opportunity for the stronger states to dominate the weaker ones.

The great powers and Italy. From 1494 to 1516 Italy became the battleground of Europe. All the great powers of Europe had interests in Italy. In 1486 the Habsburg Maximilian I of Austria ("king of the Romans," *i.e.*, the emperor-elect of the Holy Roman Empire), possessor of Austria, Styria, Carinthia, Carniola, the Tirol, Trieste, Vienna, and the city of Trent, wanted control of the Alpine valleys that opened onto the Adriatic and the Po Valley. He dreamed of receiving the imperial crown in Italy from the hands of the Pope and of joining northern and central Italy (including the Papal States) to the empire.

Ferdinand II the Catholic, king of Aragon, held Sardinia and Sicily, and the western Mediterranean was virtually a Catalonian lake. As a kinsman of King Ferrante of Naples, an Aragonese, Ferdinand had rights to the Kingdom of Naples. The Spaniards—Aragonese as well as Castilians—enjoyed a privileged position in Italy, thanks to their influence with the Pope; and many Spaniards were appointed bishops or government officials in Italy, thus paving the way for Spanish penetration.

The Swiss Confederation, too, sought to control the approaches to the Alpine passes, particularly those leading toward the states of Milan and Venetia, in one direction, and toward Burgundy, in the other. The Swiss peasants were formidable soldiers; their armies of pikemen and harquebusiers (musketeers), their light artillery, which they used in combination with infantry, and their battle formations of compact squares had revolutionized military tactics in the late 15th century. The Swiss victory over Charles the Bold, duke of Burgundy, in 1476–77 marked the beginning of the end of the dominance of heavily armed cavalry.

Swiss military tactics

The French, also, had diverse interests in Italy. The House of Anjou had laid claim to the Kingdom of Naples since 1266, when Charles of Anjou, brother of King Louis IX, had conquered the kingdom. Charles was expelled by the Aragonese in 1282, but the Angevins kept trying to re-establish their authority. In 1398 Louis, duc d'Orléans, brother of King Charles VI of France, married Valentina Visconti, daughter of Gian Galeazzo Visconti, duke of Milan. By this marriage the Orléans family inherited rights to the Duchy of Milan. The Sforza family had usurped power in Milan, but the House of Orléans had not renounced its heritage. King Charles VI had purchased Genoa in 1396. The kings of France continued to be protectors of the Genovese Republic and acted as mediators between the Genovese and their enemies. They had given Genoa to the Sforzas on condition that the latter remain vassals. The kings of France were also protectors of the Duchy of Savoy in order to assure free transit from Italy to France through the Alpine passes. At the same time, however, they also protected the Marquis of Saluzzo and the Marquis of Montferrat against the power of the Duke of Savoy.

The Italians did not resent these foreign intrusions into their affairs. They were accustomed to them and were even glad to call upon the help of their powerful neighbours to the north. Despising these "barbarians," the Italians felt that with their political adroitness they could always make use of the foreigners. But a great change had occurred without their realizing it: by the end of the 15th century, the great powers of Europe had become national states, the territories, resources, and internal unity of which far exceeded those of the regional states of Italy. What is more, those states had terminated their far-flung civil and foreign wars. France had put an end to the Hundred Years' War (1337–1453) against England, the Burgundian Wars against Charles the Bold, and the internal wars against the insurgent barons. Spain, unified by the marriage of King Ferdinand of Aragon and Queen Isabella I of Castile, had finally expelled the Moors from the country by conquering Granada in 1492. Now the great powers were capable of acting effectively in Italy.

Charles VIII's invasion of Italy. Charles VIII, at the head of his army, left Lyons for Italy on July 27, 1494.

Some historians claim that he was forced to invade the Italian peninsula by compelling economic interests in his own kingdom. In point of fact, the young king of 22, an avid reader of romances of chivalry, dreamed of the glory of the knights errant and of a crusade against the infidel, as well as of the riches of the Orient. Charles's invasion of Italy in support of the ancient Angevin interests was, in Charles's eyes, merely the first step in the launching of a great crusading venture. By yielding the conquests of his father, Louis XI, Charles had obtained the neutrality of the great powers. He returned Artois to Maximilian of Austria, heir of Mary of Burgundy, gave back Roussillon and Cerdagne to Ferdinand of Aragon, and gave large sums of money to King Henry VII of England.

Dissension among the Italian states and internal squabbles of their political parties accomplished even more than the French army's obvious superiority in heavy attack cavalry and siege guns to afford Charles VIII an easy passage through Italy. He took Florence, Rome, and, finally, Naples on February 22, 1495. The King entered the city bearing the imperial globe, the symbol of imperial authority, inscribed with the titles "king of Sicily and of Jerusalem" and "emperor." Though the inscription referred to the Byzantine Empire, all Europe took it to mean the Holy Roman Empire. The kings grew uneasy. The Germans were offended, inasmuch as they had had custody of the Holy Roman Empire ever since the time of Otto I (960). Ferdinand of Aragon was fearful of Sicily's fate and worried also about his rights to Naples.

Formation of the Holy League

Hence, on March 31, 1495, the Venetians, Maximilian I, Ferdinand of Aragon, Ludovico Sforza "the Moor" (duke of Milan), and Pope Alexander VI formed the Holy League (Venetian League) in order to restore peace and Italian independence. The League was the first coalition to maintain the balance of power in Europe. The allies massed an army. Since the person of the king of France embodied the nation, Charles VIII had to be brought to safety. Leaving an army in Naples, he began a homeward retreat. At Fornovo di Taro (July 6, 1495), Charles fought off the League's army, which strove to cut off his retreat as he crossed the Apennines. The Marquis of Mantua, general of the League, claimed victory. Charles VIII, however, returned to Lyons on November 7, 1495.

Gilbert, comte de Montpensier, whom Charles had left in command at Naples, had to fight on three fronts: against the Aragonese army led by a general of genius, Gonzalo Fernández de Córdoba (El Gran Capitán); against the Neapolitan barons; and against the common people of Naples. The French won all the pitched battles, but guerrilla warfare tore them to shreds. In June 1496 the remnants of the French army surrendered at Atella. The King of Naples returned to his kingdom, and Charles VIII signed a truce with Ferdinand. Soon thereafter, in 1498, he died accidentally.

Louis XII's invasion of Italy. His successor, Louis XII, was of the House of Orléans. On his accession to the throne, he assumed the titles of king of France, king of Naples, and duke of Milan. Since Ferdinand and Maximilian had been the winners against Charles, Louis XII soon formed allies to restore the balance of power. He came to an understanding with Pope Alexander VI, who did not relish the Spanish inroads into Italy. Louis XII married off the Pope's son Cesare to Charlotte d'Albret, whose father disputed Ferdinand of Aragon's right to the Kingdom of Navarre; Cesare was made duke of Valentinois and a *grand seigneur* ("great lord") of France in 1499.

Louis also made an alliance with the Swiss, who then were at war with Maximilian of Austria. At Lucerne, in 1499, the Swiss gave him permission to recruit from their infantry, the best in Europe. Louis next entered into an alliance with Venice, enemy of the Sforzas who, in turn, were vassals of Maximilian, and even with Philip I the Handsome, son of Maximilian, and the duke of Burgundy (who now ruled over the Low Countries), as well as with Henry VII of England.

Louis retook Milan and captured its Sforza ruler Ludovico (il Moro) at Novara (1500), sending him as a prisoner to France, where he died in 1508. Louis XII initiated the organization of the duchy of Milan as if it were a French province. He thought he would achieve a political master stroke by sharing the Kingdom of Naples with Ferdinand of Aragon. Together, their armies conquered the kingdom; the King of Naples was taken prisoner and sent to France. At the same time, Louis sent a fleet into the Aegean to launch a crusade against the Turks.

With the death of Pope Alexander VI in 1503, Louis XII's chief minister, Bishop Georges d'Amboise, posed his own candidacy for the papacy. The election of d'Amboise however, would have put the French in too strong a position, and the majority of the cardinals voted at first for Pius III, who died 22 days later. The cardinals then elected Giuliano della Rovere, who became Pope Julius II (1503).

At that moment, Ferdinand of Aragon, no longer worried about the Aragonese of Naples, thanks to his collaboration with Louis XII, broke with the latter. D'Amboise kept the French relief army too long in Rome to back his papal candidacy. The Aragonese navy controlled the sea, preventing the French troops in Naples from receiving provisions. Gonzalo Fernández de Córdoba instituted military reforms in Naples that led to the formation of *tercios*, or infantry regiments. The French lost Naples in 1503. By March 1504 the Spaniards had taken the whole Kingdom of Naples.

French loss of Naples

Against the victorious Spaniards, Louis XII now tried to reverse alliances by making an agreement with Maximilian I. By the three treaties of Blois, Louis was to marry off his 18-month-old daughter, Claude de France, to the future emperor Charles V, then only 12 months old, who was the grandson of Maximilian and the son of Philip the Handsome and Joanna the Mad. Claude's dowry was to comprise the French provinces of Brittany, Burgundy, the county of Blois, the duchy of Milan, and the Kingdom of Naples. If Louis XII should produce no sons, all these territories would ultimately remain with Charles.

The King's council, refusing to sanction this act of folly, forced the King to sign a will in May 1505 whereby his daughter would marry François d'Angoulême, the future Francis I. In May 1506 Louis summoned the States General of the Realm. The latter insisted on the proposed marriage with d'Angoulême "because Monsieur François is completely French." As a result, the Blois treaties were broken.

The coalition against France. Pope Julius II sought to humble Venice, which was expanding at the expense of the Papal States, and then to expel all foreign powers from the peninsula. Julius formed the League of Cambrai (March 1508), which was ostensibly directed against the Turks but was in reality aimed against Venice. Joining the League were Ferdinand of Aragon, Emperor Maximilian, Henry VII of England, and, to the surprise of everyone, Louis XII, who turned against his Venetian allies because they had been delinquent in carrying out the agreements of 1499.

It was Louis himself who pulled the chestnuts out of the fire for the coalition. The French army defeated the Venetians at Agnadello in 1509; Venice gave back to the Pope the towns in Romagna it had seized. Then, on October 4, 1511, the Pope formed the Holy League against France, with Venice, Spain, and the Swiss led by Matthäus Schiner, cardinal of Sion. The Swiss, eager to acquire the Duchy of Milan, felt that Louis XII was not paying them enough and was not taking their interests sufficiently into account. Although an outstanding French general, Gaston de Foix, won several victories, notably at Ravenna on April 15, 1512, the Swiss attack against the French lines of communication forced the French army to retreat.

Then, exploiting religion for political ends against the Pope, Louis XII forced the convocation of an "ecumenical" church council at Pisa. Pope Julius II was ordered to appear before the council to be tried as an enemy of

Christianity. On April 21, 1512, the council suspended Julius from his holy office. Julius countered by summoning another council at the Lateran on April 19, 1512. This council declared that the pope alone could call an ecumenical council.

Julius excommunicated those who claimed that such a council had higher authority than the pope. He condemned the Gallican religious articles of the Kingdom of France, which were embodied in the Pragmatic Sanction of Bourges (1438). At this juncture Pope Julius II died, in February 1513. The French were driven out of Italy, but the other "barbarians" remained. The Spaniards dominated the Italian peninsula. Contrary to his own intentions, Julius II had worked for them.

Spanish control of Italy

Louis XII remained isolated against the coalition. The Swiss gained a decisive victory over the French at Novara (June 1513). They then invaded Burgundy and laid siege to Dijon. The English, in turn, landed at Calais in July 1513, but the distance involved made it difficult to supply their armies. The belligerent powers negotiated. Louis XII, a widower, married the young Mary Tudor of England (October 1514) but died on January 1 of the following year.

The various kings were short of money and none felt like fighting for the Swiss or the Spaniards. Louis XII's successor, Francis I, was therefore able to break up the coalition. He bought peace with England and negotiated the betrothal of Louis XII's second daughter Renée to young Charles of Austria, duke of Burgundy, who had just come of age. Francis renewed his alliance with the Venetians and was able to defeat the Swiss at Melegnano (September 13–14, 1515). The League powers were forced to negotiate. Pope Leo X, a Medici, proffered Francis an alliance with Florence (October 1515) and signed the Concordat of Bologna with him (August 18, 1516). Charles, having become King of Spain by the death of Ferdinand of Aragon, concluded the Treaty of Noyon (August 13, 1516): by its terms, Francis I renounced all his rights to Naples in favour of his infant daughter Louise, who was betrothed to Charles. In November 1516 Francis I made a "lasting peace" with the Swiss, which gave him the right to recruit soldiers in the Swiss cantons. By the Treaty of Brussels (December 1516), Maximilian returned all his conquests to the Venetians. By the Treaty of Cambrai (March 1517) between Francis, Maximilian, and Charles, the powers agreed to settle their differences and to undertake a crusade against the Turks.

The Treaty of Cambrai

The effect of the Italian Wars on Europe. With the signing of these treaties the Italian Wars had, in effect, come to an end. They had been decisive for Europe. Policy had expressed and strengthened a change in human attitudes: the natural laws emanating from God were forgotten; the specific laws described in the Old and New Testaments were neglected. Foreign policy henceforth took into account only the desire for domination and, to satisfy it, only the most effective means, whatever those means might be. The shifting relationship of forces was constantly under assessment. The Italian Wars had exposed to broad daylight the ruin of European unity and the division of Europe into jealous, antagonistic states and nations, each striving its utmost to gain pre-eminence in Europe. These wars had transferred the pattern of Italy to that of Europe as a whole: instead of a Christian republic under the jurisdiction of emperor and pope, there now came into being a system of states in balance, with equality and freedom, in principle, but, in fact, the freedom of the strong to dominate the weak. Europe still had a chance for unity in its adherence to a universal church, which might serve as a rallying point for a crusade against the Turks. But, within a generation after the Treaty of Cambrai, the unity of Christian Europe was permanently shattered by the Protestant Reformation, which completed the division of Europe along religious and ideological lines (see REFORMATION).

THE STRUGGLE BETWEEN THE HABSBURGS AND THE FRENCH

During the turbulent period from the imperial election of 1519 to the Peace of Cateau-Cambrésis in 1559, the House of Austria and the House of France contended for mastery of Europe. In the course of this struggle, the King of France—at first the weaker contestant—sought help against the Emperor from the Protestant princes of Germany and even resorted to an alliance with the infidel Turks. As a result, the last chances for the achievement of European unity were shattered, and the subsequent European balance of power was built on the divisions of Europe.

The election of Charles V. Maximilian of Austria, king of the Romans, died on January 12, 1519. According to the Golden Bull of 1356, his successor was to be chosen by seven Imperial electors: three of them ecclesiastics, the archbishops of Mainz, Trier, and Cologne; and four of them secular rulers, the Count Palatine of the Rhine, the Duke of Saxe-Wittenberg, the Margrave of Brandenburg, and the King of Bohemia. Since 1438 the emperors had been chosen from the House of Habsburg, and it seemed only natural that Maximilian I's grandson Charles of Ghent (duke of Burgundy [1506], king of Spain [1516], and archduke of Austria [1519]) should be elected.

From 1516 onward, however, King Francis I of France was also a candidate for accession to the imperial throne. He was acting in accordance with an old French tradition. Since the time of Philip the Fair (reigned 1268–1314), the kings of France had proclaimed themselves emperors in their own kingdom and had laid claim to imperial privileges (1302), and a long succession of French princes had been candidates for the imperial throne. In the late 15th century Louis XI had emphasized the sacred and "imperial" nature of French royalty and had made the cult of Charlemagne compulsory in France, declaring January 28, the date of Charlemagne's death in 814, a national holiday. But Francis I had another reason for his candidacy: if Charles of Ghent, whose possessions already ringed France, were elected emperor, his enormous power would threaten the very existence of the French kingdom.

The election of 1519 gave rise to much hard bargaining, in the course of which both Francis I and Charles bribed the electors. But German national sentiment ran strongly against Francis, who was unable to gain the support of the Fuggers, the leading family of bankers of Germany. Instead, the Fuggers lent large sums of money to Charles. Francis then promoted the candidacy of a German, Frederick of Saxony, head of the opposition to the Habsburgs. But Francis's action was too late. Though Frederick was elected unanimously on June 27, 1519, he abdicated three hours later, after being intimidated by a pro-Habsburg army that had surrounded the electoral hall. On the following day, Charles was elected Emperor Charles V.

Charles V—sober, lonely, introspective, and religious—took seriously his duties as emperor and as heir to the House of Burgundy. He considered himself superior to the kings and princes of Europe, whose policies he continually tried to control. Moreover, he believed that, as their commander in chief, he was to lead the armies of Europe in a crusade against the heathen, especially the Turks. On the other hand, his duty as head of the Holy Roman Empire was to restore its territorial integrity—to regain the territories that had previously belonged to it but that had been torn away during earlier struggles for power. Among these were several French provinces: the Dauphiné, Provence, Burgundy, Flanders, and the Duchy of Milan.

Charles's Burgundian inheritance

As heir to Mary, duchess of Burgundy and wife of Emperor Maximilian, Charles also inherited the vast Burgundian realm once ruled over by John the Fearless, Philip the Good, and Charles the Bold—"the grand dukes of the west." Much of Burgundian inheritance had been lost to Louis XI of France at the end of the 15th century, but Charles V considered it an act of reverence to his lineage to claim the entire Burgundian inheritance, which included Franche-Comté and the Duchy of Burgundy, Flanders, Artois, and the towns on the Somme River (at that time France's strategic boundary). Charles

V found himself in a strong position to wrest territory from France. With installations in several Alsatian territories, in Franche-Comté, and in Spain, he had begun the encirclement of France.

The contest between Charles V and Francis I. As France and the empire prepared for war, both wooed the English, who could prove a powerful ally: the English had extensive commercial ties with Flanders and supplied wool for the Flemish cloth industry; as masters of Calais, the English controlled the entry into France. Francis I, moreover, had aroused the jealousy of King Henry VIII of England by the lavishness of his entertainment at the meeting of the two sovereigns on the Field of Cloth of Gold in 1520. Henry therefore allied himself with Charles and in 1522 invaded Boulogne and Picardy and even threatened Paris. Charles also received the support of the Pope, who sanctioned the restoration of the Kingdom of Naples to the imperial crown because of Charles's active support of the Roman Church against the Reformation. (In 1521, at the Diet of Worms, Charles banished Luther and his followers from the Holy Roman Empire, ordering governing authorities and theological schools under their jurisdiction to approve all writings bearing on the Catholic faith.) Before the close of 1521 Charles had seized Tournai and his generals had reconquered the Duchy of Milan. Charles V claimed the Duchy of Burgundy and, in his will of May 22, 1522, asked that he be buried in the Carthusian church at Dijon beside the dukes of Burgundy who had gone before him. To focus his efforts more effectively, he entrusted his brother Ferdinand with the Austrian territories of the House of Habsburg.

In 1523 Charles V negotiated an alliance with Francis I's chief vassal, the Constable of Bourbon, seigneur of central France. In 1524 the Constable invaded Provence and laid siege to Marseilles in order to ensure connections between Aragon and Italy. His defeat and subsequent retreat allowed the French to reoccupy Milan and besiege Pavia. Called upon for help, the imperial army routed the French at Pavia on February 24, 1525, and took Francis I prisoner.

French defeat at Pavia

Contrary to Charles V's expectations, France did not collapse. Francis' mother, Louise of Savoy, together with the King's council, maintained order, saw to it that the armies were paid, and kept up the resistance. Charles's victories had upset the balance of power, and the states of Europe turned against him. Pope Clement VII, the Venetians, and Henry VIII of England signed with the French a first treaty of alliance, the Treaty of Moore, in August 1525. In Germany Charles V was unable to enforce the edicts of Worms or of Nürnberg (1522). He made preparations for the Diet of Speyer, scheduled for 1526. But the outlook seemed unfavourable inasmuch as he needed the votes of the Lutheran cities and Lutheran princes John of Saxony and Philip of Hesse for his proposed war against the Turks, who, under Süleyman the Magnificent, were readying an invasion of Hungary. Thus, on January 14, 1526, Charles V signed the Treaty of Madrid with Francis I. By that treaty Francis I was to "restore" Burgundy and its dependencies to Charles V, renounce all rights to Naples, Milan, Asti, and Genoa, relinquish his suzerainty over Flanders and Artois, pardon the Constable of Bourbon and his supporters, provide the Emperor with a fleet and an army for his coronation at Rome and his crusade, marry Charles V's sister Eleanor, assume all of the Emperor's financial obligations to Henry VIII of England, swear to keep his pledges on his honour as king and knight and on his faith as a Christian—and also leave as hostages the two sons he had had by Claude of France: the nine-year-old dauphin and the four-year-old Charles, duke of Orléans.

On March 17, 1526, Francis I left Spain. As early as January 13 he had sworn before a notary that he had signed the Treaty of Madrid under duress, and he declared it null and void. He asserted that he did not have the right to dispose of the Duchy of Burgundy without the consent of that province. The deputies of Burgundy, who visited the King at Cognac, protested that he could not part with his domains, and they wished to remain obedient subjects of the French crown. On May 22, 1526, Francis I formed the League of Cognac with the Pope, the Venetians, and the leading Italian princes. He renewed relations with the English. He also inaugurated a policy that envisaged striking at the Habsburgs from the rear. He entered into negotiations with Sigismund I, king of Poland, and John (János) Zápolya, voivode (governor) of Transylvania. Then, shutting his eyes to deep-seated differences in religion and culture, he took the unusual step of negotiating with the Muslim leader of Bosnia, to induce the latter to invade the Habsburg territories of Styria and Carniola in the Balkans. Finally, he made an alliance with the Ottoman sultan Süleyman. The latter, after capturing Belgrade in 1521, advanced into Hungary. He won the Battle of Mohács, where the Hungarian king Louis died in combat, and attacked Vienna, although without success (1529). Charles V's brother Ferdinand was elected king of Hungary by one Hungarian diet; but another Hungarian diet named Francis I's ally John Zápolya.

Charles V's situation was in jeopardy. His army, commanded by the Constable of Bourbon, had taken Rome, where the Constable was killed (May 6, 1527). The Emperor's Lutheran mercenaries had looted and pillaged the world capital of Christendom. Thereupon, Henry VIII and Francis I declared war on the Emperor (January 23, 1528). Charles V, outraged by what he considered treachery on the part of Francis, challenged him to a duel. The French invaded Milan and Naples. But Francis could not pay a high enough price for Andrea Doria's Genovese fleet, which went over to Charles V. Soon lack of money forced the two monarchs to negotiate. Margaret of Austria, the Emperor's aunt, and Queen Mother Louise of Savoy arranged the so-called "Ladies' Peace," the Treaty of Cambrai on August 3, 1529. The Treaty of Madrid was confirmed but its terms were softened. The French retained Burgundy, the cities on the Somme, and the counties of Guines and Ponthieu—that is, a protective screen for Paris. Francis granted the County of Charolais to the Emperor for life. He promised gold *écus* (crowns) to get back his children, who were returned to France in July 1530. He married Eleanor.

The "Ladies' Peace"

Charles V, in turn, renounced his Burgundian dream but remained, nevertheless, the leading monarch of Europe. He was master of Italy, where he maintained a secure hold over the territories of Milan and Naples. In Florence, Alessandro de' Medici was restored by Charles as ruling duke and made to marry the Emperor's illegitimate daughter Margaret. Thenceforth, Alessandro became one of his faithful followers. On February 22, 1530, Charles V had received the iron crown of the Lombard kings and two days later, on February 24, he had been crowned emperor by the Pope. On January 1, 1531, the Augsburg Diet, at his bidding, elected his brother Ferdinand as king of the Romans. Thus, he could pursue his dream of forcing the King of France into submission.

French anti-Habsburg alliances. Faced with this threat, Francis I negotiated feverishly with all those opposed to the Emperor. Appealing to the concept of the balance of power in Europe, Francis gradually outlined a policy of protecting Europe's "liberties" against the Emperor—a policy that was to become a constant in the policy of the French kings until the time of Louis XIV. Francis first came to terms with Henry VIII of England, who was hoping to annul his marriage to Catherine of Aragon and wed Anne Boleyn. Henry, needing Francis' backing in dealing with Pope Clement VII, signed the Treaty of Boulogne (October 20, 1532) with Francis, ostensibly against the Turks. The Anglo-French alliance deteriorated, however, when Henry VIII cut the Gordian knot, marrying Anne Boleyn and proclaiming a national church separate from Rome.

The Treaty of Boulogne

For his part, Francis I came to an agreement with the Lutheran princes of Germany. At the Diet of Speyer in 1529, the Lutherans had protested against the imperial decree of April 29, 1529, which guaranteed the Lutherans the principle of "*cujus regio, ejus religio*" (*i.e.*, the religion of a region or city would be that of its prince or municipality) but with tolerance for the Catholics. The

Lutherans did not want tolerance. On June 25, 1530, they presented to the Augsburg Diet their "Confession of Faith" approved by Luther—the Augsburg Confession. It had been signed by the elector John of Saxony, his son John Frederick, George, margrave of Brandenburg-Culmbach, Francis and Ernst, dukes of Brunswick-Lüneburg, Philip, landgrave of Hesse, Prince Wolfgang of Anhalt, and the delegates from the free cities of Nürnberg and Reutlingen. At the end of December 1530, the Protestant princes, together with 15 free cities, formed the Schmalkaldic League and took up arms against the Emperor. The Protestants looked to support from England and France. Charles V tried concessions to his traditional enemies and persuaded the Pope to summon a general church council. Several treaties that Francis I signed with the Protestants enabled him to recruit Protestant mercenaries, who were first-rate soldiers. But he could not derive maximum advantage from this alliance because of his suspicions vis-à-vis the French Protestants, his understanding with the Pope, and his alliances with the Turks.

Francis I also sought an alliance with the Medicis, who controlled the strategic Florentine seaports on the route of the Spanish galleys. Pope Clement VII, also a Medici, was seeking a royal marriage for his niece Catherine. An interview at Marseilles between the Pope and Francis led, on October 28, 1533, to the marriage of Prince Henry and Catherine de' Médicis.

Francis I and Charles V both sought alliances with the Muslims. In 1529 Charles V sent Jean de Baldi to Persia to try to provoke a Persian invasion of the territories of Süleyman the Magnificent. Charles had also tried to make an alliance with Moulay Ḥasan Agha, king of Tunis. Francis, on the other hand, strove above all to forge an alliance with Süleyman, who possessed the best army and best galley fleet in Europe. His armies could take Vienna—and thus the Holy Roman Empire—from the rear. His fleets could ravage the coasts of Italy and, above all, cut Charles V's lines of communication in the western Mediterranean.

The core of Charles V's armies was made up increasingly of Spanish soldiers. His navy consisted mainly of unseaworthy galleys that had to stick closely to the coasts. The best method of bringing Spanish troops into Italy, Burgundy, and Flanders was by way of Barcelona to Genoa, thence through Milan, the Alpine passes, Savoy, Franche-Comté, and Lorraine or, alternatively, via the St. Gotthard Pass, the Rhine Valley, and the principality of Liège. If the Barcelona–Genoa route were cut—an easy feat for the French with their bases on the shores of Provence—then reinforcements would have to journey along the coast of North Africa and then up the west coast of Italy.

Franco-Turkish Alliance

Francis therefore tried for an Ottoman alliance with Süleyman. The Sultan had appointed as admiral (captain pasha) the Algiers pirate Barbarossa (Khayr ad-Dīn). The latter was plundering the coasts of Spain and had seized Tunis in 1534, from which he threatened Malta, Sicily, and the Kingdom of Naples. In 1535 Charles V recaptured Tunis and installed a Moorish king there as his vassal. But in the process he had freed 20,000 Christian prisoners and had thus retained his lustre as head of Christendom against the infidels. Then, Francis I endeavoured to send Barbarossa against Genoa, and in February 1536 he signed a trade agreement with Süleyman that gave France commercial mastery in the Near East. Thereupon, Charles V invaded Provence and attempted to take Marseilles to keep open the Barcelona–Genoa route. As Charles advanced, the French laid waste to the country. Hunger and disease decimated Charles' army, forcing him to withdraw in late summer 1536. The French armies occupied Savoy and Piedmont and treated them as French provinces. Barbarossa, after occupying Bizerta, ravaged the Adriatic. Süleyman threatened Italy. Charles V was forced to come to terms with Francis.

Pope Paul III intervened to unite all the Christian forces against the Turks. On July 14, 1538, the two monarchs restored friendly relations at Aigues-Mortes, where they feasted together with a great show of amity and reconciliation. Charles then passed through France to repress the rebellious people of Ghent. In early January 1540 he stopped at Paris, where he was entertained lavishly.

A condition of the agreement between Charles V and Francis I was that the latter's second son, the Duke of Orléans, marry the Emperor's daughter Maria and receive the duchy of Milan as a fief from the Emperor. The territory of Milan, however, was indispensable for the passage of Spanish troops, and in October 1540 the Emperor gave it as a present to the Infante Philip (later Philip II of Spain). Chagrined, Francis I again became Charles V's adversary. John Zápolya, who had died on July 22, 1540, had left behind an infant son. Ferdinand, the Emperor's brother, sought to exploit the situation by capturing Buda. Francis I aided Sultan Süleyman in the latter's efforts to make peace with Venice (October 1540). On July 30, 1541, Süleyman won a notable victory at Buda and conquered eastern and southern Hungary, while the Emperor failed in an attack against Algiers. On July 2, 1541, Francis I's envoy to Süleyman, Rincon, was assassinated in the Po Valley on orders of the Marchese del Vasto; a year later, on July 12, 1542, Francis I used this as a pretext to declare war. He had William of Cleves, duke of Cleves and Gelderland, and his allies—Denmark, Sweden, and Scotland—attack the Low Countries. But he made the mistake of diverting the main body of his forces against Perpignan, a town in southeastern France, near the Spanish border, where he suffered defeat. Charles V, triumphant in 1543, restored Gelderland and Zutphen to the Low Countries, where 17 provinces were made into one entity, and the Pragmatic Sanction of 1549 defined uniform rights of succession for the region. The Diet of Speyer (1544) with the Danes allowed the latter to levy tolls on ships passing through adjoining straits.

Barbarossa's Turkish fleet, joining with the French fleet, took Nice from the Duke of Savoy (August 1543). Toulon now became a base for the Turkish fleet, which launched assaults on the coasts of Spain and Italy but missed a chance to capture Genoa (1544). Süleyman welcomed the French fleet at Constantinople with great pomp. In Italy, the Duca di Enghien defeated the imperial troops at Ceresole in Piedmont (April 13–14, 1544). But Charles V invaded Champagne and, by way of the Marne, came as far as Château-Thierry, while Henry VIII besieged Boulogne and Montreuil with a view to marching on Paris by way of Saint-Quentin and the Oise. The French followed a "scorched-earth" policy of systematic devastation. The imperial army, short of food, had to retreat. Charles V left Henry VIII in the lurch and negotiated at Crépy (September 18–19, 1544). He gave up Burgundy. Charles, duke of Orléans, was to marry either the Emperor's daughter Maria, with the Low Countries and Franche-Comté as a dowry, or Ferdinand's daughter, with Milan as dowry. These were the "alternatives." Charles V was given four months to decide. Francis I was to give his son the duchies of Orléans, Bourbon, Châtellerault, and Angoulême as an appanage (*i.e.*, feudal grant for future inheritance), renounce his suzerainty over Flanders and Artois, and restore Savoy and Piedmont to their duke. But Charles V could not make up his mind. The duke of Orléans died in 1545, and the treaty lapsed.

Francis I next sent his Scottish allies against Henry VIII (who had captured Boulogne) and even attempted an invasion of England. Shortage of funds forced Henry VIII to negotiate. By the Peace of Ardres (June 7, 1546) Francis I had to promise to abandon his Scottish allies and pay 800,000 crowns over eight years for the return of Boulogne.

Like the rest of Europe, the German Protestants had been shocked by the alliance between the French king and the Ottoman sultan and were further disturbed by the reconciliation between the King and the Emperor and by the summoning in December 1545 of the Council at Trent, which called for the reform of the Roman Catholic Church and a new offensive against the Reformation.

The Council of Trent

The Protestants therefore offered to mediate between Henry VIII and Francis I. Francis and his second son, Henry, used the occasion to renew relations with the German Protestants. Charles V, however, banished the Landgrave of Hesse and the Elector of Saxony from the empire; and, with Ferdinand I, he sought forcibly to crush the Protestants who had banded together in the Schmalkaldic League. He negotiated with the Turks so that he could bring all his forces to bear against the rebels, who were now backed by the Czechs. Francis I made preparations to come to the aid of the insurgents, but he wanted Henry VIII's prior agreement. But the latter died in January 1547, and Francis himself died a month later. In April Charles V won the Battle of Mühlberg against the Schmalkaldic League, while Ferdinand I, for his part, crushed the Bohemian revolt. The Emperor was now able to assemble a Diet of Augsburg, which, on May 15, 1548, published the 26 articles of an "Interim" proclamation, which ran counter to the religious formulations established by the Council of Trent. Although the proclamation re-established church jurisdiction, worship of the saints, and the seven sacraments, it authorized communion in the two rites and marriage for priests—at least until a council could meet, the formulas of which the Emperor could accept. The Emperor was now at odds with the Pope on matters of discipline, not dogma, and he disavowed the council, which had moved to Bologna and then came to a halt (1549).

Flushed with victory, Charles V dreamed of restoring the hereditary imperial crown to his family. After lengthy family negotiations, it was decided in 1551 that, after Charles' death, Ferdinand I would become Emperor and would have the Infante Philip elected king of the Romans; the latter, in turn, would bring about the election of Ferdinand's son Maximilian, husband of one of Charles V's daughters and regent in Spain during Philip's stay in the Low Countries (1548–50). Thus, the imperial crown would pass alternately to the two branches of the Habsburgs.

But to the Protestants of Germany, the Augsburg Interim seemed to threaten "Germanic freedoms," and, when the Protestant city of Magdeburg was banned from the empire in 1549, they turned to the King of France as the "protector" of those "freedoms." The new French king, Henry II, had betrothed his son Francis to the six-year-old Mary Stuart, queen of Scots. Mary was also the niece of the powerful French Guise family, and the threat of a Franco-Scottish alliance compelled the new English king, Edward VI, to give Boulogne back to Henry II for a small sum. With peace with England assured, Henry II was able to lend his support to the German Protestants. After consulting among themselves, the German princes decided, on January 15, 1552, that Henry II should occupy the imperial cities Charles V might seize: Metz, Toul, Verdun, and others where German was not spoken. In these cities Henry II would act as agent for the empire, defending the latter's rights against the Emperor. Maurice of Saxony forced the Emperor to flee from Innsbruck. Henry II made the "Rhine journey," occupied Toul and Metz (April 10, 1552), reached the Rhine at Haguenau, then withdrew and entered Verdun on June 12. The "Three Bishoprics" were occupied. Charles V tried in vain to recapture Metz (autumn of 1552 to January 2, 1553), which was defended by François de Guise.

The Franco-Scottish alliance

In Italy, the town of Siena revolted against the imperial garrison on July 26, 1552, and received help from a French army. A Franco-Turkish fleet incited Corsica to rebel against the Genovese (1553). Siena was recaptured from the imperial armies in 1555 and annexed to Tuscany. In July 1554 Charles V had arranged the marriage of his son Philip to the new queen of England, Mary I, daughter of Catherine of Aragon and a Roman Catholic. By terms of the marriage agreement, the children of Philip and Mary were to inherit England, the Low Countries, and Franche-Comté, thus re-establishing the old Anglo-Burgundian alliance against France. The rest of the Spanish empire would revert to Don Carlos, Philip's son. If Don Carlos died childless, the son of Philip and Mary would inherit the entire Spanish empire. But, if Mary Tudor died childless, Philip would no longer have any ties with England; and on November 17, 1558, Mary Tudor died childless.

At this point, Charles V, ill and exhausted, sent his brother Ferdinand to represent him at the Diet of Augsburg, where Protestantism was established permanently in Germany. The treaty of October 3, 1555, between Catholics and Protestants officially recognized the existence of Lutheranism in Germany. Princes and cities could freely choose between Catholicism and Protestantism. But individuals had to follow the religion of their prince or local ruler: *cujus regio, ejus religio*. Under the principle of *reservatum ecclesiasticum* (retention of church rights), secular conversions of religious property made before 1552 remained in effect. Thereafter, any church beneficiary who adopted the Reformation had to restore the property he had acquired to the Roman Catholic Church. This was a decisive defeat for Charles V in his efforts to maintain religious unity in the empire and to wage the anti-Muslim crusade that might have become the basis for a united Europe.

The Diet of Augsburg, 1555

Abdication of Charles. The Emperor's policies had failed, and Europe was hopelessly divided. Charles V summoned the States General of the Low Countries at Brussels and abdicated before them as sovereign of the Burgundian territories in favour of his son Philip (October 25, 1555). At Brussels, too, he signed an act of abdication as king of Spain on January 16, 1556, yielding to Philip the crowns of Spain, Italy, and America. On February 5, 1556, he relinquished to Philip Franche-Comté and the Charolais region. On March 28, 1556, Philip was proclaimed king at Valladolid, Spain. Next, Charles renounced the hereditary possessions of the Habsburgs in Germany and the imperial crown (September 1556) in favour of his brother Ferdinand, king of Bohemia and Hungary. On February 3, 1557, he retired to Spain, in a palace he had built in Extremadura, close by the monastery of Yuste.

His imperial abdication was formally accepted on May 3, 1558, and Ferdinand became emperor. Charles died at Yuste on September 21, 1558.

RELIGIOUS WARS AND DYNASTIC STRUGGLES PRIOR TO THE THIRTY YEARS' WAR

Lines of communication between the Habsburgs. Emperor Ferdinand had inherited from Charles V most of the strategic entry points into central Europe (Alpine passes, Belfort Gap, Vienne Gap). From 1522 on, Ferdinand was in possession of the hereditary territories of the House of Austria: the duchies of Styria, Carinthia, Carniola, the Tirol, Freiburg im Breisgau, Sundgau in Alsace, and, in principle, a part of Switzerland: Constance, Lucerne, Thun, and Fribourg, gateways to the main passes. After 1526 he was made king of Bohemia and Hungary, which had not belonged to Charles V, and also margrave of Lusatia and Moravia.

Philip II, king of Spain, had received the Netherlands, Franche-Comté, Castile, Aragon, the Balearic Isles, Sardinia, Sicily, the Kingdom of Naples, the duchy of Milan, the imperial vicariate general (*e.g.*, overall authority) over all Italy, and the enormous wealth of the Spanish possessions in the New World.

Collaboration continued between the two Habsburg branches under the guidance of the King of Spain. The poorer Austrian Habsburgs sent their younger sons and those of the leading families in their domains to the Spanish Habsburgs in order to further their careers. Thus, the Spanish Habsburgs were well supplied with military leaders, government administrators, and diplomats. The Austrian Habsburgs received financial aid, as well as Spanish and Italian troops, from the Spanish branch; the Austrians, in turn, lent imperial prestige and influence to their Spanish kinsmen. Marriages strengthened the ties between the two branches of the family.

Now more than ever, the Habsburgs needed military lines of communication. If they had Genoa and its harbour and could keep the Duchy of Milan as a strategic turntable and if, beyond that, they held the Stelvio, Ma-

The role of the Alpine passes

loja, and Engadine passes, Spanish troops in the Milan region could maintain constant contact with Austrian troops in the Tirol. For ease of communication with the Rhineland states and the Netherlands, they had to control the Splügen and St. Gotthard passes, the Valtellina, Lake Constance, the "forest" towns of the Rhine (Waldshut, Säckingen, Rheinfelden), Freiburg im Breisgau, and the right bank of the Rhine; or, by way of Turin, the Mont Cenis and Little St. Bernard passes, the Tarentaise and Maurienne regions; or, by way of Savoy, Bugey, Valromey, the Gex region and from there to Franche-Comté, Dôle, Besançon, Brisach in Alsace, and the Rhine. From the Rhine, Spanish military highways traversed the Lower Palatinate, the archbishopric of Trier, and Luxembourg; and it was over these roads that Spanish troops reached the Netherlands. All along their route the Habsburg forces could invade France. As the Atlantic and the English Channel became increasingly denied to the Spanish fleets, these land routes became the theatre of most of the battles.

The Treaty of Cateau-Cambrésis. In his "instructions" to his son, Charles V stressed his duty "never to abandon the heritage bequeathed to him." This heritage included the Habsburg dream of mastery of Europe and the concept of protecting Roman Catholicism. It signified, consequently, a constant crusade against the Turks as well as the Lutheran and Calvinist heretics. So Philip II became the champion of the Counter-Reformation.

On February 5, 1556, at Vaucelles, the Spanish government had signed a truce with France. But the new pope, Paul IV, dreamed of freeing Italy from the yoke of Spain. An attack by the Duke of Alba, viceroy of Naples, against the Roman Campagna in September 1556 led Henry II to proclaim himself "protector of the Holy Roman Church" and to send an army into Italy. The imperial forces, allied with the English, attacked France through the Vermandois Gap. Prince Emmanuel-Philibert of Savoy routed the French at Saint-Quentin (August 10, 1557). Henry II, at Compiègne, stood guard over Paris. Then, on January 6, 1558, the Duc de Guise was able to retake Calais, then Guînes, and occupy Dunkirk and Thionville, thus cutting off the English from the imperial armies.

A financial crisis throughout Europe in 1557–59 brought both rival camps to the point of exhaustion. Henry II turned his attention to the spread of Calvinism in France, while Philip II undertook the suppression of Protestants in the Netherlands. On April 3, 1559, the Peace of Cateau-Cambrésis was signed, which stabilized international relations in Europe until the Peace of Westphalia. Philip II and Henry II restored to each other their respective recent conquests. Henry II got back the cities on the Somme but returned Corsica to the Genovese; he also gave up Milan, restored to Emmanuel-Philibert of Savoy the regions of Piedmont, Savoy, Bresse, and Bugey that had been ruled as French provinces after 1536, but retained Metz, Toul, and Verdun, and kept Calais for eight years—and was to retain it for a longer period thereafter if he paid 500,000 gold crowns. On June 22, 1559, Philip II wed by proxy Elizabeth of Valois, Henry II's eldest daughter, who arrived in Spain in January 1560.

The Treaty of Cateau-Cambrésis

Campaigns against the Turks. Philip II, a Habsburg, though Spanish-oriented and Castilian to the core, hastened to Spain as soon as the war was over, never to leave the country again. His possessions constituted a Spanish empire. Spain, born of the crusade against the infidels, understood Charles V and Philip II best. A source of noblemen-soldiers, possessing the best infantry in the world, Spain guided the empire and the course of European politics—accepting the burdens as well as the glory. But the fight against heretics blended with the desire for Spanish domination.

Spanish interests and the idea of a Crusade merged. Henceforth, great events bypassed Germany, where civil wars of religion had become endemic and which now turned into the great marketplace for mercenary foot soldiers and cavalrymen for all Europe. Ferdinand I was succeeded as emperor in 1564 by his son Maximilian II, who, two years later, invaded the Hungarian territories of young Zápolya, voivode of Transylvania and vassal of the Sultan. The new sultan, Selim II, successor to Süleyman, forced the imperial troops to retreat and extracted from them a large annual tribute and an eight years' truce. Now the Emperor was nothing but a dependent of the Ottoman Empire.

Spain required security in the eastern Mediterranean for the transport of Spanish troops, for the protection of the seaborne trade of Barcelona and Valencia, and for the transport of Sicilian grain to Castile. But the only bases the Spaniards retained on the southern coast of the Mediterranean were at La Goulette (Ḥalq al-Wādī; near Tunis), Oran, and Melilla—all of which suffered constant attack from the Barbary Turks, the successors to Barbarossa. The Turks were in the ascendancy in the Mediterranean. Although they failed to capture Malta in 1565, they seized the island of Khíos (Chios) from the Genovese in 1566 and then ousted the Venetians from Cyprus, gateway to Anatolia, Syria, Palestine, and Egypt. Once in control of the eastern Mediterranean, the Sultan could effectively intervene in the western Mediterranean. Pope Pius V was unable to unite Europe against the Turkish menace. He had to settle for a Holy League comprising Philip II, Genoa, and Savoy (May 20, 1571). The three fleets, commanded by Don John of Austria, illegitimate son of Charles V, destroyed the Turkish fleet at Lepanto, at the outlet of the Gulf of Corinth, October 7, 1571. Don John was able to recapture Tunis and Bizerta. Selim II, however, equipped a new fleet, and the Turks recovered Tunis and La Goulette (1575), offsetting the brilliant naval victory at Lepanto. The Turks of North Africa thus threatened Spanish communications throughout the western Mediterranean.

The Battle of Lepanto

The Sultan, suzerain of the voivodes of Moldavia, Walachia, and Transylvania, having humbled the Emperor, intervened as far away as Poland. Mehmed Sokollu, a Bosnian and grand vizier after 1565 under Süleyman, Selim II, and Murad II, made a pact with the Berbers, who supplied the naval forces; then he tried to incite the Moriscos of Spain against the crown, to pit France against the Habsburgs, and to set Venice against Naples. Only after the death of Sokollu in 1579 did the decline of the Ottoman Empire set in.

Events in northern Europe. Emperor Maximilian II tried in vain to reassert the authority of the Holy Roman Empire in the regions of the southern Baltic. Denmark, a Lutheran country after the Diet of Speyer in 1544, imposed a levy on all ships crossing the straits adjacent to it. But the Baltic Sea became the prey of all who sought to control the export of grain, wood, hemp, and copper to the West: of the Muscovite tsar, Ivan IV (the Terrible), who captured Narva in 1558 and proceeded to lay waste to Livonia and Courland (Kurland); of Sweden, which fought strenuously to defend Reval (Tallinn) against Ivan and against Denmark; and of Poland, which, anxious to have Danzig enjoy a trade monopoly with the West, placed the estates of the Knights of the Teutonic Order under its suzerainty. The King of Denmark, in control of southern Sweden, laid claim to the Swedish crown. He allied himself with King Sigismund of Poland, while Sweden made overtures to its former enemy, Ivan the Terrible. A "Seven Years' War" (1561–68) devastated the regions bordering on the Baltic. Emperor Maximilian managed to re-establish peace between Sweden and Denmark (with both countries restoring their respective conquests) and obtained an acknowledgment of German suzerainty over Livonia (1570). Sigismund II Augustus, having united Poland and Lithuania, had the union recognized at Lublin in 1569.

But in Poland the death of Sigismund Augustus in 1572 ended the line of Jagiellon kings and created a European problem. At first the Poles appealed to the French prince Henry of Valois, who became king of Poland (April 1573). But, when Henry's brother Charles IX died on May 30, 1574, Henry quit Poland to become Henry III, king of France. Maximilian II proposed his own candidate for the Polish throne; but it was Sultan Murad who won out: Murad's vassal, Stephen Báthory, voivode of Transylvania, was crowned at Cracow in May 1576.

The French Wars of Religion. The Austrian Habsburgs did not succeed in uniting eastern and northern Europe to check the infidels and heretics; the Spanish Habsburgs fared no better in western Europe. Nevertheless, they profited from the gradual weakening of France, where in quick succession Henry's weak sons Francis II, Charles IX, and Henry III mounted the throne and where the government fell under the influence of the queen mother, Catherine de Médicis. Catherine yearned for an understanding with Spain and wanted French princes to marry Spanish princesses. Philip II had been able to turn the Spanish Embassy in Paris into an espionage centre, and the Spanish ambassador became an advisor of the French government. On the other hand, the death of Francis II and the departure of his wife, Mary Stuart, for Scotland and the spread of Calvinism in Scotland by John Knox, with resultant difficulties for the young queen, ruled out the danger of joint action by the French and the Scots. Moreover, France was weakened by the religious wars. At one moment Charles IX seemed to draw closer to the French Protestants. Their leader, Gaspard de Coligny, had great influence in the King's council. Coligny was getting ready to send a French army into the Netherlands on behalf of the Protestant rebels there and was preparing colonial expeditions in South America against the Spaniards. Coligny's assassination and the St. Bartholomew's Day Massacre (August 24, 1572) rid Philip II of these dangers. The intrigues of Henry III's brother, François, duc d'Alençon, his efforts to marry the English queen, Elizabeth I, his negotiations with the French Protestants, his attempts to introduce Protestant French armies into the Netherlands —attempts that foundered in the face of Spanish military superiority—did not pose a serious threat to Spanish policy. (For the religious wars in France, see further FRANCE, HISTORY OF: *France, 1490–1715: The wars of religion.*)

St. Bartholomew's Day Massacre

The revolt of the Netherlands. The principal problem facing the Habsburgs was the task of repressing heresy in the Netherlands. In 1568 Philip II sent the Duke of Alba to the Netherlands, where he set up the so-called "Council of Blood" and waged a reign of terror, which provoked a general uprising in August 1571. William I the Silent, prince of Orange, had escaped into Germany, where he recruited mercenary troops. The States General of Holland named William stadholder (chief of state) in May 1572. But the St. Bartholomew Massacre deprived William of Coligny's 12,000 troops and the French fleet; Alba recaptured Mons, Malines (Mechelen), and Haarlem, where savage slaughter and looting became the order of the day. The Protestants in the Netherlands held on to Alkmaar in the north and Leiden in Zeeland, threatening to open the dikes and flood the country. Philip II was forced to recall Alba, admit the failure of his policy, and attempt to negotiate. It was no use. In April 1576, at Delft, the States General of Holland and Zeeland established a federation, temporarily entrusted William of Orange with authority, and decided to appeal to the other provinces in the Netherlands to defend their liberties. The loot of Antwerp by mutinous, unpaid Spanish soldiers—"the Antwerp fury" (November 4, 1576)—was the last straw in compromising the Spaniards in the eyes of the inhabitants of the Netherlands. Alessandro Farnese, who now became imperial governor of the Low Countries, succeeded in uniting only the southern, French-speaking provinces: Artois, Hainaut, Lille, Douai, and Orchies, where the Catholic religion continued (Arras, January 6, 1579), while, by the Union of Utrecht, the provinces of Holland, Zeeland, Utrecht, Gelderland, Overijssel, Friesland, and Groningen formed the United Provinces against Philip II. The latter put a price on William the Silent's head. Then, on July 24, 1581, at The Hague, the States General proclaimed an end to the rule of Philip II as sovereign of the Netherlands. Acting in accordance with the right of self-determination, a new nation was born in Europe.

William the Silent offered sovereign power to François de Valois, Henry III's brother, who entered Antwerp as the duke of Brabant and entered Ghent as the count of Flanders (1582). But the "insolences" of the French hordes incited the Dutch burghers and peasants to revolt. François de Valois died. William the Silent was assassinated at Delft by Balthazar Gérard (1584). It was a useless crime: the Dutch–Zeeland fleets controlled the mouths of the strategic Scheldt (Schelde) River. See further LOW COUNTRIES, HISTORY OF: *The Netherlands to c. 1600: The revolt and the formation of the republic (1567–79).*

Philip II was going to have to cope with England. In Morocco the Sa'dī (Sharīfian) dynasty opted for a holy war rather than a crusade against Portugal. Sebastian, king of Portugal, was killed and his army wiped out at Ksar el-Kebir (1578). He left no immediate heir. Philip II had himself declared king of Portugal (1580). Then he nurtured the scheme of assassinating Protestant Queen Elizabeth, whose marauding sea captains were threatening Spanish and Portuguese overseas colonies; of having his daughter Isabella declared queen of England; and of building, thus, a solid Catholic front from Algarve, in southernmost Portugal, to the Shetland Islands. The murder plot was uncovered in 1583. Elizabeth then made an alliance with the United Provinces and sent them troops (1585). After another plot, she ordered her Roman Catholic rival Mary Stuart executed (1587). Then Philip II attempted an invasion of England. But his "invincible Armada" was destroyed by the English fleet, which, with its superior ordnance and greater manoeuvrability, thus acquired control of the strategic waterways (1588).

The Spanish Armada

Shortly thereafter, Henry III was assassinated in France by the monk Jacques Clément, and the fanatical Catholics of the League refused to recognize Henry of Navarre, a Protestant, as King Henry IV (1589). In the ensuing civil war Henry IV was aided by 4,000 English soldiers sent by Elizabeth, while Philip II assumed the title of protector of the French Catholics and, to help the French Leaguers, sent Alessandro Farnese and the Spanish army in the Low Countries. Next, Philip II seized the bridgehead of the Blavet River in Brittany, sent a Spanish army into Languedoc, a Spanish garrison to Paris, and demanded the French crown for his daughter Isabella. But the Parlement of Paris proclaimed France's independence; Henry IV became a Catholic and was crowned on February 27, 1594, entering Paris on March 22 of that same year. He next set out to reconquer his kingdom. As a shield for Burgundy, at Fontaine-Française on June 5, 1595, he forced the Spanish army back across the Saône River. The Pope gave him absolution on September 17, 1595, and thus disavowed Philip II's crusade. In May 1596—at Greenwich—England, France, and the United Provinces re-formed their alliance. The Anglo-Dutch fleet sacked Lisbon (July 1596). On the ground, the Battle of Amiens ended in a Spanish defeat (March–September 1597). Philip II found himself forced to deal with France. The Treaty of Vervins re-established the Treaty of Cateau-Cambrésis (May 2, 1598). Philip had to relinquish the Low Countries to his daughter Isabella and her husband, Archduke Albert. Thus ended the great dream of refashioning Catholic unity in Europe under the hegemony of Castile.

The beginning of the Thirty Years' War. After 1598 the dream of European unity faded little by little. The Habsburgs saw their European hegemony drift ever farther away. Spain was going into decline with rampant inflation while the Holy Roman Empire was being gradually pulled apart by political and religious struggles.

Henry IV of France worked assiduously against the Habsburgs. After he defeated the Duke of Savoy, the Treaty of Lyons (1601) gave him Bresse, Bugey, Valromey, and the territory of Gex. He thus became a neighbour of the Swiss, with whom, in 1602, he renewed the traditional alliance, guaranteeing him access to the Alpine passages and driving the Habsburgs' communication axis farther east. In 1594, Spain made peace with England, set up a fortress barring the roads to the Valtellina Pass, and tried to put an end to its problems with Holland. But Henry IV negotiated a defensive alliance with the United Provinces, forcing King Philip III of Spain to sign the Twelve-Years' Truce (April 9, 1609) with his

rebellious subjects, who thus took their place within the framework of sovereign states. The expulsion of the Moriscos from Spain (1609–11) completed the weakening of Spain.

On the Rhine, the succession of Jülich–Cleves, in dispute among several pretenders, gave the Habsburgs the idea of forming a Habsburg bloc including this principality and the Netherlands. As early as 1606, Spanish garrisons had moved into the territory. In 1609 the Emperor took possession of Jülich. But in 1608 the German Lutherans and Calvinists had formed a Protestant Union. In 1609 Bavaria set up a Catholic League to oppose it. Henry IV made an alliance with the Union on February 12, 1610, in order to support the Protestant heirs to Jülich–Cleves and, on April 25, 1610, another with the Duke of Savoy against Spain, so as to cut the Habsburgs' communication lines. The assassination of Henry IV on May 14, 1610, precluded the departure of the army. But Marie de Médicis's government, in 1614, under the Treaty of Xanten, obtained the Emperor's de facto renunciation of his project.

The tolerant attitude of Emperor Maximilian II, followed by that of his successor, Rudolf II, had encouraged the development of Protestantism within the empire. Rudolf II had even given the Protestants of Bohemia a "Letter of Majesty" (July 9, 1609), echoing Henry IV's Edict of Nantes, which guaranteed their liberties to them. After a period of divergences among the Habsburgs, which had gone as far as degenerating into civil war under Emperor Matthias, the Habsburg Ferdinand V of Styria, a disciple of the Jesuits, succeeded in being "accepted" as king of Bohemia (June 6, 1617). He began the persecution of the Protestants. The latter countered with the "defenestration of Prague," by throwing from the windows of Hradčany Castle two of the imperial lieutenants on May 23, 1618, the traditional date for the start of the Thirty Years' War (*q.v.*). (R.E.M.)

II. 1648–1814

EUROPE AFTER THE THIRTY YEARS' WAR

The Peace of Westphalia. For 30 years, from 1618 to 1648, western Europe was rent by war. Though the original recourse to armed strife had come in large measure from the breakdown of the religious settlement of the Peace of Augsburg (1555; see above), far more than religious controversy was involved in the Thirty Years' War. As the war broadened and deepened, the religious differences that pitted reformed Catholicism against militant Calvinism were relegated to the background, and secular issues became more important in the fighting.

The fighting of Catholics and Protestants, of course, involved the issue of religion; but another conflict was constitutional in character, somewhat akin to a civil war: to determine with whom authority would rest in the Holy Roman Empire—*i.e.*, with the emperor or with the large member states. Lastly, the Thirty Years' War was an international war among the states of Europe, a war in which the stakes were largely territorial power, strategic frontiers, and defensible boundaries.

French entry into the Thirty Years' War

France was involved in the war from the outset, giving subsidies and diplomatic counsel to common anti-Habsburg enemies. It was not until 1635, however, that the great French minister Cardinal de Richelieu made a formal declaration of war on Spain. The religious aspects of the war were not of primary importance to the Cardinal, but the territorial aspects were: with France caught between the troops of the Habsburgs in Italy and in the empire, Richelieu's first concern was to win defensible frontiers so that France could have military security. To gain that security, however, he had to destroy the military might of the enemy, first of all of Spain, whose armies could encircle France along its eastern front from north Italy to the Spanish Netherlands.

Richelieu cast the die in 1635. The moment had come, he reasoned, for France to throw all its reserve strength in the balance. His calculation was sound. After serious reverses, French military strength began to tell. Richelieu died in 1642, before the final victory was won, but the troops of Richelieu's successor, Jules, Cardinal Mazarin, compelled the Austrian Habsburgs to open negotiations.

Prolonged and acrimonious peace parleys opened in 1644 in the towns of Münster and Osnabrück in Westphalia. Not for two centuries had so many negotiators gathered to discuss so much and to reach so many compromise decisions.

The religious provisions were a clear acknowledgment of altered reality. The principle of *cujus regio, ejus religio*, set down at Augsburg in 1555, was not only reaffirmed but was extended to cover Calvinism as well as Lutheranism. Religious passions were far from eradicated in the spiritual life of millions of believers, but religious wars were no longer fought in Europe.

The territorial provisions provided generously for the victors at the expense of the empire. The French title to the strategically important bishoprics of Metz, Toul, and Verdun was confirmed. Though deliberately ambiguous clauses seemingly fell short of giving France full sovereignty in Alsace, in reality they allowed the French to extend their influence within Germany. Sweden, allied with France, emerged as an important European state, obtaining the bishoprics of Bremen and Verden and the western half of Pomerania (Hither Pomerania). Sweden thus acquired a hold over the shores of the Elbe, Oder, and Weser rivers. Of the larger German states, the electorate of Brandenburg, another ally, was confirmed in its possession of eastern Pomerania (Farther Pomerania) and in its succession rights to the archbishopric of Magdeburg. The Upper Palatinate went to the elector of Bavaria, an ally of the French, and the Rhine Palatinate went to the son of the former elector Palatine, another French ally.

The political arrangements were not less important than the territorial. The authority of the empire, already in manifest decline, was further weakened. The 300-odd states into which Germany was divided were recognized as sovereign states free to make alliances outside the empire as well as among themselves within. It was stipulated that the empire could not raise taxes, recruit troops, declare war, or make peace save with the consent of the states. France and Sweden won the right to sit and vote in the Imperial Diet. The effective authority of the Austrian Habsburgs, from whose family the emperors traditionally had been elected, were restricted to their hereditary lands of the Austrian duchies and the recently conquered kingdom of Bohemia.

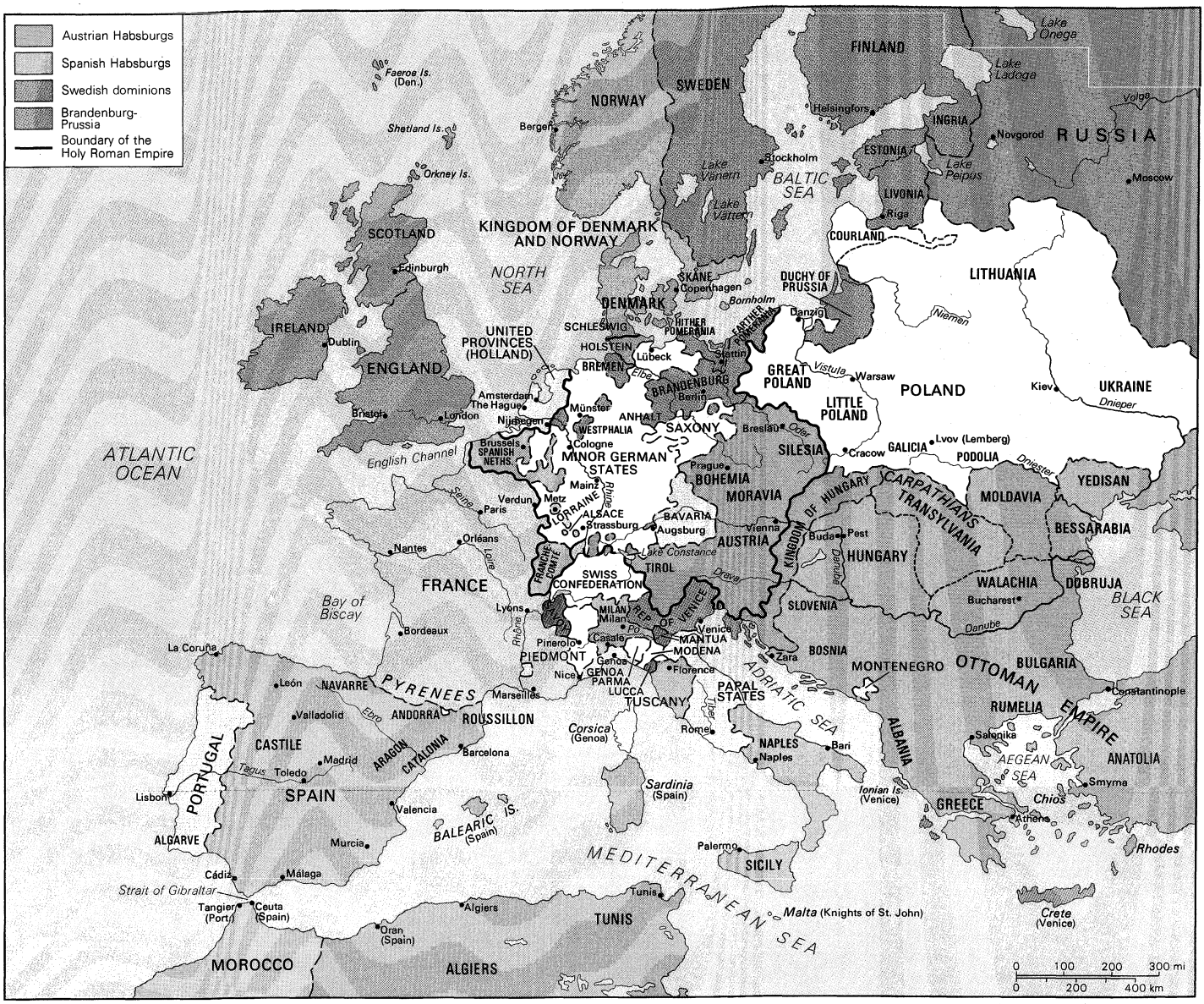
The physical devastation was appalling. Mercenary troops, and their field commanders, roamed freely over the ruined countryside. Hunger and pestilence decimated the population, in the opinion of many historians, by at least a third. The war could not have ended too soon.

The Peace of the Pyrenees. War between France and Spain continued for 11 more years after Westphalia. Almost to the end neither contestant was strong enough to win by itself. Neither Spain, beset by revolts in Catalonia, Naples, and Portugal, nor France, tormented by an aristocratic rebellion known as the Fronde, could mount an overpowering offensive. In the circumstances, Cardinal Mazarin sought to break the deadlock by diplomacy. One string to his bow was to neutralize Germany and prevent the Emperor from putting troops in the field. When Emperor Ferdinand III died in 1657, Mazarin intrigued to win the imperial throne for the young Louis XIV of France. The effort failed, for the Habsburg Leopold I was unanimously elected as Ferdinand's successor.

Neutrality of the empire

But Mazarin's intrigues were not futile. The electors wrung from Leopold an agreement to remain neutral between France and Spain so long as the imperial territory itself was not involved in the fighting. A year later, in 1658, Mazarin took another long step forward toward neutralizing Germany by having Louis XIV named "protector" of the newly formed League of Rhine Princes, which proclaimed unity and peace as its aim.

Meanwhile, Mazarin made a bid for the support of England on the seas. Though a cardinal of the church, he did not scruple to invoke the aid of the Protestant regicide Oliver Cromwell, the Puritan leader who had come



Europe in 1648.

to power in the wake of the English Civil War (1642–51). Cromwell was receptive to overtures. The transaction reached successful issue in 1657 when by a formal alliance Cromwell agreed to support France against a common Spanish enemy. With British fleets on the seas, Cromwell could look forward to acquisitions in the Spanish colonies as compensation for joining France in an attack upon Dunkirk, which was still in Spanish hands. The fall of Dunkirk and the French general Henri de Turenne's great victory over the Spanish at the Battle of the Dunes (1658) wrote a finish to the long struggle that Westphalia had only partially ended. The Peace of the Pyrenees in 1659 destroyed Habsburg Spain's long military ascendancy over Europe. Spain lost much and obtained nothing. The territorial provisions of Westphalia were reaffirmed. France retained the vital bridgeheads across the Spanish lines of communication from Milan to the Netherlands and on the southwest obtained Roussillon and Perpignan, which established the Pyrenees as the boundary between the two countries. In addition, France acquired most of Artois from the Spanish Netherlands, along with several cities in Flanders.

To seal the peace, the infanta Marie-Thérèse was given in marriage to the young French king. Aware of the implications of that marriage for the future, King Philip IV of Spain had insisted that his daughter renounce all claim to the Spanish throne. The astute Hughes de Lionne, chief French negotiator, acceded to this condition, adding, however, the all-important proviso that the renunciation be made contingent upon the payment of a dowry of 500,000 Spanish crowns. The marriage was appropriately celebrated in a splendid ceremony, but the dowry remained unpaid. And when a sickly child, the future Charles II, was born of Philip's second marriage, the shadow of his precarious existence, somehow prolonged for almost 40 years, fell athwart the political scene. Meantime France had entered upon its own ascendancy in western Europe.

The Treaty of Oliva

A year later, in 1660, the Treaty of Oliva ended the ruinous fighting along the Baltic known as the First Northern War, which had been fought since 1655 between Sweden on the one hand and Poland, Austria, and Brandenburg on the other. For the first time since 1618, all of Europe, in the east as well as in the west, was at peace—at least on a formal treaty basis. But no formal apparatus or mechanism existed to keep the peace. In its absence the states of Europe sought to compose the fears, rivalries, and ambitions that set them apart by combining two approaches: one was to invoke procedures that implemented the concept of balance of power; the second, employed concurrently, was to make the state strong enough to deter attack. Hence, to maintain a balance of power abroad while making their state strong at home became the leitmotiv of the makers of policy. Every-

where on the Continent they looked toward France as the model to be emulated.

France's position in Europe. They could do so with good reason. Just as its leadership of western Europe became established in the political affairs of the Continent, so within the span of two generations the cultural ascendancy of France became undisputed. By its language and its literature, its arts and its philosophy, France was the recognized leader, the standard bearer, and the envied model of the civilization that prevailed in the western world from London to Vienna.

Cultural predominance of France

France was not only foremost in the arts of peace, it was also unmatched in the concentration of royal power that had been achieved both for affairs within and state relations outside. There was first the king himself, the "Sun King" Louis XIV. Not exceptionally endowed in intellect, only exceptionally vain and avid of flattery, Louis systematically sought to establish his mastery within the state that he would make master in western Europe. Considering himself the instrument of France's destiny, he rejected, at least in outlook, all inherited feudal conceptions of the nature of the state and the relation of the king to it. He was the embodiment of the dominant political concept of his century, the conviction that to the king belonged plenary control over civil and military policy and regulation. It was this deeper sense of his position and not mere vanity that lay at the core of his repeated statements that he was the state: *L'état c'est moi.*

Policy lines were laid down by the few ministers and secretaries of state in whom he had confidence and with whom he worked in close collaboration. Enforcement came from the civil servants, the *grands commis*, as Louis XIV called them. Not complete enforcement, however, for the old "constituted bodies" of *parlements* and other sovereign courts, provincial estates, and municipal corps remained and often obstructed the implementation of policy at the summit. Within the administrative revolution that was being effected there were margins for resistance. The fact remains, nevertheless, that Louis had power possessed by no other monarch. He was the supreme example of royal absolutism.

Colbert's reforms

Another increment of state power came from the stimulus that Louis's finance minister, Jean-Baptiste Colbert, gave to French economy. Colbert instituted a sustained and comprehensive program to link the reform of the taxation system with a massive government-directed development of industrial and commercial activities. The fiscal program broke down before the opposition of the privileged classes and the debilitating demands of war, but the thrust of his endeavour was neither completely blunted at home nor ignored by France's enemies or rivals abroad.

They appreciated his intent to stimulate economic development and end France's dependence upon the import of vitally needed goods manufactured abroad. Colbert's full program was frustrated; indeed, it was too demanding to be capable of realization, but what was achieved under its aegis was considerable.

Revolutionary changes in military and naval organization and practice were not less radical than those effected in civil administration and economic behaviour. Under the direction of Michel Le Tellier and his son François-Michel Le Tellier, marquis de Louvois, the crown created a new formidable instrument of power. In large measure the disorderly rabble of Louis XIII was transformed into a trained, disciplined, and centrally controlled standing army. Many of the colonels and generals who had purchased their commands were replaced, and new commissions came from the King. A ladder of promotion by merit was introduced. A sustained attempt was made to integrate the branches of service, the infantry, cavalry, and the artillery, the latter under civil control. In increasing measure military intendants, government appointed, attended to multiple details concerning recruitment, and upon them devolved enlistment, feeding, equipping, clothing, and paying of new recruits. From a force of 72,000 in 1667, the soldiery increased to 120,000 five years later and to nearly 300,000 at the beginning of the War of the Spanish Succession (1701–14). Nowhere in Europe did a comparable military power exist.

It was not only in its military apparatus that France could not be matched. Except for the United Provinces of the Netherlands, the states of western and central Europe were strikingly weak in all the respects that made France strong. If France disliked and feared the Dutch Republic, that feeling derived in great measure from irritation over the extraordinary strength of a state that was both republican and Protestant. Dominated by businessmen possessed of a business ethic, the republic had extraordinary wealth and power, well beyond its national resources, population, and territorial extent. It carried a large part of the seaborne commerce of Europe. It was a leading colonial power, and Amsterdam was a world financial mart. Louis XIV disliked the spectacle of Dutch power that presented itself; Colbert had an obsessive fear of it and favoured a preventive war. From England there was no challenge. Allowing for Louis XIV's exaggeration, his verdict was accurate enough: "England could hardly breathe again after her past ills." The Italian peninsula, once so important and powerful in the politics of the states, had been relegated to the margins of history. The great days of the Austrian Habsburgs and their own Danubian kingdom were yet to be. As for the empire itself, though enfeebled by the recent wars, it still had to be reckoned with regarding alliances made and unmade, and reckoned with not least by France.

France's relations in the north and east

The wars in northern and eastern Europe, waged separately from those in the west, were not of first moment to France in the play of balance-of-power policy and diplomacy. Still, by its traditional policy of maintaining a tier of allies along the eastern frontier of the empire, France was marginally involved in those struggles that the Peace of Oliva ended. The states engaged in that power struggle in the years immediately before and immediately after the beginning of the personal rule of Louis XIV in 1661 were Sweden, Denmark, Poland, Brandenburg, Russia, and the Ottoman Empire. Though Sweden was to decline catastrophically during the second half of the century, its position was still strong in 1660 when the Treaty of Oliva ended the war with Poland over the Swedish throne and left Sweden in control of the Baltic. France judged that Sweden could still be counted upon as the northern link of its alliance system in eastern Europe. A role in French calculations was also cast for Poland in order to keep eastern Europe from becoming a second theatre of operations if France should have a war in the west on its hands. Here French confidence was little justified, for Poland's decline was accelerated when its wars against Russia and the Ottoman Empire ended disastrously with the loss of the eastern Ukraine to Russia. Russia's entry into the power relations of Europe was not to begin for some years after 1661. Its steady advance against the Ottoman Empire notwithstanding, Russia did not count in the strategy of wars and diplomacy of the West in 1661, nor did developments in the Ottoman Empire, still regarded as a reliable pro-French state.

In 1661, then, all seemed relatively quiet and secure for France when Louis XIV turned to consult his interests in the West.

THE AGE OF LOUIS XIV

The War of Devolution. During the early years of Louis XIV's rule, French relations with the powers were largely focussed on the problems of the succession in Spain. The normal difficulties attending such matters were accentuated by the presence of the wretched Charles II on the Spanish throne from 1665 until his death in 1700. Afflicted by multiple ills of body and mind (not quite imbecilic but, what was more injurious to the interest of his state, certainly impotent), he was the problem child of international relations.

Since Charles could have no offspring, the Spanish branch of the Habsburgs was doomed to expire with his death. The future of the Spanish monarchy was therefore at stake. Involved also was the disposition of the

personal estates of the royal family in the Spanish Netherlands. This Spanish question was at the heart of the jealousies, rivalries, intrigues, and combinations that worked themselves out in diplomatic manoeuvres and wars.

Bourbon claims to the Spanish throne

By virtue of his marriage to Charles's sister, Louis XIV established claims and fashioned designs on the great heritage, arguing that his wife's renunciation was invalid since the stipulated dowry had not been paid (see above *The Peace of the Pyrenees*). One consideration of his policy was to advance French frontiers eastward to the Rhine by annexing the Spanish Netherlands and Franche-Comté. Another thrust of his policy, which was not defensive, was to pool the resources of France with those of Spain, both in Europe and overseas, by getting the entire Spanish inheritance exclusively for his kingdom. To attain these ends, Louis intrigued with all the states, large and small, giving subsidies with fine impartiality, so long as they served his purposes.

Louis XIV chose not to fight in support of his claim to the Spanish throne in 1665, when the heritage passed from Philip IV to his son Charles II. Louis bided his time, isolating Spain diplomatically and familiarizing his fellow rulers with the term devolution, which he applied to his claim. The term derived from legal principles and procedures in private law prevalent in Flanders, according to which property rights of the deceased devolved to children of the first marriage to the exclusion of the children of the second. This criterion was certainly not relevant to royal heritages, and to use it as Louis did in 1667, when he declared war, was to set up a facade to validate his claim. Behind that legal front were his troops, poised for action.

His generals Turenne and Sébastien Le Prestre de Vauban struck in 1667 and drove deep into the Netherlands. This offensive galvanized the state system into action. England and the United Provinces came together in an alliance to wage war on France if France rejected their mediation. Sweden joined and made it a triple alliance; Emperor Leopold, lacking the military strength to do otherwise, covered himself by a secret treaty of partition to go into effect on the death of Charles II—an event regarded, however mistakenly, as impending.

The Emperor's hedging and the vigorous mediation of the alliance gave Louis pause. Reasoning that time was on his side anyhow and reassured of his strength by the Great Condé's spectacular conquest of defenseless Franche-Comté, Louis found an opportunity to get out with honours. By the Treaty of Aix-la-Chapelle (1668), Louis withdrew his troops, retaining only some 11 border cities in Spanish Flanders.

The Dutch War. French relations with the Dutch Republic steadily deteriorated after the war on Spain, and a second major conflict broke out in 1672. On the Dutch side there were deep, well-grounded fears of French aggression and expansion, while the French, for their part, had their grievances, particularly resentment over the republic's initiative in having contracted an alliance with England. These differences might have been composed had the will existed. Even Colbert's almost pathological fear of Dutch commercial and financial strength need not have led to war. The ultimate responsibility lay with Louis XIV himself. Without even the pretext that the Dutch army threatened French security, he deemed war necessary for his personal glory.

The fighting began in 1672 when Turenne dramatically crossed the Rhine and then penetrated deeply into Holland itself. The victorious French troops, however, delayed their advance against Amsterdam, providing a delay that gave the retreating Dutch time to elect the unyielding young William of Orange stadtholder. The delay also gave the defenders time to make a desperate decision. They breached the dikes and Holland lay flooded. The water did not freeze, and the invaders were forced to abandon their advanced position and evacuate the territory they occupied.

The flooding of Holland

The war changed character when, after the retreat of the French, the Emperor came into the fighting by the side of William of Orange. The latter, in his new capacity as elected executive officer of the state, bestirred himself to form the broad Coalition of The Hague, to which adhered such disparate members as Spain, the Holy Roman emperor, England, most of the Rhenish states, Denmark, and Brandenburg.

As the network of alliances that France had formed began to rip apart, it became the turn of the French to defend their now threatened lines of communication as best they could. Their armies fought brilliantly on several theatres of operations, but, despite great victories in the Netherlands, spectacular successes along the Rhine, and the occupation of Franche-Comté, they could not mount an overpowering offensive. Nor could the coalition troops break the French defense. The Treaties of Nijmegen (1678–79) terminated the long-drawn, inconclusive fighting.

Though the French seemed victorious, the Dutch were far from conquered. In fact, in the larger context of French designs, the war could be held a blunder. France surrendered its conquest of the Spanish Netherlands, obtaining in exchange a number of cities on the northeast frontier. From Spain it also obtained recognition of its conquest of Franche-Comté, so necessary for France's defense. From the Dutch themselves France obtained no territory. The restrictive tariffs were quietly abrogated. In a larger perspective the war had significantly altered state relations. The Dutch republic had withstood the best war machine in Europe. The great maritime powers of England and Holland had come together, and in William of Orange France had an implacable and unyielding foe.

War of the League of Augsburg. The Peace of Nijmegen did not close the chapter of French aggression. Once again a coalition of states was required to check Louis XIV in his course. Blocked along the Lower Rhine by Holland, France nibbled at the territory of the Holy Roman Empire along the Middle and Upper Rhine. For a long time French jurists had been gathering data concerning territorial dependencies that had been or might not have been ceded to France by the deliberately ambiguous provisions at Westphalia. In 1680 courts of claims called *chambres de réunion* were set up to make rulings on doubtful or disputed territory. Since the judges were French, most of their verdicts were favourable to their country. After scattered small bits of territory, largely in Lorraine, had thus been incorporated into France, another ruling in 1681 held that the important imperial city of Strassburg was also a dependency of France. It was at once occupied by French troops.

The intent animating this *réunion* policy was clear: to sever such bonds as still linked the occupied territory to the empire and, by integrating them completely under French sovereignty, to establish a territorial boundary that France regarded as essential for its defense. The annexationist tactic went further. It fitted into the larger pattern of taking what could be acquired one way or another from the enfeebled empire. Direct territorial despoilment in the west also fused with the French conception of embarrassing the empire in the east by giving generous subsidies to Magyar rebels in Hungary and encouraging Turkish attacks.

Renaissance of Habsburg military strength

This classical ploy of *Realpolitik* misfired because of a remarkable and decisive renaissance of Habsburg military strength. After containing the Magyar and Turkish threats, Leopold became free to turn to affairs in western Europe and make himself the leader of a Catholic union of states against Louis. In those same years and immediately following, William of Orange was playing upon the fears and indignation of Protestant states who were menaced not only by Louis's military aggression but also by his decision in 1685 to revoke the Edict of Nantes, which since 1598 had extended toleration to Protestants within France. These anti-French combinations were strong enough to force Louis to accede to the Treaty of Regensburg (1684). By its provisions Spain recognized the annexations in Luxembourg and the Netherlands, while Louis, retaining Strassburg, renounced further annexations.

Louis kept Strassburg, but he did not keep his word.

After his troops again occupied the Palatinate, the two separate anti-French groups centred around Leopold and William of Orange pooled their strength into the defensive League of Augsburg (1686), adhered to by all those states threatened by France and England—the Holy Roman Empire, Spain, Sweden, the electorates of Bavaria, Saxony, and the Palatinate, and the Dutch Republic. In 1688 William of Orange also became king of England as William III. After that great power also joined the League, it was transformed into an offensive alliance, the Grand Alliance. An open war then began, and France faced a reconciled Europe alone. If France had had even a semblance of a case for its aggressive measures in a war-oriented polity of states where the line between defense and expansion was not always clear, Louis spoiled it by his domineering tone and unilateral decisions.

The Grand Alliance

He had hoped, by early decisive victories over each adversary in turn, to obviate the necessity of waging a general war with all at once. That hope disappeared as the allies quickly organized their forces, coercing France to fight on a broken front on the Continent and at the same time almost everywhere on the high seas. The coalition was at least equal in strength to France in effectives, leaders, and financial resources. The armies in the field were immense, as were their losses. By 1693, the fourth year of the war, it was clear that no decisive military solution was possible. France itself was safe from invasion, but the French could not penetrate into the heart of the empire or shatter the joint naval might of the English and the Dutch. There were no decisive engagements on the seas, and a French attempt to invade England broke on the defeat at La Hougue in 1692, where the fleet of Anne-Hilarion de Cotentin, comte de Tourville, was annihilated.

To maintain 200,000 men in the field and two squadrons on the seas strained French finances to the breaking point. Moreover, the crop failed in many regions of France in 1693, and famine and disease swept the country. Famine prices, collapsing wages, and unemployment came together in a chain reaction of suffering unparalleled since the terrible years of the Fronde. Among the allies, also, there was keen misery and war weariness. The Grand Alliance crumbled, and in 1697 the war ended with the two treaties of Rijswijk, the first between France, England, Spain, and the United Provinces, and the second between France and the empire.

By the treaty provisions France gave up its conquests in Italy and returned the occupied strategic fortresses of Pinerolo and Casale. In Germany, France abandoned its claims to the succession in the electorates of Cologne and the Palatinate. Alsace was retained but not the *réunions* effected prior to 1679. The French also retained Strassburg, but Lorraine was restored to its duke. On the right bank, too, all the land seized by the French reverted either to the empire or to individual princes. The United Provinces also gained by the terms, winning commercial concessions together with the right to garrison the Barrier fortresses between France and Holland. Both France and England restituted colonial territory, and France formally accepted William of Orange as king of England. Louis's *réunion* policy was completely blunted.

The War of the Spanish Succession. No mutually acceptable arrangements had meantime been worked out among the rival claimants for the disposition of the Spanish heritage. Between 1698 and 1700 Louis XIV, understandably wishing at all costs to avert war, took the diplomatic initiative to draft partition treaties acceptable to France, England, and the United Provinces while defeating Leopold's design to be sole heir. Neither of the two treaties that France proposed proved acceptable to the Emperor. The moribund Spanish king then tossed a bombshell into the already tense situation by proposing his own arrangements for the Spanish succession. He had been persuaded by pro-French counsellors that only the House of Bourbon possessed the military strength required to hold the Spanish inheritance intact both on the Continent and overseas.

In the autumn of 1700, shortly before his death, Charles drafted a will embodying that conviction. It bequeathed the whole inheritance to Louis's grandson, Philippe, duc d'Anjou, stipulating, however, that if Louis refused for his grandson, the entire heritage would go to the archduke Charles, son of the Emperor. In either contingency there would be no partition of the heritage. The future lay with Louis, and, after long discussions with advisers, he decided that he could not reject what seemed like a gift from heaven. He accepted Charles's will, enjoining upon Anjou to be a good Spaniard.

The terms of Charles's will

The maritime powers bowed to reality, not without a protest from the Emperor and deep misgivings on the part of William III, who argued that England had everything to lose and nothing to gain by acceptance. All seemed propitious for France when Louis committed a series of blunders that confirmed William III in his fear that Madrid was to be governed from Versailles. At the very moment that France should have been tactful and conciliatory, Louis acted as though he wished gratuitously to alienate and defy the powers. The Parlement of Paris was forced to register a royal letter patent that reserved Anjou's rights to the French throne. French troops occupied the Barrier fortresses, temporarily driving out the Dutch. The French government also advanced the claims of French companies to share in the English trade with the Spanish colonial empire—another display of arrogance that outraged public opinion in England. And in a final belligerent move, Louis formally recognized the son of the ousted James II of England as James III.

Perhaps Louis intended by this peremptory behaviour to impose an image of his overwhelming strength upon the powers; but he succeeded only in heightening apprehensions and bringing the powers closer together in the Grand Alliance for which William III had so tenaciously laboured. William died in 1702, and the union was his final bequest to block French aggression. Facing the greatest threat to the political balance of Europe since the days of Emperor Charles V, William had fashioned a coalition that was to be France's undoing. The War of the Spanish Succession began in 1701.

The allies gradually established their domination, largely because of their two great field commanders, the Duke of Marlborough and Prince Eugene of Savoy. By 1708 France was open to invasion, and the year 1709 was a catastrophe at home. It was the year of the "great winter," which brought suffering that surpassed even that of 1693–94. War weariness, exhaustion, and the threat of bankruptcy generated a peace movement. But the terms offered by the victorious allies were so onerous that Louis XIV would not accept them, and military operations were renewed. Marlborough won his last great victory at the bloodbath of Malplaquet in the Netherlands, and peace negotiations were soon reopened under the insistence of the pro-peace Tories, who won the general election of 1710. After Marlborough was relieved of his command, England virtually withdrew from further military operations, leaving its allies to fend for themselves as best they could. Encouraged by England's example, they, too, made their own arrangements. A series of peace treaties in 1713–14, of which Utrecht and Rastatt were the most important, brought peace to Europe.

Allied victories

The Peace of Utrecht. The French paid heavily for their military reverses: they abandoned their designs to annex the Netherlands and also gave up all conquered territory east of the Rhine. Of Louis's conquests only Franche-Comté and some towns and cities in Flanders remained in French hands. After a half century of aggression this territorial reward was not striking. With respect to the troubled successions, the English recognized the Bourbon Philip V, the former duc d'Anjou, as king of Spain, while France accepted Anne on the English throne. Several bastions were set up against possible future French aggression. The treaties recognized Prussia as a kingdom with possessions on the Rhine. The Dutch again occupied the Barrier fortresses. In the southeast of the line set up against France, the new kingdom of Savoy stood on guard, retaining occupation of Nice and acquiring Sicily. French expansion was contained, but France lost no territory of its own on the Continent and remained the richest, the best-governed, and the strongest

country in Europe. Though France was forced to cede valuable possessions to England in North America (Newfoundland, Nova Scotia, and sovereignty in the Hudson's Bay territory), the country's extraordinary vitality and rich resources nevertheless remained. Within a few years after the peace France embarked upon a remarkable, century-long economic expansion that was to make the country a formidable rival of England for maritime and colonial supremacy.

The territorial provisions of the peace settlements fell heavily upon Spain, which lost the Netherlands, Milan, and the kingdom of Naples to the Austrian Habsburgs. The greatest victors of all were the English, who occupied Gibraltar and Minorca, which had been conquered early in the war and which opened the way toward British control of the western Mediterranean. By the Methuen Treaty (1703), Portugal became a virtual commercial satellite of Britain, giving England valuable trading rights with Brazil. The cessions of French colonial territory noted above, together with those acquisitions from Spain and Portugal, contributed significantly toward the building of the first great British empire. Of all these changes, the most immediately valuable, at least for the merchants, was the *asiento* agreement. By its provisions England won a 30-year monopoly to import Negro slaves into the Spanish-American colonies and, also, the exclusive right to send one ship a year via Portobelo, on the Isthmus of Panama, to provision the colonies.

The *asiento* agreement

Peace in the North. While the War of the Spanish Succession was being waged in the west, the final great struggle was fought in the east to end Sweden's control of Baltic shores. So long as France's allies, Sweden and Poland, were strong and stood on guard as buffers against involvement, there had been no occasion for either France or England to be seriously concerned. The military collapse of Sweden in the Great Northern War (1700–21) against a broad coalition of states that included Poland, Denmark, Saxony, Prussia, Hanover, and, most important of all, Russia, significantly altered the military and diplomatic pattern. Sweden, France's outpost in the east, lay prostrate, and Russian troops had reached the Baltic.

Peacemaking between Sweden and the lesser states to settle rival territorial claims was speedily effected. The settlement of the Russo-Swedish boundary limits and the reorganizing of their relations proved more difficult. France and England, for instance, had to be reckoned with. France was concerned to keep Russia from acquiring full Baltic control in the place of Sweden. England was anxious to gain assurances of unmolested shipment of needed naval stores from Scandinavia.

With these considerations in mind, France and England, linked in alliance since 1716, planned to put pressure on Russia in the settlement of what was becoming a general European problem. Peter I the Great was not inclined to be so influenced, and he had in a benevolently neutral Austria a decisive supporter of his views and intentions for Russia. Hence he was free to impose, almost to dictate, territorial terms to Sweden in the Treaty of Nystad (1721). That peace stripped Sweden of its provinces of Karelia, Ingria, Estonia, and Livonia as well as some small islands.

The Treaty of Nystad

All Europe was now at peace from west to east; and Russia, in uncontested control of the Baltic, had entered into the diplomatic relations of the Continent.

PEACE AND WAR, 1713–89

Postwar problems, 1713–40. After Utrecht and Nystad many problems and issues still remained to threaten the peace and challenge the firmness and ingenuity of the directors of state policy. The solutions were made at several congresses of diplomats, at ministerial conferences, and all under the aegis of ad hoc alliances. Among these alliances the first was that of France and England in 1716, which two years later out of necessity developed into the Quadruple Alliance, including the United Provinces and Austria.

One postwar problem concerned the succession to the thrones of France and England. In France, where Louis XV was still in his minority, two rivals advanced their claims to the throne: Philip V of Spain, whose rights Utrecht had denied, and Philippe, duc d'Orléans, Louis XIV's nephew and regent for the young king. The Quadruple Alliance validated the rights of Orléans, but since Philip persisted in maintaining his claims, tension over this matter remained. In England the danger arose from the Jacobite supporters of the Old Pretender, the self-styled James III, who refused to recognize George I as king. The government crushed the Jacobite rebellion in 1715, but plotting to seat his son, Charles Edward, the Young Pretender, continued with sporadic foreign aid until 1745.

There were also diplomatic clouds over the troubled situation in the Barrier fortresses, where the Emperor refused to allow the Dutch to garrison those posts. Here, too, a congress meeting was required. Problems in Italy made for trouble because Philip V remained unreconciled to the Utrecht provisions that ratified Spain's heavy territorial losses to Austria. Two local wars were waged in Italy between Spain and Austria until, in 1720, the Quadruple Alliance successfully mediated the disputes, compelling Spain to abandon its Italian claims and, at the same time, inducing the contesting states to endorse the solution of the succession difficulties in both England and France. There was difficulty, too, in the unfriendly relations between Spain and the maritime states concerning the settlement in Italy, as well as over the tireless manoeuvres of the Emperor to obtain signatories to the Pragmatic Sanction, which provided for the powers to accept inheritance by his daughter Maria Theresa of all the hereditary Habsburg possessions. In the east, there was the War of the Polish Succession (1733–38) over the election of the king of the Polish republic and, also, the liquidation of the wars waged by Russia and Austria against the Ottoman Empire. Austria lost heavily to the Turks by the Treaty of Passarowitz, and Russia, too, was temporarily checked in its advance.

The Pragmatic Sanction

Nevertheless, Europe as a whole gained because a general European conflagration was averted. These settlements in both the west and east preserved the peace, but barely. The built-in predisposition toward war had not been eradicated. Warfare and competitive militarism were neither condemned nor rejected. The procedures animating the balance settled crises; they made no contribution to lasting peace. As the range of state competition widened and as its substance deepened, to increase security became the primary concern of all policy makers. In 1740 a generation-long respite from general war ended with a series of conflicts that gripped Europe until 1763. In these wars, the balance again revealed itself, in the words of a contemporary, as a mechanism for

> recognizing the strength, resources, interests, rights, fears, and hopes of the different powers, so that as the occasion warrants it, we may appease these powers, divide them, defeat them, or ally ourselves with them, depending on how they serve our advantage and increase our security.

War of the Austrian Succession. All appeared quiet after the last settlements of 1739. The pent-up resentments erupted in the following year, and the balance was shattered by two wars fought more or less simultaneously, one colonial and commercial overseas, the other dynastic and territorial on the Continent.

The former, an Anglo-Spanish conflict, had its origins in the repeated violations of the *asiento* agreement. English privateers, going beyond the provisions stipulated in the Peace of Utrecht, engaged in a lucrative smuggling trade, which Spanish naval officers repeatedly broke up by a rigorous policy of search. After many mutually injurious incidents and outrages, the clamour of the merchants forced England's chief minister Robert Walpole to abandon his longtime policy of peace and declare war on Spain—the War of Jenkins' Ear (1739–43), so-called after an English seaman whose ear was amputated by the Spanish patrols. France, because of its heavy financial interests in colonial trade, prepared to support Bourbon Spain by dispatching two powerful squadrons to fight the English in the West Indies. These instructions were quick-

Anglo-Spanish hostilities

ly rescinded when Frederick II the Great of Prussia invaded Habsburg Silesia without warning. Frederick's aggression brusquely and brutally undid the Emperor's prolonged diplomatic efforts to have all the powers sign the Pragmatic Sanction. Frederick found in the accession of the young, inexperienced Empress a tempting opportunity to round out the strategic Prussian boundaries and take for his kingdom another step toward military hegemony in Germany. The invasion of Silesia and Frederick's early triumphs stimulated other states to press their claims against Maria Theresa. His invasion also gave impetus to the anti-Habsburg coterie at Versailles in its plan to repudiate the policy of peace that the wise nonagenarian André-Hercule de Fleury, Louis XV's great minister, had been pursuing for almost two decades. This hostile faction succeeded in bringing France into the Austro-Prussian War at the head of a coalition against the traditional continental enemy Austria. Bound to France by the Pacte de Famille, Spain entered into the war.

The fighting might well have ended in 1742 when, by a separate agreement with Maria Theresa, Frederick withdrew, with Silesia in his possession. Instead, it took on European dimensions. The Empress declared war against France. In England, Walpole fell from power, and then England chose, like France, to engage in a limited war in Italy, with the end in mind of weakening Spanish total resources by a military diversion on the Continent. This English strategy of a holding operation provoked, in turn, a needless French declaration of war against England prompted by panic in high governmental circles. Thus the two wars were joined into the single War of the Austrian Succession, though again in 1745 Frederick's straitened resources forced him to make peace with Austria and withdraw from the fray; but, with Silesia still in his hands, the now pointless fighting continued. On land the French troops were in the main victorious. Outside Europe the British fleet and privateers destroyed the French escorts and drove both French and Spanish merchantmen off the seas.

Treaty of Aix-la-Chapelle

The financial exhaustion of all the powers finally brought sanity into play, and France and England undertook secret peace negotiations that resulted in the Treaty of Aix-la-Chapelle (1748). The signatories confirmed the Pragmatic Sanction, and Austria formally accepted the cession of Silesia to Prussia. As for Frederick, his *démarche* paid off. Silesia was his. Overseas nothing essential was changed. France and England made mutual restitutions of their conquests. The *asiento* was not altered. This War of the Austrian Succession, the first round of the mid-century wars, had settled nothing. "As stupid as the peace," cried the wits. The war only bequeathed a legacy that was the seed of future conflict, the legacy of the Austrian determination to compel Frederick to disgorge Silesia and restore the older balance in place of the new one.

The "Diplomatic Revolution." In the years immediately following Aix-la-Chapelle, the only force that seriously threatened the peace was the studied resolution of the Austrian chancellor Wenzel Anton, Graf von Kaunitz, to win back Silesia. To be sure, an undeclared war between England and France was being fought in America (the so-called French and Indian War), but, far from desiring to have it spread to Europe, England shrank from that prospect. The English were fearful that in the event of war on the Continent either France or France's ally Prussia might overrun the Duchy of Hanover to which the English king held title and, also, perhaps occupy the Netherlands. As England's own ally, Austria, could not give adequate assurances of military aid, England turned to Russia. By the Subsidy Treaty of September 1755 England purchased the services of 50,000 Russian soldiers, who aligned themselves on the border of East Prussia, poised to advance into Prussian territory. The English did not regard the subsidy arrangements with Russia as incompatible with England's existing friendly relations with the Habsburgs. The arrangement was designed to give England double assurance.

Developments belied intentions. Prussia saw in the arrangements not a defensive move but a threat. Frederick already knew of a secret Russian–Austrian understanding against Prussia. His sense of security gave way to a deep fear that his state might well be sucked into an unwanted war against Russia. The way out, he reacted in his fear, was to undercut Russia by offering England the assurances concerning Hanover. Since Prussian troops would be appreciably closer to Hanover than the distant Russians, England readily accepted Frederick's proffered assistance. The two states accordingly signed the Convention of Westminster (January 1756), in which Prussia agreed to deny entry into and passage through Germany of any foreign troops (*i.e.*, Russians). England obtained the desired guarantees, and its relief was great. Frederick's was equally great. He saw in the Convention an instrument to keep the peace.

The Convention of Westminster

Westminster, however, did not produce the desired effect. It was a bombshell. Vienna was outraged, or professed to be, by the "treachery" of its English ally. Louis XV and Versailles bristled with indignation over the seeming defection of Prussia, and, overwhelmed by fears over what seemed an English–Prussian combination, France turned toward Austria as a shield.

Reversal of Bourbon-Habsburg enmity

France first allowed its alliance with Prussia to lapse and then contracted a defensive alliance with Austria, the First Treaty of Versailles (May 1756), by which France agreed to come to the assistance of Austria if the Habsburgs were attacked. To defend Austria, so French calculations ran, was also to protect themselves. Thus, out of this reflex of fear, France took the first long step toward a reversal of alliances and the abandonment of its centuries-old anti-Habsburg policy.

St. Petersburg's reaction was, if anything, more violent still. To Russian eyes, Westminster was clearly an unscrupulous repudiation of earlier English policy. Thus, a third set of fears obtruded on the diplomatic scene—Russia's fears that it might have to fight alone against Prussia covered by England. Russia thus turned to Austria for self-protection; in May 1756 Russia agreed to give Austria limited military aid in the event of a Prussian attack.

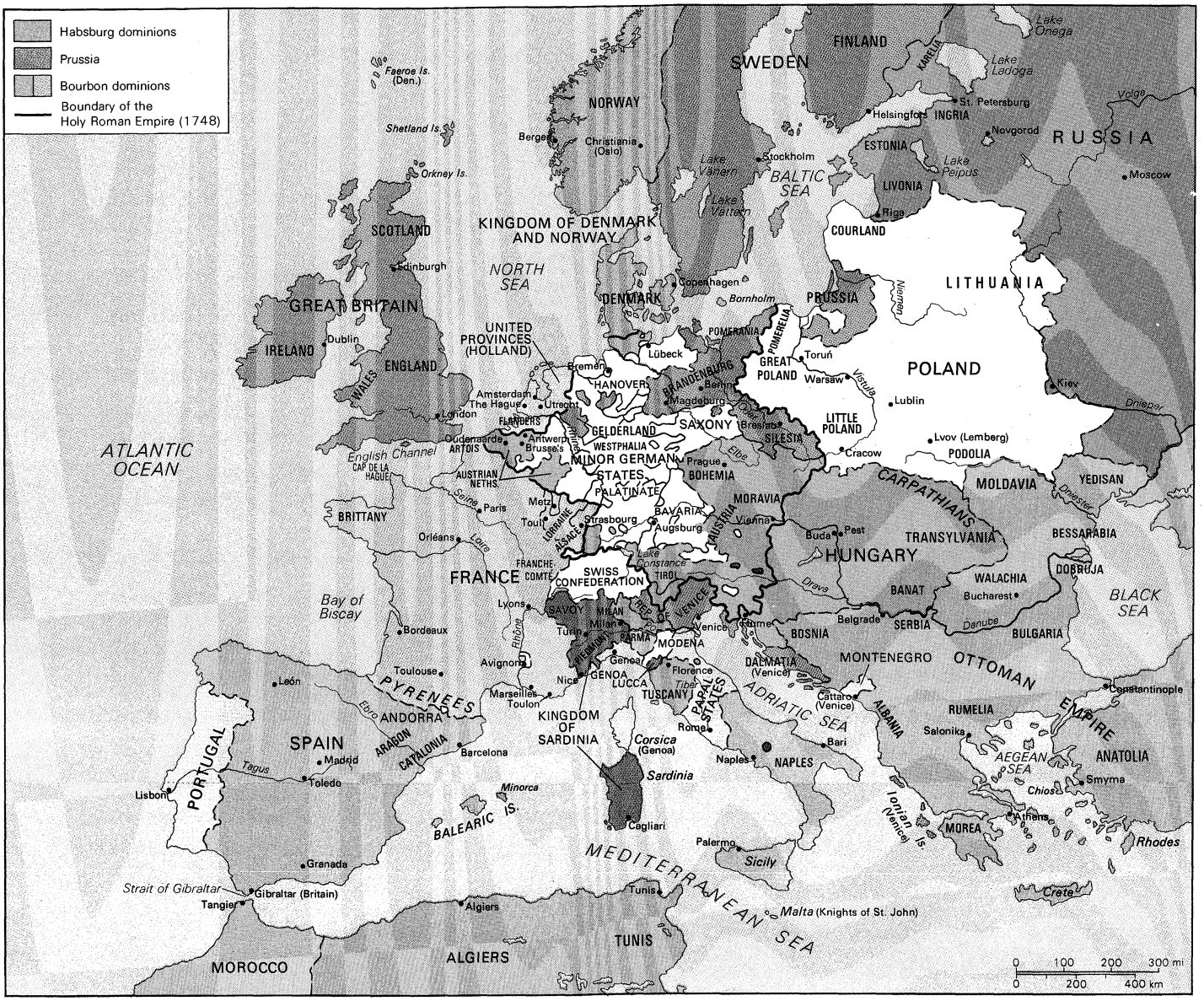
Frederick did not wait to be attacked first. In August 1756 he invaded Saxony, which he considered the centre of plotting against him. It was now too late for diplomacy: the generals had taken over. France contracted a full military alliance with Russia in the fall of that year. In the following spring, before the campaign began, France signed the Second Treaty of Versailles with Austria, a full offensive–defensive military accord. By this agreement, the last of the three bilateral pacts, the reversal of alliances had become a *fait accompli*.

This diplomatic revolution was a lesson in futility, revealing the fatal inadequacy of the balance of power. Seeking peace, the great states were led to create a situation beyond their control. Obsessed by reciprocal mistrust, they made the appeal to arms inescapable.

The Seven Years' War and the Peace of Paris. Frederick's invasion of Saxony opened the war on the Continent, which merged with the commercial–imperial conflict into the Seven Years' War (*q.v.*; 1756–63). In the European theatre, the Austrians at once invoked the provisions of the military agreements that bound them to France and Russia. Frederick, for his part, demanded and obtained English aid in the form of large subsidies and a small task force dispatched to the Continent. French tactics envisaged a rapid campaign for the conquest of Hanover. Military calculations in England were initially divided, veering after Frederick's great victories at Rossbach and Leuthen in 1757 toward a full military understanding with Prussia and the use of Hanover as the base of operations for a large-scale military diversion to contain the enemy.

Frederick's stubborn defense

Frederick's conduct of the war was heroic for its daring, flexibility, and dauntless courage against overwhelming odds. It was victorious, too, for a brief span of years. Operating on inner lines until 1759, he prevented the Austrian and Russian armies from joining forces. In 1759 the Russians inflicted a shattering defeat upon him at the Battle of Kunersdorf, and his future seemed hopeless. Yet somehow he held on, suffering defeat after de-



Europe in 1748.

feat but refusing to sue for peace. For three years the war in Europe was a protracted and ruinous combat in which the Prussian and Austro-Russian armies confronted each other in the east, while French and Anglo-Hanoverian troops effectively neutralized each other in the west. In Prussia's exhaustion Frederick could not have resisted indefinitely, particularly since England withdrew its financial support, but on January 5, 1762, "the miracle of the House of Brandenburg," as patriotic Prussian historians call it, or "her sacred majesty, chance," as Frederick put it, saved him. Tsarina Elizabeth died and was succeeded by her deranged nephew Peter III, an unmeasured admirer of Frederick. With Peter's accession the anti-Prussian coalition fell apart. After signing immediate peace with Prussia the Tsar restored all Russian conquests and placed the Russian army at Frederick's disposal. Though Peter was assassinated within a half year, Peter's wife and successor Catherine II the Great wanted no renewal of a war in which her country would be fighting against Austria for the benefit of Prussia. Prussia's eastern front was at last secure. "Thanks to providence," Frederick jubilantly cried out, "our back is free."

The major war in the Germanies now virtually ceased, but the fighting for empire continued on the seas in America, the West Indies, Africa, and India. The British conquests outside Europe propelled William Pitt the Elder, Britain's war leader, into the pages of history as the great empire builder. On the other hand, in France the dictatorship of Étienne-François, duc de Choiseul, not less embracing and intransigent than Pitt's, failed to win security for France. With rising opposition at home, defeats on the high seas, and the failure of a direct invasion of England and Scotland, Choiseul made overtures for peace early in 1760, not without first playing the Spanish card in his negotiations. Fortunately for England, Pitt's fears of Spanish might were exaggerated. Spain suffered crushing defeats on all fronts in its brief military venture.

After these reverses nothing remained but for Choiseul to sue for immediate peace, which was signed on February 10, 1763. The end of the war between England and France did not come, however, until John Stuart, 3rd earl of Bute, the chief English negotiator, had already taken his country out of all fighting on the Continent. He broke completely with Prussia, cutting off all subsidies and leaving his erstwhile ally to its own resources. Austria, too, was then ready for negotiations, for after Catherine's withdrawal the Habsburgs stood alone in Europe.

The Treaty of Paris was less severe than might have been expected in view of Britain's overwhelming victories. The peace was the handiwork of Lord Bute and of still lesser men who had forced Pitt from office and were prepared, to secure their advancement, to sacrifice

if necessary some of the imperial interests that Pitt had so tenaciously and energetically pursued. The guiding consideration of Bute was to turn the diplomatic negotiations to the account of his young royal master, George III. The tactics of the imaginative Choiseul, the chief French negotiator, were to drive a wedge between the British peacemakers and exploit to the full their fear that Pitt might be returned to office and reopen the war. For Choiseul the peace was to be only a respite from actual fighting, an armistice so to speak, which France would terminate when it was again prepared to reopen the duel for overseas empire. Though he made bitter concessions to the victor, he saved France from the status of a second-rate power, which Pitt had planned. Though Britain did not obtain the Carthaginian peace that Pitt would have imposed, it had emerged as the leading commercial and colonial power in the Western World.

The emergence of Britain's colonial supremacy

In North America, Britain became master of the vast territory east of the Mississippi as well as "the frozen acres" of Canada. Patriots angrily clamoured that their country also retain the captured West Indies sugar islands; however, despite the great strategic naval importance of Guadeloupe and Martinique, the sugar bloc in Parliament sacrificed imperial interests to financial profits and voted their return to France. Moreover, France was permitted to keep the island of Gorée, commanding the slave trade of Senegal, along with a number of trading posts on the eastern coast of India.

Spain, too, fared better by the peace terms than might have been expected by virtue of its overwhelming defeat. The Spanish lost Florida to England and gave up their old claims to the Newfoundland fisheries, but they kept their South American colonies and islands in the West Indies. The victor restored Cuba and the Philippines to Spanish rule. Choiseul, in order to place a lien upon his Bourbon ally's goodwill, ceded to Spain that part of the French territory of Louisiana, including the port of New Orleans, which lay west of the Mississippi.

Prussia also ended the war a victor, for the separate Treaty of Hubertusburg (February 15, 1763) formally recognized Frederick's conquest of Silesia. In Prussia there was great devastation and suffering, but that did not last. What lasted was "the miracle of the House of Brandenburg," the survival of Prussia, its new status as a great power, and the continuing confrontation of Prussia and Austria for mastery of Germany.

Renewal of Anglo-French competition. Choiseul was the directing intelligence behind his country's foreign and colonial policy during the 1760s. He counted heavily on the Habsburg alliance of 1756 and on the network of Habsburg–Bourbon dynastic marriages to cover French security on the Continent and leave France free for action outside Europe. Working via the Pacte de Famille with Spain, Choiseul looked ahead toward a joint Bourbon thrust against England and bent every energy toward restoring the shattered might of French armed services and strengthening the bases for naval operations. This aggressive policy could not be implemented, however, and under Charles Gravier, comte de Vergennes, who took over in 1774, it was toned down. On the Continent Vergennes relied less on the Habsburg alliance (which he feared would entangle France in Austrian intrigues) and more on developing friendly relations with Prussia, with the German states along the Rhine, and, most of all, with Russia. There was appeasement in this approach, but he was prepared to pay the price of lessened French prestige and influence to obtain economic benefits. "In the present state of affairs," he declared, "commercial questions are political questions and as such are within the province of foreign affairs."

The U.S. War of Independence

Vergennes's great opportunity to strike directly against England came in 1776 with the armed resistance of the American colonies to British rule. Initially, he adopted a policy of open non-belligerency against England and more or less secret assistance for the Americans. The decisive change from secret aid to open alliance came after the U.S. Declaration of Independence in July 1776 and Gen. John Burgoyne's defeat at Saratoga in 1777. In February 1778 Louis XVI signed treaties of alliance and commerce with the Americans, and within a few months France was openly at war with England. Vergennes had not miscalculated on the Bourbon ally. In 1779 Spain contracted an open alliance, contributing with the United States, 40 ships of the line that gave the associates naval superiority. In return, Spain obtained from Vergennes a reluctant agreement not to withdraw from the war until Gibraltar had been retaken from the British.

In 1780 England declared war against the United Provinces, who had built up a lucrative trade conveying supplies to the colonists and allowing American privateers to refit in Dutch ports both in Europe and the West Indies. But Vergennes appealed to the neutrals to encompass England's defeat by turning them against "the tyrant of the seas." Here his warm relations with Russia stood France in good stead, and Russia was persuaded to sign the declaration that established the League of Armed Neutrality. By the terms of the declaration, contraband was narrowly restricted to munitions and arms, and the armed neutrals regarded themselves as free to navigate along the coasts and to the very ports of all belligerents. Vergennes had gone far toward arraying the Continent against England.

The peace treaties of 1783, signed by England, France, Spain, and the Netherlands, ended the fighting. The treaties recognized the independence of the colonists. Spain retained Minorca and Florida. To France England ceded Tobago and Senegal. England retained bitter memories of a long and humiliating series of naval reverses. Though Adm. George Brydges Rodney's victories in 1782 restored English superiority, the antiwar mood of shame and humiliation was deep. The English wit Horace Walpole caught it in a striking phrase: "You must be happy now," he wrote, "not to have a son who would live to grovel in the dregs of England." England's enemies were jubilant, more jubilant than reality warranted. "Thus is the great power . . . fallen completely and forever," wrote Leopold of Tuscany, ". . . sunk to the ranks of a second class power. . . . France, delivered for all time of that formidable adversary, thereby doubles her intrinsic strength, her commerce, her authority, and her prestige. . . ."

The growth of French commerce

France did not double its "intrinsic strength," least of all in the military affairs of eastern Europe. With respect, however, to the French commercial position between 1783 and 1789, Vergennes's policy did add greatly to the strength of the country. He negotiated a trade treaty with Russia that gave French merchants terms as good as those given to the English. By opening trading facilities to foreigners in its West Indies colonies (1784) and concluding a practically free-trade treaty with England in 1786, France accelerated the remarkable increase in the volume and value of its commerce. And England was far from "fallen completely and forever." By skillful diplomacy England formed a triple alliance with Prussia and the United Provinces. As an ally of the United Provinces, England was assured of navigation on the Scheldt and a point of entry for trade with the Continent. As an associate of Prussia and in possession of Hanover, England could build a trading clientele in the Germanies. A dominant member of the alliance and Russia's greatest customer, England pursued active trade relations in the North Sea and the Baltic. With Gibraltar in their hands and with the support of Portugal, the English also held Spain in check. By supporting the territorial integrity of the Porte (Ottoman government) they had access to the eastern Mediterranean, and their share in the trade with free America surpassed the total reached before 1776. On the eve of 1789 the two maritime states still confronted each other in the competition for empire.

Aggression in the East. For the better part of a decade after the Peace of Paris, a civil war within Poland and renewed warfare between Russia and Turkey again threatened the peace of Europe. The immediate crisis began in 1764. After the death of the Polish king Augustus III in the preceding year, Prussia and Russia reached an understanding to secure the election of a native nobleman, Stanisław August Poniatowski (Stanisław II), who presumably could be counted on to fit the

role of convenient tool for their ambitions. They also agreed to "defend the peace of the Republic" and protect the religious rights of Greek Orthodox Catholics. By virtue of this understanding, Russian troops occupied Poland and imposed upon the Poles a pact that named the empress Catherine II as "Protector" of Polish laws and liberties. Rejecting this intervention, a group of ardent Polish patriots and Catholics organized a political and religious uprising against the Russian "Protector." While civil war raged within Poland, the Russians inadvertently became involved in war with the Ottoman Empire (1768), whom France had encouraged to support Polish liberties.

Veteran Russian troops quickly crushed the Turkish forces and occupied the two Danubian provinces of Moldavia and Walachia. Here was an immediate threat: should Russia's neighbours, particularly Austria, permit Russia to cross the Danube and overrun the Balkans without getting any territorial compensation? If Austria ordered its troops to attack the flank of the invading Russians to enforce its claim to a share of the Turkish territory, most likely Russia would have to fight Austria alone without the military support of her Prussian ally, for Prussia needed peace to recuperate from the Seven Years' War. With that consideration in mind the European forces veered away from the idea of despoiling the Ottoman Empire toward the project of partitioning enfeebled Poland instead. Russian troops, therefore, did not cross the Danube; Turkey was neither totally crushed nor partitioned; and the general peace was preserved. There would thus be no need to fight over the spoils. An amicable arrangement, mutually advantageous, could solve the Polish problem without involving the powers in war.

The first partition of Poland

A solution along those lines was casually suggested by Prussia, and within a year and a half, in the summer of 1772, the details were worked out; three separate treaties were signed, which specified the respective shares of the contracting powers. By this first partition, Poland lost nearly one-third of its territory and almost one-half of its population. Frederick's share was the smallest in land and population, but it was valuable for its strategic location and political worth. It closed the gap that separated the Prussian provinces, so that Prussia became one continuous stretch of territory between the Elbe and Lithuania, with control of the basin of the Lower Vistula and its enormously important grain trade. To Catherine went the eastern part of the old state of Lithuania. The chief importance of that territory was military, but it also gave Russia political as well as economic control over those eastern Polish provinces that had been penetrated by Russian traders for a full century. To the Habsburgs went 2,750,000 new subjects and a huge stretch of land that included Galicia, the northern slope of the Carpathians, and a corridor through Silesia to the new possessions.

Nevertheless, the partition of Poland set precedents for similar action against the Ottoman Empire. Continuing Russian military successes in the war begun in 1768 compelled the Porte to sue for peace. By the treaty of Kücük Kaynarca in 1774, the sultan recognized the independence of the Crimea, which the Russians had conquered, and ceded three of its great fortresses. The sultan paid a heavy war indemnity, opened the Black Sea to Russian navigation, and gave Catherine II a commercial treaty with the most-favoured clause. Russia restored the Danubian principalities that its armies occupied, but Catherine reserved the ominous right to intervene in their behalf. The Russians also won the right to build an Orthodox church in the foreign quarters of Constantinople, along with a vague and therefore doubly dangerous right to make representations in behalf of its worshippers.

In 1775–76 the Holy Roman emperor Joseph II acquired Bukovina and Fiume, which gave Austria outlets for trade to the north and west. In 1777 he troubled the peace again by seeking to acquire succession rights to the duchy of Bavaria. The attempt failed because Vergennes ruled that the alliance clause was not involved and because Prussia did its utmost to block a move that would have given Austria military roads from Bavaria to the Tirol and north Italy. Russia obtained additional prestige and influence from this dispute, serving as mediator (1779) between Prussia and Austria—that is, mediator within the internal affairs of Germany.

Stopped in that venture, Joseph allied himself by treaty with Russia (1781). Temporarily blocked in the west, he was associating himself with Catherine II in her "Grand Plan" for the partition of European territory of the Ottoman Empire. During a lull in the pressures in the east, Joseph again felt his way in the west. He sought to win consent to his effort to have the Scheldt opened to Belgian commercial navigation. He was thwarted in that venture as well as in the concurrent move to exchange the Belgian provinces for Bavaria. Neither Vergennes nor Catherine supported him, while Frederick of Prussia organized a League of Princes (1785) to block an arrangement that could have rounded out the Habsburg kingdom.

Russian occupation of the Crimea

Meanwhile, the Russian attrition of Turkey continued. In 1783 Russia occupied and unilaterally annexed the Crimea, which the Ottoman Empire was forced to recognize by treaty. A year later the Russian colossus occupied Georgia, there, too, declaring that it was "Protector." By 1787 the Ottoman Empire had had enough and declared war on Russia when the latter rejected a Turkish ultimatum for the restitution of the Crimea. Russia appealed to its new Austrian ally, and Joseph responded by deploying his troops on the Turkish line of the Danube. Joseph himself was soon to hasten home when word came that a revolt against his rule had broken out in Hungary, but the two predators were on the loose. This is where matters stood when the outbreak of the Revolution in France in 1789 threw old diplomacy and issues into the discard. All the great powers had fared well in the late 18th century; the only victims were the lesser and weaker states and the home populations that footed the bill for militarism, navalism, and balance of power diplomacy.

THE FRENCH REVOLUTION AND NAPOLEON

War with Europe. The Revolution in France revolutionized the nature and scope of international relations. The deputies of the National Assembly (1789–91), the Revolutionary government of France, had no intention of making war on the great powers. On the contrary, they voted a decree (May 22, 1790) formally renouncing all wars of conquest and promising that new France would never use its troops against the liberty of others. Out of the great reconstruction, however, arose frictions that within three years led to a war that was to be waged intermittently for almost a quarter of a century.

Conflict with the papacy

There was antagonism between the papacy and the assembly over the Civil Constitution of the Clergy (July 12, 1790), which sought to harmonize the relations of the secular clergy with the new civil administration. After protracted negotiations, the papacy condemned the constitutions and later declared the entire Revolution anathema. Pius VI also condemned the incorporation of the papal enclave of Avignon into France as an unjustifiable application of the principle of popular sovereignty in the conduct of international relations.

Other issues drew objections from the Holy Roman Empire. Various powerful German princes, who held feudal rights over certain territorial possessions in Alsace, appealed to the Emperor in protest against the application of the antifeudal decrees passed by the National Assembly. From the French point of view, since Alsace was politically part of France, the princes were pleading for special rights of extraterritoriality. So the National Assembly ruled against them.

Disputes notwithstanding, the powers did not use them as pretexts to intervene in the internal developments of France. Nor did they so use repeated secret appeals for help from the French rulers and the émigrés. Though the Emperor signed the preliminaries of a defensive military alliance with Prussia, he otherwise limited himself to a war of nerves, repeating that the restoration of the authority of the French monarchy was a matter of common concern to the powers.

Direct responsibility for going beyond this war of nerves lay with the bloc of Girondin deputies (then known as Brissotins) in the new Legislative Assembly (which replaced the National Assembly) who deliberately initiated a policy to provoke Austria into intervention and war. They were not alone in wishing that course to be followed. They themselves counted on a victorious war that would unmask what they called "the treason" of Louis XVI and Marie Antoinette and give them control over French policy. Military leaders like the Marquis de Lafayette, Louis, comte de Narbonne, and Charles-François du Périer Dumouriez, who also desired war, did so out of their conviction that France would be defeated and a grateful nation would then turn to them.

Leopold, while still holding to his tactic of verbal denunciation of "the pernicious sect of Jacobins," nevertheless tightened his military alliance with Prussia, while the Prussian king Frederick William II went further and drew up a plan of campaign. The Brissotins then presented an ultimatum to the Emperor, summoning him not only to renounce the Austrian–Prussian alliance but also all other treaties that threatened "the sovereignty, independence, and security of the nation." They counted on Leopold's rejection, and when the response of Leopold's son and successor, Francis II, was held "unsatisfactory" on April 20, 1792, France declared war on "the king of Bohemia and Hungary." The deputies defined the war as "the just defense of a free people against the unjust aggression of a king."

The early fighting against the Austrians was disastrous on the northeast front, and Prussia's entry in the war greatly intensified the military danger. As the steady Prussian advance brought their troops onto French soil and as the great fortress of Verdun fell into their hands, it seemed as though only a miracle could save Paris and the entire Revolution. The "miracle" occurred on September 20 at Valmy in the passes of the Argonne. The raw French recruits held firm against the Prussian line while their artillery fire stopped the enemy advance. Though only a minor skirmish, Valmy was in effect the decisive watershed of the war. The Prussians yielded their position and began a long retreat back to the frontier. As they retreated, the French advanced in a three-pronged offensive. Savoy fell to them, the Middle Rhine, too, and most important, the Austrian Netherlands, which Dumouriez overran.

The fall of Verdun

The problem then arose in all the conquered territory of future relations between the Gallic liberators and the pro-French native sympathizers. The solution came in a momentous declaration of policy. By the decree of November 19, 1792, the new assembly of the French Republic declared "in the name of the French nation that it will bring fraternity and aid to all peoples that wish to recover their liberty. . . ." This memorable decree threw down the gauntlet to the powers. It proclaimed an ideological crusade against the old European regime of kings and privileged order of society. A complementary decree (December 15) put teeth into the declaration of policy. It stated that in all occupied territory the system of taxation, the tithe, the titles of nobility, and all special privileges and feudal dues would be abolished, and all property belonging to local rulers was to be taken over by the agents of the French Republic. These self-styled "Friends of Liberty" would institute the new administrative system, and all who resisted them would be treated as enemies.

This mystique of a crusade, coupled with the older *politique* of conquest and annexation, laid down the bases for the war of Revolutionary France against all European states, immediately against England. Threatened by the French occupation of the Austrian Netherlands and by the opening of the Scheldt to French commerce, England prepared to fight. The Convention declared war on February 1, 1793, including the United Provinces in the declaration. A month later Spain, too, was at war, then (in rapid succession) Sardinia and Naples, the lesser Italian States, Denmark, Sweden, and the Ottoman Empire. England contracted treaties of alliance and subsidy with the member states of this First Coalition.

War with England

From a limited war between France and Austria, the conflict had broadened into a confrontation over two conceptions of human relations.

The defeat of the First Coalition. The fighting was a succession of defeats on all fronts. With Paris again exposed to direct attack and England declaring all France under a state of blockade, the Convention rescinded the decrees of November and December. The allies refused to negotiate: they were now discussing plans for a partition.

At this desperate juncture the Convention voted the decree of the *leveé en masse* (August 23), which was designed to transform France into an armed camp. "All Frenchmen," it read, "both sexes, all ages are called by the *patrie* to defend liberty." Implemented systematically, this declaration of total war released untapped vital energy, and the tide of battle turned to victory. By the last days of 1793 the invaders were beaten back on all fronts. Toulon was recaptured from the English, a young captain of artillery, Napoleon Bonaparte, distinguishing himself in the final attack. The Spanish withdrew to the Pyrenees and Savoy was cleared.

The élan of 1793 carried the Republican troops to victories in 1794. As the armies repulsed the invaders, they took the offensive. The line of the Rhine was regained. The Republican armies overwhelmed all of Italy. The Dutch fleet surrendered. Prussia was the first of the allies to sue for peace in the autumn of 1794. The supporters of a separate peace in Prussia reasoned that to give up territory along the left bank of the Rhine and later to receive compensation within Germany by the terms of a general European peace was a satisfactory end of the fighting.

For France the Peace of Basel (April 1795) was fully acceptable: it kept open the line of armed advance along the Rhine. To Prussia it brought the end of war and the highly advantageous status of being a neutral. By the Treaty of The Hague (May 1795), the United Provinces was compelled to cede strategic bits of territory and pay a heavy war indemnity. The reorganized Dutch government, called the Batavian Republic, was then bound to France by an alliance that drew the Dutch and their great resources into the war against England. A month later, in June, it was the turn of Spain, which had to recognize the French Republic and cede to it the Spanish part of Santo Domingo (Hispaniola). The final step taken by the French annexationists (in October) was to vote the incorporation of the Austrian Netherlands as an integral part of the territory of the Republic. The language of the French crusade to liberate enslaved peoples was kept in all these transactions; but the deeds of the liberators resembled nothing so much as the play of old-fashioned power politics.

Annexation of the Netherlands

The government of the Directory, the conservative government that came to power in France in 1795, set its sights on a general European peace. The two allies still in the field were also eager for a comprehensive settlement, but they were not ready to accept the preestablished terms set by France. So the Directory prepared to continue the fighting. Since victory on the seas was unlikely, the French plan of campaign called for limited naval operations. The main thrust was to be on the Continent. The major attack was directed against Austria, where in a broad pincer movement two great French armies were to converge upon Vienna. A minor supporting campaign was envisaged in Italy, where the assignment given to Bonaparte, who was in command of the Army of Italy, was to conquer Lombardy. The ultimate political aim of these military operations was to compel defeated Austria to retain Lombardy in return for its acceptance of France's retention of the annexed Austrian Netherlands.

It was not only the military genius of Bonaparte but also his views concerning the territorial settlement that completely upset the plans of the Directory. As he was winning his resounding victories in northern Italy in 1796 and 1797, his thinking ranged far beyond the objectives of his government. He agreed that the conquered Rhineland would be retained. He would not,

however, acquiesce in installing Austria in conquered Lombardy. He intended to retain French military control over the territory that he was reorganizing as the Cisalpine Republic and to use it, along with French-occupied territory on the Adriatic and in the Ionian Islands, as a stepping-stone toward Egypt. For he planned to turn next to the east and strike at England by severing its imperial communications.

Treaty of Campo Formio

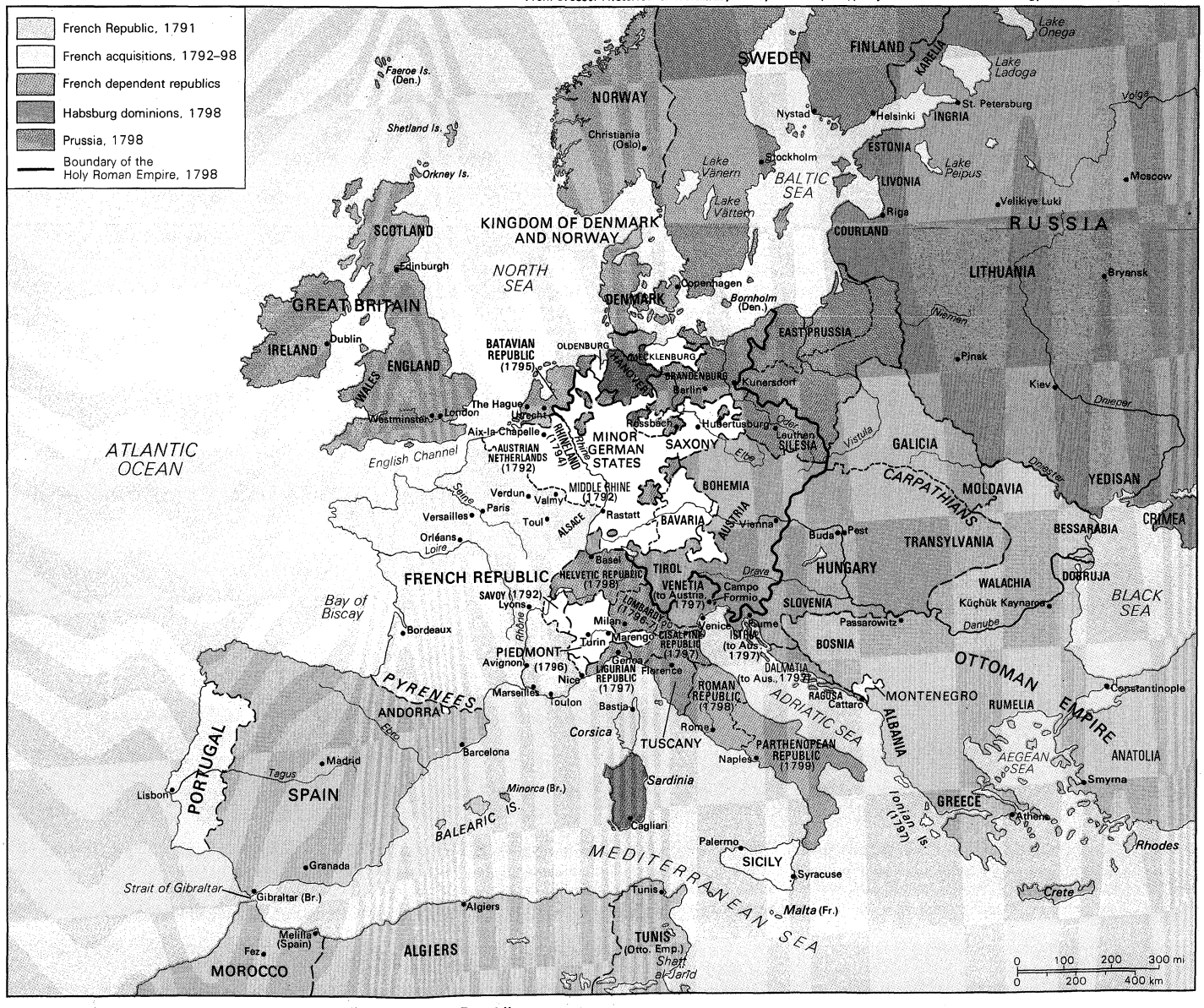
With such projects in mind he carried on negotiations with Austria in the summer of 1797 that eventuated in October with the Treaty of Campo Formio. Austria recognized the French annexation of the Austrian Netherlands and agreed (in secret articles) to cede approximately two-thirds of the imperial territory on the left bank of the Rhine, provided that this cession were ratified at an imperial congress to be held at Rastatt. The ousted German princes were to receive compensation elsewhere within the empire. Prussia, however (and this, too, in a secret article), was not to share in the dismemberment as the Treaty of Basel had stipulated. In Italy, Austria received Istria, Dalmatia, and the Venetian mainland in return for the loss of Lombardy. The Cisalpine Republic was formally recognized by Austria. The Ionian Islands that Bonaparte was so anxious to gain went to France.

Such were the main provisions of Campo Formio that the Directory reluctantly ratified. It was an imperialistic peace, a sorry end to the crusade for liberty. The French statesman Emmanuel-Joseph Sieyès spoke wisely when he said: "This treaty is not peace; it is a call to a new war."

The Second Coalition, 1798–1802. The Peace of Campo Formio was an interlude. While Bonaparte went off on his romantic odyssey to wrest Egypt from the Porte (Turkish government), in his absence the agents of the Directory mercilessly victimized the satellite republics set up by the military commanders. Threatened again by all these acts, Austria, England, and Russia formed the Second Coalition (1798), to which the Porte, Naples, and Sardinia adhered. Though in the early months of 1799 the allies pushed back the French on several fronts and by mid-year republican defenses were crumbling, the allied victories were transitory. By early fall, when Bonaparte returned from his Egyptian fiasco, French victories had dispelled the military danger. France was secure against invasion and could negotiate for peace from strength.

The great majority of Frenchmen yearned for peace, but not Bonaparte. He needed a major military victory to consolidate the power he had seized by the coup d'etat of Brumaire in November 1799. He drew up plans for the campaign of 1800. Once again, as in 1796 and 1797, his spectacular victories in Italy were decisive. And again, as four years earlier, Austria asked for peace. Negotia-

From *Grosser Historischer Weltatlas*, vol. 3, *Neuzeit* (1972); Bayerischer Schulbuch-Verlag, Munich



Europe in 1798 and the expansion of the French Republic.

The Treaty of Lunéville

tions ended with the Treaty of Lunéville (February 1801). The salient terms of Campo Formio were reaffirmed. The piecemeal ousting of Austria from the Italian peninsula continued, but more important than its expulsion from the Duchy of Tuscany was the Habsburg recognition of the new Italian republics as sovereign powers.

Only England remained at war, and the English government had to yield to the insistent clamour for an end to the struggle. Peace talks were begun a month after Lunéville. The preliminaries of October 1801 were converted into the definitive Treaty of Amiens on March 27, 1802. The provisions were harsh. Egypt was restored to the Ottoman Empire. Both Malta, which went to the Knights of St. John, and the Cape of Good Hope were restored under conditions that made their future security precarious. Save for Ceylon and Trinidad, England restored its maritime conquests. Moreover, England tacitly accepted the main territorial changes on the Continent as they had been agreed upon at Lunéville. Amiens contained no provisions for a favourable commercial treaty that would reopen the continental markets for English goods.

In the circumstances, a chorus of recrimination fell upon the government. England obtained, so it seemed, only what Bonaparte gave it: peace on his terms. After ten years of war England's great rival and competitor emerged as the most powerful state in Europe. "One may state without the slightest exaggeration," Talleyrand was to write later, "that at the time of the Peace of Amiens, France enjoyed abroad, thanks to its military supremacy, such power, glory, and influence that the most ambitious person could desire nothing more for his country."

Bonaparte desired more. The conqueror who had needed peace in 1801 no longer needed or desired it two years later. His head was filled with vast projects, and to realize them he opted for war. War, he often stated, was essential to his plans.

Napoleon's grand design. The provocations by which Bonaparte deliberately forced England in self-defense to rupture the Peace of Amiens in 1803 reopened the coalition warfare that did not cease until the Treaty of Tilsit in 1807. Within the span of those four years Napoleon (as he was known after he established the French Empire) won his greatest victories at Ulm, Austerlitz, Jena, and Auerstädt, crushing Austrian and Prussian resistance. He also defeated the Russian armies at the battles of Eylau and Friedland, but he could not vanquish Russia itself.

Napoleon's Continental System

In those years he pursued against England that ruthless commercial warfare that he called the Continental System. To strike at England's prosperity and power by denying it entry of colonial products and manufactured goods was not new. What was new was his project, by a self-blockade, of destroying England completely by closing the markets of the entire Continent that had fallen under his domination. He would reduce England by "glutting her with excess." Such was the sense of the decrees of 1806 and 1807 by which he coordinated commercial war with conquest in Europe. The British tried by one means or another—smuggling, licenses, coercion of neutrals—to keep the channels of trade open.

In these circumstances, the role of Russia was crucial, and it was to enlist Russian support that Napoleon made peace and arranged an alliance with the colossus whom he could not decisively defeat in battle. This reversal of alliances was made at Tilsit in 1807. For Alexander I the gains of Tilsit were immense, or seemed so at the time. His position of a half-defeated enemy was transformed into that of an ally of the French emperor. The immediate gains of Napoleon were greater still: in return for recognizing Russia's special interests in eastern Europe he obtained an associate for his intention; the preamble to his treaty with Alexander put it "to assure the happiness and tranquillity of the globe."

For a "grand design" was maturing in Napoleon's mind. He had become the embodiment—and the victim—of a seductive vision. He brought to that design the force of his powerful personality, his traits of character, and his early intellectual conditioning. Into it entered a Plutarchian conception of the hero of classical antiquity. The example of Alexander the Great, who had striven to attain world domination, was a component of his grandiose dream; and perhaps most of all the idealized memory of the career of Charlemagne, who had restored the Roman Empire in the West, fashioned his thinking. He conceived himself as a new Charlemagne, the head of a European empire permeated by the sober ideals of the French Revolution, as he saw it, and resting upon an enlightened institutional rearrangement of state and society that would do away forever with the injustices and inequities of the lingering old order in Europe.

Napoleon's vision

He referred publicly to the Grand Empire that was to replace the moribund Holy Roman Empire. While kingships, duchies, and fiefs were established (none of them to survive his defeat) for members of his immediate family and his high civilian and military subordinates, the most momentous changes took place in the Holy Roman Empire itself. They had been initiated in 1803, when he secularized the ecclesiastical state. They continued with the extinction of the medieval city-states and the possessions of the imperial knights. The climax came in 1806. The surviving large secondary states became members of the Confederation of the Rhine of which he was protector, and it was these states that renounced their loyalty to the empire. There was now an empire without subjects, and on August 6 the Austrian emperor Francis II foreswore his titles and prerogatives. Now there was no old empire. The Grand Empire of Napoleon replaced the 1,000-year-old Holy Roman Empire.

For several years more Napoleon extended the boundaries of imperial France and the Grand Empire. In 1810 imperial France stretched from the Elbe to the Pyrenees and from the English Channel to Rome. In the south it included Genoa, Tuscany, Parma, Piacenza, the states of the church, and the Illyrian Provinces on the Adriatic. In the north it had incorporated parts of Hanover and Prussia, the Hanseatic cities, and all of the Netherlands. Outside imperial France, northern and central Italy fell under the authority of the Grand Empire, as did Spain, Naples, Westphalia, Frankfurt, and the new Duchy of Berg on the Rhine. Napoleon himself was king of Italy, protector of the Confederation of the Rhine, and mediator of the Swiss Confederation (now called Helvetia). Through the king of Saxony he held effective control over the Grand Duchy of Warsaw. Prussia and Austria, both nominally independent, were bound by treaties of alliance. Only England, Russia, Portugal, and the Porte lay outside Napoleon's control.

This vast Grand Empire rested on no firm structural foundation. With each apparent increase of its power Napoleon underscored its personal character. He never fully elaborated any single principle of political organization except that of hierarchical paternalism. The son whom the Habsburg princess Marie-Louise bore him in 1811 was to be the guarantee of the permanent rule of his new dynasty. But by then the tide was running sharply against him.

Resistance and opposition. The working of the Continental System inflicted heavy losses, but the grim years of English endurance were also years of increasing hardship for the French, too, years of swelling opposition. Resistance and opposition rose even more outside France.

Revolt of Spain

The first great revolt came in Spain, where religious and patriotic revolt broke out against French rule. Napoleon had counted on a small task force to wrest and hold military control. The nationalist uprising, the "cancer," as he called it, ravaged Spain for six terrible years. To contain it he had to invest many thousands of his veteran soldiers.

The reverses in Spain opened the Iberian Peninsula to an English expeditionary force under Sir Arthur Wellesley (later duke of Wellington), who, from his fortified lines in Portugal, methodically penetrated Spain.

Defeats in Spain and the sapping of French military strength ignited combustible hatreds in central Europe. They fired the fighting of 1809, when the Austrians pro-

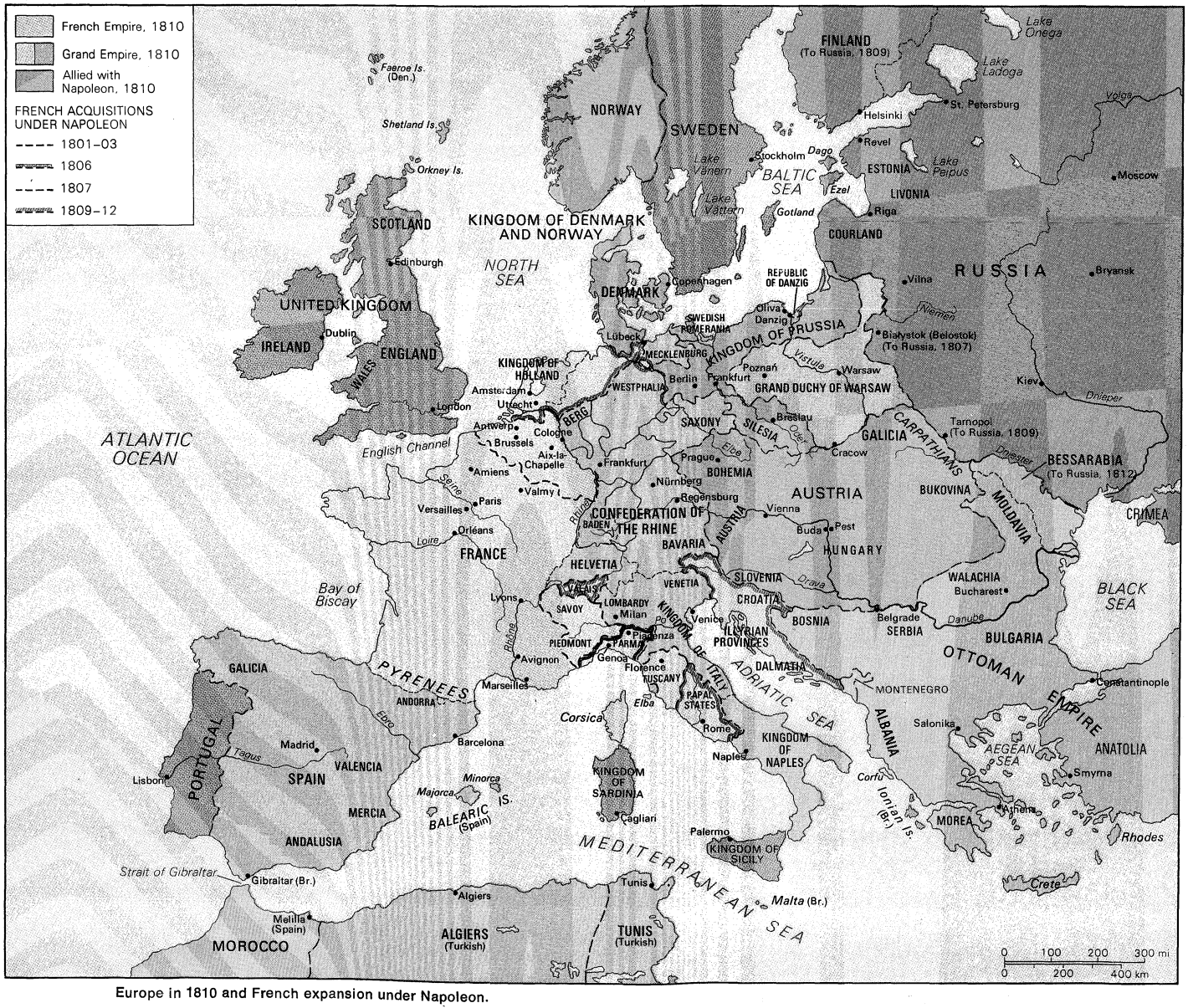
claimed a "war of national liberation." To his troops, as he took the field, the young Archduke Charles addressed a ringing appeal: "The liberty of Europe has taken refuge under your banners . . . your German brothers, now arrayed in the ranks of the enemy, await liberation at your hands."

The Austrian uprising was premature. Napoleon overwhelmed the enemy at Wagram and dictated onerous peace terms. He could not, however, arrest the preparations that Prussia was making for its emancipation. While the German philosopher Johann Gottlieb Fichte was penning his resounding *Reden an die deutsche Nation* (*Addresses to the German Nation*), resolute civilian reformers, among whom Karl, Freiherr vom Stein, and Karl von Hardenberg were the foremost, put down the foundations for a remarkable reconstruction of Prussian administration and pursued a policy of social regeneration. Under the guidance of Gerhard Johann David von Scharnhorst and August Wilhelm Anton, Graf Niedhardt von Gneisenau, military reformers rebuilt the armies that Napoleon had earlier shattered.

Break with Russia

The break with Russia was the catalyst of the final collapse of Napoleonic domination. The seeds of conflict had been sown in the very arrangements of Tilsit, by which France agreed to further Russian interest vis-à-vis the Porte and Sweden, and Alexander, conversely, promised his support of Napoleon's great design. Though it became apparent soon enough that neither partner was acting in good faith, strained relations were patched up after a fashion at a meeting at the summit at Erfurt in 1808. The Tilsit provisions were reaffirmed; so, too, were misgivings and secret hostile reservations. Relations were maintained on that level of mutual mistrust until the final break in 1812. The specific grievances concerned the future settlement of Poland and the European territory of the Porte, but the break came because in the last analysis there was not room enough in Europe for two competing imperialisms. Napoleon never doubted that he would win out. "In five years," he declared, "I shall be master of the world. Russia alone remains and I shall break her in pieces." In that expectation (of 1811) he invaded Russia on June 24, 1812.

Napoleon's collapse. The invasion of Russia ended with the holocaust of the Grande Armeé. Of almost 400,000 men who had crossed the Niemen into Russia, fewer than 100,000 struggled back to safety. The rest were dead or wounded, deserters or prisoners. With incredible rapidity Napoleon raised a new army with which he planned to renew the fighting. The enemy gave him no pause. Alexander made the fateful decision not to conclude a separate peace but to continue the war and liberate Europe. Prussia renounced its alliance and, in the early spring of 1813, before the campaign opened, declared war on France. The Austrians moved more



Europe in 1810 and French expansion under Napoleon.

slowly, Chancellor Metternich combining military preparations with diplomatic manoeuvres to gain Napoleon's acceptance of the peace terms that he offered. When those negotiations broke down during the armistice between the spring and autumn campaign, Austria, too, declared war. The spring campaign of 1813 was indecisive. The fall campaign was a disaster as calamitous as the Russian debacle. The troops that succeeded in escaping from Leipzig barely cut their way back to the Rhine.

The victors at that "Battle of the Nations" (Battle of Leipzig), their contacts established with the pro-peace leaders within France, invaded France in December—not, however, without first addressing an appeal to the nation over Napoleon's head. Meanwhile, victorious at Vitoria in Spain, Wellesley stood ready to invade France from the southwest. The net was tightening. In the spring campaign of 1814, fought on French soil, Napoleon won brilliant victories, but he lost the war to the allied invaders, whom the English foreign secretary, Viscount Castlereagh, barely kept united. By the Treaty of Chaumont (March 1814) Castlereagh also laid the foundations of the Quadruple Alliance, which, it was hoped, would enforce the peace once it was made. When Alexander and King Frederick William entered Paris on April 1, the true fighting ended. Napoleon's final effort at Waterloo a year later, after his escape from Elba, was in the nature of a quixotic game too late to have either lasting political or military significance.

For the military events of the period from 1789 to 1814 see FRENCH REVOLUTIONARY AND NAPOLEONIC WARS.

(L.Ge.)

III. 1815–1914

THE AGE OF METTERNICH AND THE CHALLENGE OF 1848

The Congress of Vienna. For a quarter of a century France had kept Europe in turmoil. The French Revolution and its Napoleonic sequel had introduced a new liberal ideology into Europe and had boosted the national consciousness of its peoples. In 1814 France was finally defeated, and peace was restored by the Treaty of Paris. The Bourbons returned as the legitimate heirs to the French throne.

The restoration of the "legitimate" order—that is, the order existing in Europe before 1789—was the avowed intention of the victors who met in Vienna in 1814. The principal delegates to the Congress were four: Metternich for the Habsburgs, Viscount Castlereagh for Britain, Tsar Alexander I for Russia, and Talleyrand for France. Napoleon's brief return to power during the Hundred Days did not materially affect the decisions of the Congress. A second Treaty of Paris, somewhat more severe than the first, once more made peace with France.

In broad terms the Congress restored the former order, yet not without some qualifications. The Holy Roman Empire, ended by Napoleon in 1806, was replaced by a simplified German Confederation. The ancient republics of Genoa and Venice were absorbed: the former into the Kingdom of Sardinia, the latter, together with Milan, into the Habsburg domains.

The Habsburgs were thus compensated for their loss of the Austrian Netherlands, which was by this time joined to the United Provinces as compensation, in turn, for that country's loss of certain overseas possessions to Britain. Tsar Alexander's ambitions in Poland almost wrecked the Congress until a compromise was reached, and Poland, save for Cracow, remained divided between Austria, Prussia, and Russia. Prussia acquired part of Saxony and territory in the Rhineland. The King of Saxony correspondingly suffered a territorial loss, as did the Danish king; Norway was severed from its Danish connection and joined to Sweden in compensation for the latter's loss of Finland to Russia. Britain retained Ceylon and the colony of the Cape of Good Hope and gained Malta and a protectorate over the Ionian Islands—these last two an expression of Britain's heightened interest in the Mediterranean. Ottoman affairs were not dealt with at Vienna, and Russia retained Bessarabia, which it had acquired from the Turks in 1812.

One of the main concerns of the peacemakers was the preservation of the balance of power, a concern that caused Castlereagh and Metternich to enlist Talleyrand's support in opposing Russian encroachment. Indeed, Castlereagh staunchly defended France, which was deprived of its conquests but not otherwise diminished, for France was an essential component of the European equilibrium.

Preservation of the balance of power

This relative absence of a vengeful spirit, perhaps surprising to a later age, may be credited to the fact that rulers rather than the representatives of the peoples contrived the settlement of Vienna. All agreed that the task of ruling was the proper appanage of a small class; yet there were differences. The three eastern powers—Austria, Prussia, and Russia—were absolute autocracies, whereas Britain was a constitutional monarchy. The charter that Louis XVIII granted to his subjects foreshadowed a cleavage between liberal (Britain and France) and conservative (Austria, Prussia, and Russia) Europe. After 1815 "constitution" became the shibboleth of liberals everywhere. Metternich was concerned with the preservation of the social order, and Tsar Alexander invented the Holy Alliance, a somewhat mystical rulers' international, which Britain declined to join. Castlereagh set greater store by the Quadruple Alliance, a more concrete and limited instrument, designed to contain France. Whatever ideological affinity there may have been between the two countries, at the level of power the Anglo-French rivalry was solidly established.

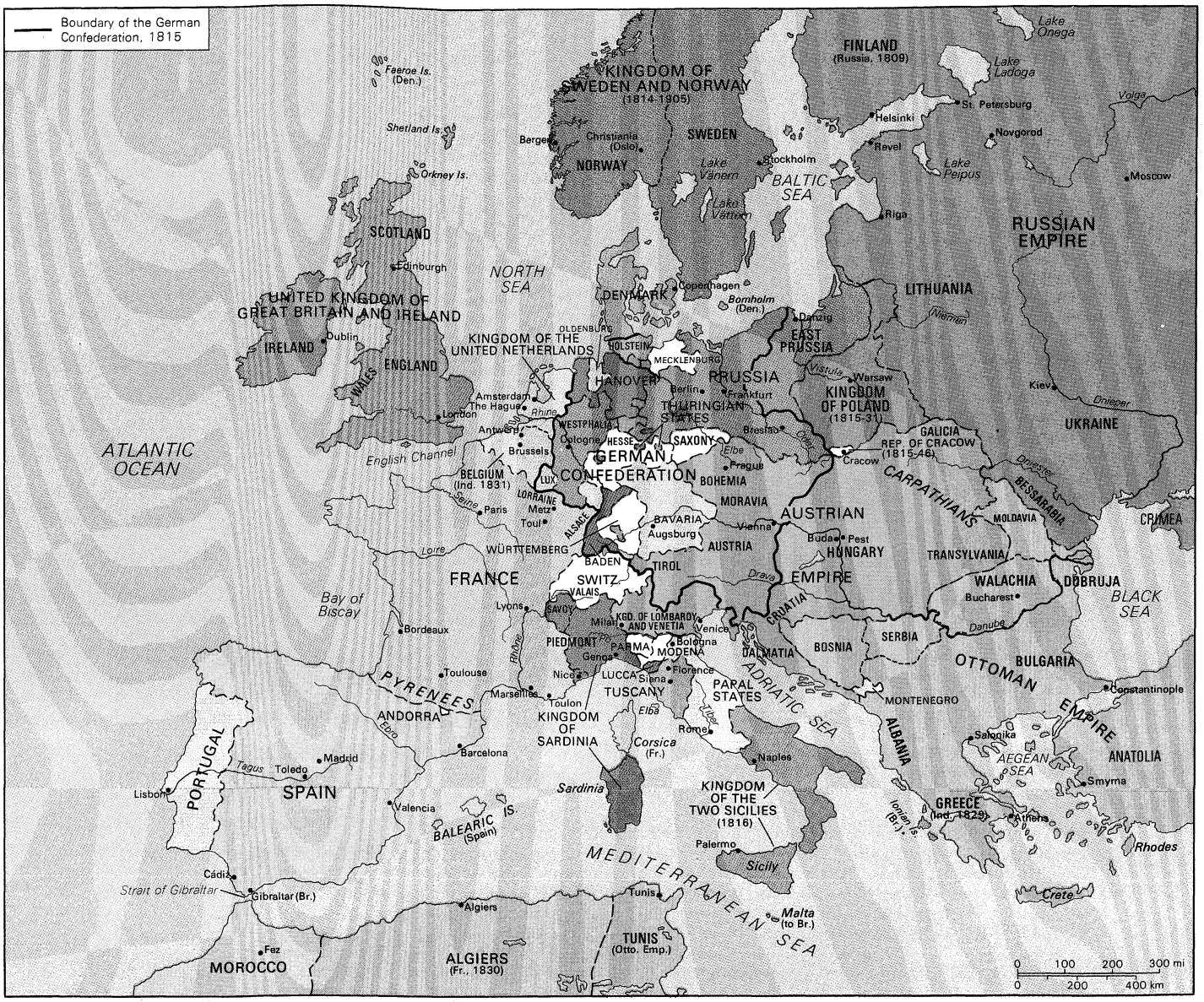
The Restoration period. On the basis of a somewhat ambiguous understanding the European order was maintained for a time. The shock of the Napoleonic Wars was strong inducement for the powers to cooperate to maintain the peace. France gave no evidence of aggressive ambitions and was soon admitted as a member in full standing to the Concert of Europe, an informal organization, yet one not devoid of reality, which for a century had perhaps greater effectiveness than the more formalized 20th-century attempts to organize international order. In one respect the European concert was far simpler: it assumed it to be both the right and the duty of the Great Powers to impose their collective will on those states not admitted to the charmed inner circle.

The peacemakers took the position that a disturbance anywhere in Europe was the legitimate concern of the powers. But on this score a cleavage soon appeared. When there were liberal risings in Italy in 1820, Metternich summoned the powers, which met at Troppau and later at Laibach. From them he received the mandate to restore order in Naples with an Austrian army. But a constitution in Naples little concerned Britain, whose consent to the Austrian intervention was rather due to its view that Italy was an Austrian sphere of influence.

An extremely important issue was involved here—that of the right of outside intervention in the internal affairs of another state. The revolutionary contagion of 18th-century France or 20th-century Russia gives profound reality to the question, which has remained a vital one to this day.

A similar problem arose in connection with Spain, where an army mutiny in 1820 grew into a rebellion, but the cleavage was sharper among the powers when they dealt with that problem at the Congress of Verona in 1822. None of the powers was keen on Russian intervention, which the Tsar would have willingly provided; France was the logical candidate for a European mandate to restore order in Spain. France itself was hesitant, and George Canning, who had replaced Castlereagh as the British foreign minister, opposed such an intervention mainly because it would be French. In the event, a French army marched into Spain, restored King Ferdinand VII's autocracy, and returned home. Europe had enforced its collective will.

Canning did not carry his opposition to French intervention so far as to threaten the use of force, but he did make it clear that Britain might use force to prevent the extension of European assistance to Spain in putting down rebellion in its American possessions. Canning



Europe after the Congress of Vienna.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

The Monroe Doctrine

proposed to the United States a joint Anglo-U.S. policy statement against European intervention in the Americas. But the United States decided to act independently. The result was the Monroe Doctrine, proclaimed in 1823. Although it was a U.S. declaration, it was dependent for some time for its enforcement on British naval power.

Reaction was on the ascendant everywhere in Europe around 1820: the British Parliament passed the notorious Six Acts (1819) against radical activities; Metternich procured the Carlsbad Decrees (September 1819) suppressing student societies and freedom of speech; and in France the reactionary aristocracy inaugurated the White Terror against republicans. The last Bourbon ruler of France, Charles X, a thoroughgoing reactionary, came to the throne in 1824. But, by the end of the decade, at least in France and Britain, the temper was changing. Britain was showing signs of returning to a more liberal practice—civil restrictions against Roman Catholics were removed in 1829—and there was much agitation for parliamentary reform and the broadening of the franchise. Across the Channel the liberals' pent up frustration with Charles X was reaching explosive proportions.

In July 1830 a revolution in Paris drove Charles X from the throne and replaced him with Louis-Philippe, who styled himself "king of the French" rather than king of France. France was approximating ever more Britain's constitutional monarchy, and with the enactment in 1832 of the great Reform Act in Britain, the political ordering of the two countries achieved considerable similarity: the rising class of commercial and financial bourgeoisie had gained substantial power. But in both countries, especially in France, workers felt denied and frustrated.

France still possessed its revolutionary tradition. Liberals everywhere turned to Paris for inspiration, and a rising in France contained contagious possibilities. There were again risings in Italy and some agitation in the Germanic world, but these movements received no assistance from France. Metternich's control remained secure in his domain; the Prussian king was content to follow Metternich's lead in Germany in exchange for Austrian consent to the Prussian-promoted policy of economic integration, the developing Zollverein, or customs union, from which Austria was excluded.

One particular problem arose in connection with the July Revolution that subjected the Concert of Europe to a severe test. Belgium (roughly the former Austrian Netherlands) had since 1815 been a possession of the Dutch king. Belgium, however, possessed a distinct identity that caused it to resent a somewhat less than tactful—and for that reason all the more apparently alien—rule. Here was the meeting ground of the French

and Germanic worlds, a fact reflected in the linguistic division between Flemings and French-speaking Walloons, the former staunchly Catholic, the latter more influenced by French liberal thought. But the two groups finally joined in a demand for greater control over their own affairs.

The rising in Brussels in August 1830 and the intransigent attitude of King William I resulted in a Belgian proclamation of independence in October. A double question arose: would the powers respond to the Dutch king's appeal for assistance in putting down unquestionably rebellious subjects, and would Louis-Philippe aid the rebellion, which enjoyed not a little sympathy in France?

Louis-Philippe was no revolutionary and determined not to intervene in affairs outside the borders of France. He had to overcome British doubts, for the freedom to trade on the Scheldt River in Belgium rated high among British interests. But once Talleyrand persuaded the British of Louis's good intention, the policies of the two countries coincided, for Britain felt little concern for Metternichian legitimacy. Metternich himself and the Prussian king had no stomach for a confrontation with France, and the Tsar was busy with his own Polish rebellion. This meant Belgian independence unless King William could deal with the situation with his own means.

The powers, which met in London to deal with the problem, accepted a de facto situation: they collectively sanctioned Belgian independence, settled the matter of its frontiers, and underwrote the perpetual neutrality of the new country. They also found a ruler for it in the person of Leopold I of Saxe-Coburg, an uncle of the future Queen Victoria. King William bowed to the inevitable and Belgium endowed itself with a liberal constitution.

Belgian independence

Clearly, here was also an instance of nationalistic success—a breach in the existing European order. But it had been effected nevertheless in peace through the agency of the Concert of Europe.

The Eastern Question. The powers meeting in London to deal with affairs in The Netherlands were also involved in the discussion of Ottoman affairs, which had not fallen within the purview of the Congress of Vienna. The Tsar's characterization of Turkey as "the sick man of Europe" was an apt one. The still vast Ottoman domain was beset by general decline and maladministration that resulted in a centrifugal tendency, most marked in its peripheral marches; in its European section, the Balkans, the forces of nationalism were gaining increasing strength. In addition, the Balkan peoples were overwhelmingly Eastern Christians, but the Sultan was caliph of Islām; the theocratic nature of the Ottoman state was an impediment to its modernization. The impotence that went with decadence created a temptation for the Great Powers, who could either maintain the status quo in the Ottoman lands or agree to their partition. Inability to agree on the latter resulted in an attempt to adhere to the former, an attempt attended by a number of crises. The Eastern Question, as it came to be known, remained a perennial of the 19th-century diplomacy of Europe. Among the powers, Russia, Britain, and France all had important interests in the Near East; Austria's were more narrowly confined to the adjoining Balkans; Prussia had none.

The Greek revolt

The combination of Turkish maladministration and of a reviving Greek national consciousness resulted in a revolt and, in 1822, a Greek proclamation of independence. The revolt dragged on inconclusively, a temptation to the Tsar in particular, torn between his devotion to legitimate rule, in this case Turkish, and the possible enhancement of the purely Russian interest.

The 1827 Treaty of London between Russia, Britain, and France was the outcome; it expressed their agreement in exerting pressure in Constantinople to grant concessions and reforms. But things went wrong at this juncture, for the accidental destruction of a Turco-Egyptian fleet by a mainly Anglo-French one in Navarino Bay in 1827 and open war between Russia and Turkey the next year created a new situation. The Russians eventually defeated the Turks, and imposed peace on them, but they were willing to join in the subsequent collective arbitrament of Europe.

In this instance also the powers managed to agree, with the consequence that, as in the case of The Netherlands, an independent Greece was born whose frontiers the powers established and on the personality of whose ruler—Otto I, the second son of the King of Bavaria—they likewise agreed. Again, the European concert had proved its effectiveness, and nationalism registered a success. But, from the standpoint of the latter, the story of Balkan emancipation was only beginning.

It was a measure of Turkish decline that during the Greek revolt the Sultan had enlisted the assistance of his vassal Muḥammad ʿAlī Pasha, an able soldier of fortune who had established himself in control in Egypt. The aftermath of the Greek episode was a falling-out between the Sultan and his vassal. In war Muḥammad ʿAlī was successful, but Russian pressure rescued the Sultan in 1833. When, six years later, the Sultan sought revenge, he was once more defeated by the Egyptians. But matters this time took a different turn. Britain was now the chief supporter of the Sultan, while France was playing the Egyptian card. It came close to a dangerous confrontation, but once Louis-Philippe had replaced the belligerent Adolphe Thiers with the more accommodating François Guizot as foreign minister, agreement among the powers was restored. Muḥammad ʿAlī had to remain content with the rule of Egypt alone.

The Anglo-French confrontation inserts itself as an episode in the long record of imperial rivalry between the two countries. But in general that activity was at a relatively low ebb during this period. The French did, hesitantly at first, lay the bases of their North African establishment, and the British, with the French and others in tow, began to open up China, in whose trade the United States was also interested. The British and the French were also active elsewhere in Asia (Malaya and Aden) and in the Pacific (New Zealand and Tahiti), and both kept an interested eye on Mexican and Texan affairs.

The events of 1848–49. Despite their rivalry, there was a measure of agreement between the two powers, and in mid-Europe Metternich had things under control. But liberals everywhere were also active; once more, in 1848, an explosion occurred, and again its Parisian phase was the signal for others. In France the Second Republic was established, but for Europe as a whole the main question was whether the new French regime would extend assistance to revolution elsewhere.

Revolution in Paris

The news of the Parisian uprising was the spark that set off liberal demands in mid-Europe. In Vienna and Berlin and in the small German states the cry was for constitution, but in Prague, Budapest, and Milan the nationalistic force equally asserted itself. Metternich's system everywhere was in collapse; he himself fled into exile, while the frightened rulers, taken by suprise, granted their subjects constitutions.

On two occasions, in 1848 and again in 1849, there was war between the Kingdom of Sardinia and Austria, the former being defeated both times. Prague was subdued and so likewise was Budapest, though it took a little more time—and Russian assistance—to effect the latter result. The army had remained loyal, and under the stern but able leadership of Felix, Furst zu Schwarzenberg, Habsburg rule was everywhere re-established; the momentarily frightened Prussian king took back his concessions, while the representatives of the Germanic world, assembled in Frankfurt am Main, gave proof of the ineffectiveness of mere high-handed talk in the absence of concrete power. But, mainly as a consequence of France's abstention, the disturbances did not spread into a major international conflict.

In France the February revolution was followed by another in June, in which the forces of order reasserted their control. The majority of the French people refused to follow the lead of the more radical Parisian workers. By the end of the year, the election to the presidency brought to that office Louis-Napoleon, a nephew of the first Napoleon, who exploited with skill the widespread

desire for order and the by-now legendary associations that were connected with the name. But Louis-Napoleon continued the French policy of nonintervention, and reaction was unimpeded in reasserting its control in mid-Europe in 1849. By the end of that year the revolutionary urge was spent and had apparently accomplished little.

What international complications there had been were limited and did not threaten the peace among the powers: a brief Prussian incursion in Denmark, a Russian one in Hungary in assistance to the Habsburg ruler, and a French force in Rome protecting the Pope were the sum total of these interventions. Yet there was also warning for the future. The impotent Frankfurt National Assembly in 1849 offered the German crown to the Prussian king; in fear of a sufficiently restored Austrian power, he declined, and, when a year later he attempted a scheme of his own to somewhat similar effect, he withdrew in the face of an Austrian ultimatum. Nevertheless, the issue of German unity and of the bases of it, whether Grossdeutschland or Kleindeutschland (*i.e.*, including or excluding Austria) had been seriously agitated. The events of 1848–49 may be summed up as a challenge to the existing order, an unsuccessful one at the time but a challenge that cast a shadow on the future.

The Frankfurt National Assembly

REVISION OF THE SETTLEMENT OF VIENNA

More than three decades after the Vienna settlement, Europe was still organized essentially as it had been in 1815; and the Great Powers had proven themselves capable of acting in accord. France, to be sure, remained a question mark, especially after 1852, when Louis-Napoleon ended the Second Republic by a coup d'etat and became Emperor Napoleon III. The Emperor's declaration that "the empire means peace" was only partially convincing to the Great Powers; in fact, Napoleon's 18-year reign was to witness an extensive revision of the settlement of 1815, a process in which France played an important part, though not always a directing hand.

The Eastern Question again: the Crimean War. The first major shock to the Concert of Europe after 1848 was the Crimean War. The war grew out of a minor squabble over the respective rights of Eastern Christians and Roman Catholics in Jerusalem, which was under Ottoman control. Napoleon III, for reasons of prestige, undertook the vigorous defense of Roman Catholic interests in the Near East. The effect, however, was to provoke a Russian reaction, ostensibly in defense of the Eastern Christians but also an attempt to gain further concessions from the Turks. Thus, the quarrel of the monks escalated into a larger competition. The Turks adopted their familiar tactic of encouraging differences among the powers, which they did with total success.

The Franco-Russian difference soon turned into an Anglo-Russian one, for the British feared Russian expansion into the Mediterranean. The crisis was slow in unfolding, to the accompaniment of poor diplomacy, but eventually a Turkish declaration of war against Russia at the end of 1853 was followed by British and French declarations early in 1854.

Both sides sought support from Austria, which, however, remained neutral. Austria earned thereby the discontent of all, but Vienna remained the centre of diplomatic activity throughout the conflict. The Austrians occupied the Danubian Principalities in the Balkans, thus preventing land contact between the belligerents, save in the Caucasus. Making use of their naval power, the western states carried the war to the Russians in the Crimea.

The war itself has been described not unfairly as a contest in incompetence; but in the end, the more technically advanced western powers were able to defeat the Russians on their home ground. The advent of a new tsar, Alexander II, in 1855 finally made accommodation possible, and the belligerents met in Paris in 1856.

The Congress of Paris

The Congress of Paris is an important landmark in the history of the Eastern Question. The Russians did not pay an exorbitant price for their defeat, as the war had been one of limited aims. The congress took place in a friendly atmosphere of festivities; the Concert of Europe had momentarily broken down, but the fundamental consensus of the powers persisted.

Europe, collectively, replaced Russia in the assertion of supervisory rights in the principalities, and a small loss of Bessarabian territory deprived Russia of access to the Danube, but otherwise the status quo ante-bellum was restored. The Black Sea was demilitarized, a clause that rankled in Russia. Also, the Ottoman Empire was admitted to membership in the Concert of Europe, though, curiously in a sense, simultaneously placed under what was tantamount to a collective European protectorate.

The powers gathered in Paris used the occasion to deal with some other matters, such as diplomatic procedure and international law, especially its maritime aspects, including the perennial concern over the rights of belligerents and neutrals at sea in wartime. One question, however, that received little more than passing attention was that of conditions in Italy, where the quality of administration in the Kingdom of the Two Sicilies and in the papal domain was not much better than in the domain of the Turkish sultan.

The making of Italy. There were two reasons for raising the Italian question at the Congress of Paris: the interest of the French emperor in Italian affairs and the presence of Count Cavour, who represented the Kingdom of Sardinia. Sardinia was not a great power but had sent a contingent of troops in support of the allies in the Crimean War.

Cavour is properly credited with the making of a united Italy; he was responsible for Piedmontese participation in the Crimean War, precisely in order to bring the Italian problem to the attention of Europe. His reward in Paris was small, but the interest of Napoleon III was not; and the consequences of it were soon to be apparent.

Since the days of ancient Rome, Italy had been little more than a geographical expression—more recently a battleground of rival influences, Spanish, French, and Germanic. But the consciousness of a common culture, especially among the literate and the educated, persisted. The French Revolution and its Napoleonic sequel had introduced nationalism as an active political force in the Italian world. The settlement of Vienna, making Austria the paramount power in the peninsula, inevitably gave Italian nationalism an anti-Austrian point.

Nationalism was to a considerable degree associated with the liberal agitation for constitutional reform, one of the chief spokesmen for which was the Italian patriot Giuseppe Mazzini. Mazzini hoped to make Italy a unitary republican state by doing away with all the existing ruling houses. But his was an extreme solution; a more moderate approach was that of federation, which had the advantage of leaving unimpaired the temporal power of the Pope, who might be called upon to head a federal Italy.

Other nationalists looked for leadership to Piedmont rather than to the papacy, for the popes were often reactionary. Piedmont in northern Italy was united with Sardinia under the House of Savoy as the Kingdom of Sardinia; the country was the most advanced independent Italian state.

Austria had successfully suppressed Italian risings in 1820 and 1830, and attempted Piedmontese leadership in 1848 fared little better; but the events of 1848–49 did have a clarifying effect; the republican and the federalist tendencies—Mazzini and the Pope—both lost their credibility. Piedmont, though defeated in war, had stood up to Austria and even managed to retain its new constitution, the Statuto of 1848.

Cavour's policy

Also, the 1848 disturbance brought Cavour to power in Piedmont. Essentially a practical man, Cavour had little use for romantic revolution or Mazzinian plotting. Once in charge of Piedmontese affairs, he proceeded on the realistic assumption that the only way to evict the Austrians was with the help of a foreign force—which France alone could supply.

In July 1858, Napoleon III and Cavour met at Plombières and worked out a secret agreement whereby France would gain territorial concessions in Italy in return for French military help in throwing off Austrian

rule. Cavour was forced to a compromise, for the well-meaning but irresolute French emperor was, after all, primarily in charge of the interest of France, from whose standpoint a divided Italy was preferable to a united one. Also, Catholics were powerful in France.

In April 1859 Cavour provoked a war with Austria, which found itself confronted with France as well. The brief war was successful for the allies; but Napoleon, concerned over Prussia's mobilization, put a sudden end to it at a meeting between himself and the emperor Francis Joseph I at Villafranca on July 11. Cavour, feeling betrayed, resigned in anger.

But it was too late for the limited stabilization that Napoleon and the Emperor tried to devise at Villafranca. The war precipitated nationalist revolts throughout Italy. Venice remained in Austrian hands, but the rest of the Italian structure toppled. Risings and plebiscites in the central states, and Giuseppe Garibaldi's colourful and successful expedition against the Kingdom of Naples, essentially brought Italy together. In March 1861 the Kingdom of Italy was proclaimed, which became a centralized, unitary state patterned on the French model. Sardinian constitutional arrangements were extended to the whole under the Savoy dynasty. The pope, however, continued to rule in a much reduced domain under French protection. Within another decade Venice and Rome were incorporated into the main body of Italy as footnotes to larger events.

The appearance of united Italy on the map of Europe was unquestionably a success of the nationalistic force, the undoing of which was hardly possible, even though most Italians had been merely passively acquiescent. France had acquired Nice and Savoy on the plea of ethnic affinity, but the manner in which the war with Austria had been settled and the Roman issue dealt with added little to Napoleon's credit in either Italy or France. As a new major power, even if the weakest, a united Italy was welcomed by Britain as an additional potential counterweight to France.

The making of Germany. The success of Italian nationalism had been followed with special interest in Germany. There were both similarities and differences between the two situations. The most fundamental fact was the existence of a German people. A claim to universality comparable to the Roman and expressed in the institution of the Holy Roman Empire had long been voided of content, and, as in Italy, there was a long record of particularistic divisions and loyalties. Two rivals contended for possible leadership in the Germanic world—the Habsburgs and the Hohenzollerns. The very nature of the Habsburg state, with its agglomeration of diverse nationalities, could not but make it antagonistic to national aspirations, which, conversely, Prussia could more easily espouse. A new component—nationalism—was thus introduced in the 19th century into what had been hitherto a dynastic contest. In Germany, as in Italy, the impact of French influence, revolutionary and Napoleonic, had given a fillip to national aspirations.

For nearly 30 years Metternich had successfully presided over the Germanic world, and his departure from power in 1848 was soon followed by a seeming return to the prerevolutionary order. The events of 1848–49, not least those in Frankfurt am Main, were a liberal failure, even though a Prussian constitution was promulgated in 1850.

The whole Germanic world remained uninvolved in the Crimean War. Order prevailed in it during the 1850s, but neither the Habsburg–Hohenzollern rivalry nor liberal hopes had been lastingly eliminated. It was a domestic Prussian problem, the issue of army credits, which developed into a dispute between the crown and the Landtag, that finally in 1862 brought Otto von Bismarck to the minister-presidency of Prussia. One is reminded of Cavour after 1848, and there are points of similarity between the two men, their diplomatic skill above all. But there is also contrast. Whereas Cavour believed in liberalism, Bismarck was a typical representative of the Prussian Junkers, the landowning aristocracy that formed the conservative backbone of the Prussian state.

The rise of Bismarck

Contemptuous of the idle speeches and resolutions of the ineffectual revolutionaries of 1848, Bismarck, believing in "blood and iron" methods, solved the constitutional issue through the simple device of ignoring the constitution. His first great accomplishment was to settle clearly the Austro-Prussian competition in Germany through the radical solution of ousting Austria from the Germanic world outside the Habsburg domain proper; Bismarck was a Kleindeutscher.

The operation was conducted with masterful skill. The method was to isolate the issue from larger international complications, then to deal with Austria through the test of military force. Bismarck secured Russian neutrality by cooperating with Tsar Alexander II in putting down a Polish rising in 1863. The grateful tsar, already incensed by Austria's neutrality in the Crimean War, shifted his support from Austria to Prussia.

Danish affairs offered Bismarck his next opportunity. Posing as the defender of oppressed German nationality in the Danish duchies of Schleswig and Holstein, Bismarck persuaded Austria to join him in war against Denmark in 1864. The duchies were put under joint Austro-Prussian control, an arrangement highly suitable for later bickering. The British and the French could not agree on joint action and left Denmark to its fate.

Having beguiled Napoleon III into neutrality through hints of possible compensation, Bismarck then proceeded to pick a quarrel with Austria over Schleswig–Holstein, and the Austrians unwisely went to war. The Austro-Prussian War of 1866 lasted only seven weeks, the Prussians defeating the Austrians at Sadowa (Königgrätz) in July. Austria was evicted from Kleindeutschland, which was reorganized in 1867 under the Prussian-dominated North German Confederation, while the South German states were left undisturbed. Austria was not otherwise punished, but its loss in the war pointed up the need for internal reform. In 1867 the Habsburg domain was reorganized as the Dual Monarchy of Austria-Hungary, under which the Habsburg lands in Hungary achieved constitutional parity with Austria—another success of the nationalistic force. Italy, which had joined in the war, though she fared poorly in it, acquired Venetia as a result of the Austrian defeat.

The Seven Weeks' War

Events had moved too fast for Napoleon III to intervene, but the ambition of Bismarck was clear. Bismarck did not think that France would stand idly by while German unification under Prussia was completed. Realizing that another war would be necessary, Bismarck prepared for it in the same fashion that he had for the showdown with Austria: the conflict must be isolated. The Tsar was still sympathetic; England was on the whole indifferent; and the reorganized Austria-Hungary was increasingly Balkan oriented.

This time it was Spain that provided Bismarck with his opportunity. A revolution in that country resulted in the search for a new ruler, in the course of which the candidacy of a member of the Hohenzollern house was advanced. France, understandably, objected to the possibility, and Bismarck brilliantly manipulated the controversy to provoke a military showdown. Napoleon III, himself not desirous of war but by now a sick and an ever vacillating man, fell victim to the trend of circumstances. France declared war in July 1870 and was ostensibly the aggressor, but Bismarck had led the game all the way.

The war remained confined to France and Prussia (or rather Germany, since the south German states joined Prussia). The outcome was shortly apparent. A large French army, with the Emperor himself, surrendered at Sedan in September. The succeeding French provisional government could only prolong the futile contest for a time. Paris was besieged and surrendered in January, and the French had no alternative but to accept the harsh terms of the victor: a large indemnity and the surrender to Germany of Alsace and a part of Lorraine.

The Franco-Prussian war

Peace was restored by the Treaty of Frankfurt in May 1871, while the new French government established in February was harshly putting down the episode of the Paris Commune. Thus, in hesitant fashion, the Third French Republic was born.

Meanwhile, in January at Versailles, the German Empire had been proclaimed, which, having successfully passed the test of force, took France's place as the first power on the Continent.

There were two footnotes to the Franco-Prussian War. In September 1870 the Italians entered Rome, henceforth their capital. (The Pope declared himself a prisoner in the Vatican, refusing to acknowledge the *fait accompli.*) The Russians also took advantage of the war to denounce the demilitarization of the Black Sea imposed upon them in 1856.

The Franco-Prussian War was the major international event of the period 1815–1914, which it divides into two nearly equal parts. But some other events had occurred since the Crimean War.

The Eastern Question was relatively quiescent during this period. What complications arose in connection with it were settled by the powers in agreement. Between 1858 and 1864 the Principalities (Moldavia and Walachia) achieved their union and their virtual independence as the country of Romania (then Rumania).

The French interest in Egypt persisted and was enhanced by the successful construction of the Suez Canal by a French company, despite some English obstructionist attempts. The canal, inaugurated in 1869, soon proved to be a great commercial and financial success.

In 1863 the Poles revolted against Russia. They were subdued, though not without difficulty, and Poland became, in effect, a Russian province. The Polish hopes of external assistance were frustrated, and, as a consequence of the episode, a Russo-Prussian understanding replaced the previously existing Franco-Russian (see above).

Events outside Europe. Outside of Europe some developments of considerable significance, especially for the longer term, also took place. The long self-imposed isolation of Japan came to an end when, in 1854, the American commodore Perry succeeded in negotiating the Treaty of Kanagawa. After a period of hesitation, in 1868 Japan deliberately embarked upon the path of Western imitation as the only manner in which it could retain its independence.

The opening of Japan

In China, a process of decline and disintegration continued reminiscent of that of the Ottoman Empire. The Taiping Rebellion eventually led to foreign intervention, mainly Anglo-French, the forcible opening of China, and the imposition upon that country of discriminatory treaties. The Russians were expanding in Siberia, securing the boundary of the Amur River. Vladivostok, on the Pacific, was founded in 1860.

In Southeast Asia the British influence was reaching out from India, reorganized after the great Indian Mutiny of 1857, toward Burma and Siam, while the French established themselves in Cochin China and Cambodia, thinking of the Mekong River as a possible southern route of access to China.

The American continent was supposedly out of bounds for European intervention, but the British and French maintained interests in Mexico and in Texas; Britain's maritime and commercial empire made the possibility of an isthmian canal of equal interest to Britain and to the United States. Manifest destiny meant territorial expansion for the latter, mainly at the expense of Mexico, but domestic issues exploded into civil war between the North and the South in 1861. Aware of the potential rivalry of the United States, the British and the French, at the governmental level at least, tended to be sympathetic to the fragmentation of U.S. power and toyed with recognition of the South. The war gave rise to some incidents—the "Trent" affair and the "Alabama" claims—but the ultimate success of the North preserved the unity of the country while laying the bases for its subsequent astounding growth. The emergence of the United States, Germany, and Japan within less than a decade is worth noting.

The French in Mexico

The U.S. Civil War facilitated French intervention in Mexico, and Napoleon III succeeded in placing his candidate, Archduke Maximilian of Austria, on the Mexican throne in 1863—an action highly unwelcome in the United States. The termination of the Civil War and the menace of a united Germany induced Napoleon III to recall the French force from Mexico, and, lacking local support, luckless Maximilian was captured and executed. The Mexican episode contributed to the impression of the French emperor as a meddler, a factor that, in turn, contributed to British and American indifference during the Franco-Prussian War.

THE IMPACT OF ECONOMICS ON POLITICS

During the 19th century economic change had an overwhelming impact on international affairs. At the opening of the century, Britain was the foremost commercial state in the world, and British wealth was the mainstay of the coalitions against Napoleon. Britain was the first nation to undergo the Industrial Revolution, and its early lead in this field, combined with its far-flung empire, ensured its primacy of power in Europe and the world for the rest of the century. Britain was also dedicated to the policy of free trade, adopted in the 1840s.

But the Industrial Revolution did not remain confined to Britain. It appeared in Belgium and in northern France, then gradually spread eastward. By 1870 Germany had equalled France in both population and industrial development, but between 1870 and 1914 Germany rapidly outstripped France in both respects.

The early 19th century was the age of coal and steam, perhaps best symbolized by the railway. The main lines of the network of railways in Britain, France, and Germany were laid during the three decades preceding 1870. It is said that the Franco-Prussian War was decided by the Germans' superior use of the railroad. Railroads required large amounts of capital; the business of dealing with money as a commodity—banking—assumed correspondingly large proportions; the financing of the construction of the Suez Canal is another illustration of its role.

These economic changes profoundly altered the structure of society. The machine, increasingly located in the factory, needed tenders. An essentially new class arose, a largely propertyless proletariat, which tended to congregate in cities. Under the system of free enterprise, labour was looked upon as a commodity. The labourer himself was dependent upon the fluctuating cycles of the economy, and his lot was usually miserable.

The workers had been important contributors to the Parisian revolutions of 1830 and 1848. At that time, however, they were neither sufficiently numerous nor adequately organized, and bourgeois capitalists were the beneficiaries of whatever changes occurred. The growth of proletarian numbers, hence of their possibilities of action, was powerfully aided by the subsequent course of economic development.

The "second Industrial Revolution"

After 1870 one may speak of a "second Industrial Revolution"—that is, a great expansion in such traditional industries as coal and steel and the appearance of new ones, such as the chemical and electrical industries. All the while, industry was spreading geographically, until one could speak of an "inner" Europe, the area enclosed by a line running from Glasgow through Stockholm, Stettin, Trieste, and Barcelona—the Europe of industry. "Outer" Europe was predominantly agricultural, but the "inner" was steadily encroaching on the "outer."

The loss of the American colonies had a profound impact on British economic thinking. Why seek to impose political control on others when trade was what mattered? Free trade was more important than colonial conquest. Imperial activity was at a relatively low ebb during the first half of the 19th century. But, as others turned to a policy of protection and as the European market showed signs of saturation, imperial activity revived. Economic motivation is not the exclusive, but certainly is the most important, driving force of the "new imperialism," which registered notable achievements during the closing decades of the period (see below). In the 30 years after 1880, virtually the whole of Africa was divided among the European powers, and the sharpening competition for colonies became a major concern of the chancelleries of Europe and a decisive factor in shaping international alignments.

The uneven growth of industry affected decisively the power standing of states. After 1870 France declined from second place to the bottom of the list of advanced industrial nations. Britain retained its primacy, but the margin of its advantage was diminishing, challenged in the main from three quarters—the United States, Japan, and Germany. But the impact of the first two is essentially a 20th-century story; the German, on the other hand, was quite clear, sharpened before 1914 by the German policy of naval construction; thus Germany, after unification, lay at the centre of international politics.

The European record over a 100 years seemed to belie the gloomy predictions of overpopulation made by Thomas Malthus at the beginning of the century. Although some 50,000,000 Europeans swarmed to other continents, the population of Europe roughly trebled. But the condition of the masses generally improved. The fundamental reason for the change, allowing for political agitation and some humanitarian pressure, lay in the increase in the production of wealth, industrialization most of all. The 19th-century experience of Europe on this score constitutes a perhaps exceptionally fortunate accident, for industrial Europe could draw on the resources of the outer world for food and raw materials.

This growing wealth had other repercussions besides that of raising the standard of living. The combined impact of the rights of man and of scientific and technological development was favourable to the rise of the masses, one aspect of which was the spread of literacy, education in general, and a cheap press. The business of government became less and less the appanage of a restricted privileged class, though the conduct of foreign affairs continued to be a largely esoteric preserve. Yet, at election time and as the franchise spread, the temptation could not be resisted to appeal to the masses. The obverse facet of the democratic coin is demagoguery. How to conduct foreign policy in a democratic milieu has been an increasingly delicate and complicated problem. In the age of nationalism, the simplest way to arouse the masses was to appeal to their national prejudices and emotions. The enthusiasm with which the peoples of Europe responded to the outbreak of World War I in August 1914 bears out the point.

Outside of the United States and Japan, Europe was the mistress and the powerhouse of the planet. The state of European culture, both in the scientific domain and in the letters and the arts, was added grist to the mill of those who confidently believed that Europe *was* civilization, even if this civilization was an increasingly materialistic, godless one. As to the proliferation of arms, it was used as an argument for confidence in the lastingness of peace, ensured by what is now called the balance of terror, although the phrase itself is of later coinage.

There were Cassandras to be sure, and in the end they were, at least in a sense, to be proved right; but, at the time, they were mainly voices crying in the wilderness. When they went to war in 1914, the peoples and the governments of Europe had little inkling that they were initiating the closing chapter of the European Age.

THE BISMARCKIAN PERIOD, 1870–90

Bismarck's system of alliances. Bismarck was a practical man rather than one dedicated to abstract principle. His social bent was undoubtedly conservative, but he understood the active forces of the modern world. Having achieved his purpose of Prussian enhancement in the form of uniting Germany under the Hohenzollern crown, he was now satisfied and entertained no further aggressive designs. He believed that the maintenance of peace and the status quo would provide the best climate in which the new Germany could thrive.

Bismarck's peace policy

During the 20 years that he remained in charge, his skillful diplomacy dominated the international scene.

Germany had no enemies save France, which had been humiliated in 1870 and stripped of Alsace-Lorraine. Bismarck had no fear of France if France could be kept in isolation. In France he preferred the republican government and even favoured colonial ambitions, which he hoped would divert French attention away from the Rhine and revanche while possibly embroiling France with others. As for the British, Bismarck was content to leave to them their monopoly of empire.

His relations with Russia were good, and in the new hyphenated Austria-Hungary the Hungarian half was satisfied with the *Ausgleich* (Compromise of 1867) and would emphasize a compensating interest toward the southeast—the Balkans. This would, however, give added point to possible Austro-Russian rivalry. Bismarck's own solution for this problem was simple: divide the Balkan apple of discord between the two contenders—he would be Austrian in Serbia and Russian in Bulgaria. Moreover, the three great conservative states of Europe shared an outlook reminiscent of the Metternichian.

These thoughts bore fruit in 1873 in the Dreikaiserbund (League of the Three Emperors). But agreement was not long lasting, for by 1876 local Balkan events resulted in a falling out between Austria and Russia. Bismarck's control had limits. If he must choose between the two, though he would rather not, he would opt for the wholly European and partly Germanic Dual Monarchy. The Austro-German Alliance was concluded in 1879, henceforth the cornerstone of German foreign policy. In Bismarck's eyes it was undoubtedly a defensive alliance, designed as much to reassure as to restrain Austria.

Within two years Bismarck was able to revive the tripartite connection, but the second Dreikaiserbund foundered, like the first, over local Balkan differences—this time over a war between Bulgaria and Serbia in which Serbia was rescued by Austrian diplomatic intervention.

This drove Bismarck into an extremely complicated scheme. His alliance with Austria was a fixed point, and he even informed the Russians of its terms. But in order to insure against the nightmare of the war on two fronts —a Franco-Russian coalition—he contrived an understanding between himself and Russia, the Reinsurance Treaty of 1887. This treaty has been described by some as Bismarck's greatest diplomatic achievement, by others as an attempt at political bigamy. It was a tightrope-walking act and lasted only as long as Bismarck himself was in power. The Austro-Serbian Alliance and the connection with Romania rounded out the Bismarckian network of eastern alliances.

The Reinsurance Treaty

Italy had a low power rating in Bismarck's eyes, but, in order to secure Austria's rear and to prevent a possible Franco-Italian connection, he enlisted Italy as well in his camp. The Triple Alliance of 1882, made possible in part by Italian fears of French designs, implied a stabilization of the still-existing Austro-Italian antagonism (a legacy of the Risorgimento) and a muting of the Italian irredentist claim against Austria. Italy also fitted into Bismarck's desire for general stability in the south.

Balkan problems. Ottoman affairs, particularly in the Balkans, which were still in the process of achieving national emancipation, continued to be a sensitive and unstable area. In the mid-1870s, revolts flared up in Bosnia and in Bulgaria, out of which ensued a Russo-Turkish war in 1877–78. The victorious Russians imposed the Treaty of San Stefano on the Turks. The treaty called for a drastic rearrangement of the Balkans, whereby a large Bulgaria would be created under the protection of Russia. But the Russian gains in the treaty alarmed the powers. In the face of determined British opposition, the Russians therefore agreed to a collective European revision of the treaty. The powers met at the Congress of Berlin in 1878 and produced a new Balkan arrangement: a much smaller Bulgaria was created and a series of compensations arranged for some of the powers, both large and small. The Balkan provinces of Bosnia and Hercegovina were put under Austrian occupation and administration, and Cyprus was "temporarily" occupied by the British. The Treaty of Berlin was a source of dissatisfaction to the Russians, who, after all, had won the war, and the Dreikaiserbund was a victim.

Seven years later, the Balkans flared up again, though the fighting, Serbo-Bulgarian, remained confined to their locale, and Bulgaria, joined by Eastern Rumelia, achieved full independence from the Sultan. As mentioned above, these episodes dissolved the second Dreikaiserbund.

THE NEW IMPERIALISM

Events in the more peripheral regions of the Ottoman Empire are perhaps better placed within the purview of colonial activity. At Berlin in 1878 it had been suggested to the French, by both the Germans and the British, that Tunisia might be suitable compensation for them. Italy was also interested in Tunis, where the reality of an Ottoman connection was cloudy.

In France there were those who insisted that revanche and the German danger should be the first cares of French policy, while a colonial party thought this a sterile attitude that merely prevented France from taking part in Europe's growing interest in the wider world. In 1881 the latter group scored a success with the establishment of a protectorate over Tunisia. Italy was annoyed and joined the Triple Alliance, but there were no repercussions in Constantinople.

Anglo-French involvement in Egypt

In the 1870s both the French and British became deeply involved in Egypt. The French interest was of long standing, and the Suez Canal was mainly used by British shipping. The Khedive's finances were mismanaged and the British and the French intervened to impose a joint Anglo-French supervision of Egyptian finances. Local Egyptian conditions led to further Anglo-French interference, and in 1882 the British took military action, the French having declined at the last moment to join them. This was the beginning of the effective British control of Egypt, for the supposedly temporary occupation of the country proved to be but the initiation of a continuing and deepening involvement.

The cases of Tunis and Egypt may be conveniently regarded as the launching of the new imperialism, a vigorous recrudescence of Europe's imperial activity. The longer established British and French priority in the imperial domain was being emulated by others. Even initially anticolonial Bismarck yielded to the pressure generated by the rapidity of German economic growth; by the mid-1880s Germany claimed various areas in Africa, and even Italy was developing an interest around and in Abyssinia.

In modern garb the age of discovery and conquest had returned, and the partition of Africa was under way. With a view to avoiding conflict, acknowledging the equal right of all, the powers met in Berlin in an attempt to put some order in the process. Actually, the Berlin West Africa Conference (1884–85), the result of a Franco-German initiative, affirmed the freedom of navigation and trade in the basins of the Congo and Niger rivers. The Free State of the Congo, the outcome of a private initiative headed by King Leopold II of the Belgians, achieved international recognition, and general rules for the establishment of colonial claims were agreed upon.

Penetration of Asia

There were manifestations of imperial activity (primarily British, Russian, and French) in Asia as well. The penetration of China was continuing, as was Russian expansion in Central Asia. This last, pushing toward the western and northern back doors of India, was a source of concern to the British, who found themselves involved in protective, "defensive" expansion. The two influences met in Afghanistan. Likewise, on the opposite side of India, a British protectorate over Burma was proclaimed in 1885, while the French became involved in war with China as a result of their activity in Tonkin.

This colonial activity was largely conducted by a handful of men with small bodies of troops. The peoples of Europe to a large extent were unaware of this aspect of the activity of foreign and colonial offices and had in general no consciousness of being involved in war. While there were also critics of imperialism among Europeans, theirs was but a small voice.

The conscious leaders of Europe's imperial urge nurtured ambitious schemes. Granting the priority of the economic interest and the rationalization of the white man's burden, or *mission civilisatrice*, the factor of national interest, of prestige and pride, should not be minimized. Control of Egypt meant control of the vital artery of the Nile. It was not long before the British were involved in The Sudan, where they suffered a temporary setback. They were also active at the southern tip of the African continent, where Cecil Rhodes, businessman and statesman of ruthless vision, was at work. From these two points as bases, Cairo and the Cape, plus British East Africa, the vision of a vast empire, roughly the eastern half of Africa, symbolized by the Cape-to-Cairo railway project, was born.

The French could entertain a comparable dream, an extension of France across the Mediterranean all the way to the Congo—the northern half of the continent. Around the sources of the Nile, British, French, German, and Belgian influences met, to which the Italian may be added. Here was incipient rivalry and a potential source of conflicts, but these did not flare up until the 1890s, when exploring activity had laid sufficient bases for overlapping political claims.

Meanwhile, the focus of international relations, over which Bismarck presided, was still predominantly in Europe. But already the seeds of disintegration of the Bismarckian system were planted. The combination of the Triple Alliance, his own treaty with Russia, and the Mediterranean agreements constituted a fragile edifice, undermined by inner inconsistencies. Russia and France were the two revisionist states that Britain was willing, to a degree at least, to help contain. But difficulties, of an economic nature at first and not necessarily irretrievable, began to appear in the Russo-German relationship. Bismarck's dismissal in March 1890 opened a new chapter in the relations among the European powers.

THE GERMAN CHALLENGE

The period of transition, 1890–1904. It is a proper measure of the authoritarian nature of the German state that Bismarck could be dismissed from office by the emperor; no constitutional issue was at stake. Bismarck's own place in the roster of statesmen is secure; if personally not a very attractive character, his rank nevertheless belongs among the first: like a good craftsman, he used the proper tool for the right job at the right time; war having served his purpose, he endeavoured to maintain the peace, all the while directing the energies released by the very fact of his unifying success.

Assessment of Bismarck

One reason for his dismissal was the advent, in 1888, of a new emperor, William II. Friction developed between the crusty old man and his young, arrogant, touchy, and romantic new master. The last Hohenzollern proved to be, in a sense, the most adequate and, for that very reason, the most unfortunate embodiment of the insecure, arrogant success that was Wilhelmine Germany, all too effectively conveying to others an impression of aggressive intent out of proportion to the reality of that intent.

A major foreign-policy decision faced the new German administration—whether or not to renew the Reinsurance Treaty with Russia. There were solid arguments pro and con, and the Tsar was desirous of renewal; but the German answer was finally refusal. This decision compounded the uneasiness that had already begun to cloud Russo-German relations before 1890, one consequence of which was the fact that Paris took the place of Berlin in the financing of Russia.

France had been anxious to escape from the isolation in which Bismarck had successfully kept it confined. But France and Russia had few common interests. Russia felt as little concern for French revanche as France did for Russia's Central Asian expansion, with its consequence of Anglo-Russian friction. The focus of France's rivalry was also Britain, to be sure, but it lay mainly in Africa, where Russia had no interest. The only possible binder of a Franco-Russian connection was common fear of Germany, which in the Russian case could arise from the Austro-German alliance.

The Franco-Russian connection began with a series of loans, launched on the Paris market, from 1888 on. A Russian purchase of French arms in 1889 was the next step, followed by military conversations, the elaboration of a contingency plan. The Tsar felt little attraction for what he considered a godless republic, but, eventually, in 1893, he underwrote the results of the military discussions, while the French, on their side, evaded their own constitutional issue through the contention that no formal

Europe, 1871–1914.

Franco-Russian alliance

political alliance was involved, thereby bypassing parliament. Yet from 1893 it is proper to speak of a Franco-Russian alliance as existing. This was the first major breach in the Bismarckian system, made possible in part by the confident but incorrect German assumption that such a possibility was excluded. Like all such instruments at this time it was defensive, and the terms of it were secret.

But the situation was still very fluid, and the focus of European rivalries was becoming increasingly extra-European. The Russians were pleased when, in 1893, friction developed between France and Britain in Southeast Asia. The difference was composed, however, and Siam, as a buffer, retained its independence.

By this time the Japanese had made sufficient progress in their imitation of European ways to have developed imperial ambitions of their own. They turned to China, with which they dealt successfully in war, imposing upon it their demands in the Treaty of Shimonoseki in 1895. This Japanese success was of concern to Russia, and an unexpected combination ensued, for Germany—this was its way to weaken or undo the newly formed Franco-Russian connection—gave Russia support in opposing Japan, while France, rather embarrassedly, followed suit. The tripartite combination, Britain excluded, was successful in frustrating Japan of some of its gains; it had special appeal to the German Kaiser, who again and again returned to his project of a continental league.

Africa was at this time also the locale of a similar imperial competition. When in 1894 the British arranged for the purchase of a strip of the Belgian Congo that would connect their African holdings in the north and the south—the Cape-to-Cairo scheme—joint Franco-German pressure caused them to abandon the plan.

The Fashoda incident

But, on the entire course of the Nile, Britain would not allow any control other than its own. When in 1895 the French launched from Gabon an expedition led by Capt. Jean-Baptiste Marchand that was to march across Africa, effecting a connection with their East African base of Djibouti, a cautionary warning was issued in London. But in September 1898 Captain Marchand reached the Nile at Fashoda, near where British forces under Lord Kitchener had been conducting operations. The confrontation of two handfuls of men on the Nile grew into a major clash between Paris and London.

It was a delicate passage. In Paris the French foreign minister, Théophile Delcassé, a convinced colonialist, came to the conclusion that war with Britain was not desirable. In the face of British intransigence, there was nothing to do but yield. French feeling ran high, but French prestige was seriously injured. Yet out of this situation was to develop one of the most important changes in European power relationships.

Britain's interest in South Africa had been enhanced by the discovery of gold and diamonds in the Transvaal, and Britain was thus brought into collision with that little Boer republic and its twin, the Orange Free State. The South African War began in 1899, after the Fashoda incident had been settled. The British underestimated the guerrilla capacity of the Boers. As a consequence, the protracted conflict entailed a far greater effort on Britain's part than had been thought necessary. In the end the Boers could not stand up unassisted against the resources of Britain. In 1902 the Boer republics were annexed into the colony of the Cape.

The episode had important repercussions. In Britain it led to soul-searching, arousing both strong criticism and nationalistic emotions. Of greater significance was its im-

pact on Britain's international position. The worldwide sympathy for the Boers drove home to Britain the fact of its diplomatic isolation. Britain's traditional stance, that of preferring to remain uncommitted and to eschew alliances in peacetime, might stand in need of reconsideration in the light of the range of Britain's worldwide commitments and of challenges from a variety of quarters. Britain's margin of advantage was diminishing in the face of other rising powers.

Three powers—the United States, Germany, and Japan —were growing at a faster rate than others but at the moment it was Germany that mattered. Since the traditional rivalries with France and Russia still existed, the possibility of a German one being added was a proper source for concern in Britain, especially if it should develop into the Kaiser's plan of a continental league. Preservation of the divisions of Europe, the balance of power, had long been the first directing rule in the defense of Britain's interest.

The German challenge had two aspects. The first was commercial, an expression of the faster German economic growth. It could hardly be expected that Germany should deliberately restrain its expansion in that domain. The other, not unrelated to the first, was in essence political, reflecting the German desire for recognition by others of its right to a place in the sun. Such expressions as *Weltpolitik* and "our future lies on the water," fondly espoused by the Kaiser, expressed it to perfection. In concrete terms, it took the form, in 1898, of embarking on the building of a large navy; it was a considered decision, taken in full awareness of its political implications, the credit—or blame—for which was shared by the Emperor himself and Adm. Alfred von Tirpitz, the father of the German navy. It could hardly be expected that Britain would remain indifferent to such a decision.

Thus, the British reaction to the shifting relationships of power was to re-examine the merits of isolation; put in different form, to reduce the range of commitments by coming to terms with one of the rivals. Bearing in mind Fashoda and the rivalry with Russia across the breadth of Asia, not to mention the currently popular view of the racial affinity and superiority of the Nordic peoples and the tradition of friendly relations with Prussia, Germany was the logical candidate for an approach.

Joseph Chamberlain, a convinced imperialist, opened negotiations with the Germans. Discussions took place on three occasions between 1899 and 1901 but resulted in failure to reach an understanding, a failure that was in large measure due to German psychological misjudgment. The advice of the German chancellor, Bernhard, Fürst von Bülow, to the Kaiser to hold the bid high, regarding Britain as a demandant from whom a political commitment of abstention from continental involvement could be extracted, had a counterproducing effect. Proud Britain may have had difficulties but, after all, had had its way with both the French and the Boers. Even Chamberlain, a Britisher above all in the last analysis, had cautioned his German interlocutors that there were possibilities other than Germany open to Britain.

The first possibility materialized in an unexpected quarter. In 1902 an Anglo-Japanese Alliance was concluded that was not even secret. Though the alliance represented a break in the traditional British avoidance of peacetime alliances, it did lighten the burden of British responsibilities in the Far East and thus enhanced Britain's freedom of action in Europe.

Franco-Italian relations

The opening years of the century witnessed a marked shift in the power relationships of Europe. Following nearly a decade of unfriendly relations with France, Italy put an end to the tariff war with the conclusion of a commercial treaty; it also reached an understanding with France over Tunis. Francesco Crispi's dream of an East African empire was frustrated by the defeat of Adowa in 1896. In 1900 Italy further acknowledged France's interest in Morocco in exchange for a reciprocal French acceptance of Italian hopes in Tripoli. The ambiguous Interpretation of the Triple Alliance that Italy gave France in 1902 to a considerable degree voided that pact of content, though it was simultaneously renewed. Bülow urbanely turned a good face on Italy's potential infidelity, but Germany entertained few illusions on the score of Italian dependability; Italy's position has aptly been described as being "on the fence."

In Serbia a domestic upset resulted in the murder of the pro-Austrian king, an Obrenović, and his replacement by the representative of the rival Karageorgević house; henceforth, Serbia abandoned its subservient Austrian allegiance and turned to Russia for protection.

The most important diplomatic shift at the opening of the century, however, was in Anglo-French relations. Fashoda convinced Delcassé that France should eliminate its traditional rivalry with England. He was ably assisted in this policy by Paul Cambon, his ambassador in London, who found receptive ears in Britain, especially after Chamberlain's failure in Germany.

Unlike the Germans, the French did not insist on a political commitment from England. Protracted Anglo-French discussions resulted in the Entente Cordiale of 1904, whereby France agreed to abandon its withering Egyptian hopes, Britain, in turn, sanctioned French ambitions in Morocco, and a number of other important differences were settled. In the context of the imperial activity of the day, this was a fair and reasonable quid pro quo. The understanding contained no hint of an alliance; it was a public document, but secret clauses were attached to it.

The meaning of this in 1904 was problematic, and it might have proved to be no more than a passing flirtation. Time and succeeding events proved it otherwise, however, and it became the opening of a significant new chapter in the affairs of Europe. The British and the French continued to differ on many things, but what gave solidity to the Entente was the fundamental fact that, allowing for differences of power in 1904, the two countries were increasingly finding themselves placed in a similar defensive position vis-à-vis the rising powers. With a time lag of roughly 50 years, the relative decline of the French power position was followed by a comparable British decline. Here was not a passing condition but rather a long-term change in the position of both countries.

The last decade of peace, 1904–14. Two fundamental assumptions of German policy—the impossibility of either a Franco-Russian or an Anglo-French connection—had proved mistaken. The precise nature of the second was uncertain, but Bülow soon had the opportunity to clarify its meaning. The proximity of two dates is significant: in February 1904 the Japanese, without warning, attacked and destroyed a Russian squadron in Port Arthur; in April the Entente agreement was concluded. France's ally, Russia, was at war with Britain's ally, Japan. The Anglo-Japanese alliance had a restraining effect on the possibility of French assistance to Russia; Britain and France shared a common interest in Russo-Japanese accommodation, and the Far Eastern conflict remained localized.

The Russo-Japanese War

To the surprise of much of the world, the "little yellow men" inflicted a series of defeats on a major European power. To be sure, Russia was handicapped by logistics —the difficulty of supplying a theatre of war over a still incomplete Trans-Siberian Railroad—while the Japanese enjoyed the converse advantage of proximity, hence, of shorter lines of communication. They controlled the sea as well; the pathetic odyssey of Russia's Baltic fleet, sailing halfway around the world, ended in its destruction in the Battle of Tsushima Strait (May 27, 1905). There was nothing for it but to acknowledge defeat, all the more so as internal Russian conditions were taking on a revolutionary hue in 1905. The interposition of American good offices brought the belligerents together at Portsmouth (New Hampshire), where Japan collected the fruits of its competent preparations, having passed the test that insured great-power status. Manchuria and Morocco are far apart, but the simultaneity of the two situations created a connection between them. Britain could indeed renounce its own claims in Morocco; it could certainly not dispose of the rights of third parties. In the form of an ostentatious visit that the German kaiser paid to the Sul-

tan of Morocco, Germany asserted its claim to an equal position in Morocco. The French tactics had been to isolate Germany, and this had been one of their purposes in reaching understandings with Britain, Italy, and Spain in regard to Morocco. On both sides of the Rhine there was complete understanding of the facts and the realities of the Moroccan situation; while Germany felt that Morocco logically "belonged" to France, it would claim a price, as others had, for its consent to that outcome. The issue was, precisely what price? The highest possible was the German view; the smallest was the French.

But Morocco was also a pretext. By asserting equal rights in Morocco, Germany could discover the depth of Britain's commitment in support of France. Bülow himself described his stance as that of "the Sphinx": keep others in uncertainty, and frighten the French; either or both of two results might thereby be achieved: a wrecking of the budding Anglo-French Entente, or the securing of a high price for Morocco.

The Moroccan crises

This is the substance of the First Moroccan Crisis, which meant on the German side raising the spectre of possible war. The French were indeed frightened, to the extent that Delcassé found himself alone among his Cabinet colleagues in opposing a policy of accommodation, insisting that Bülow's tactics were bluff. Properly, he resigned. In what had come to be a duel between two men, the German chancellor had scored the major victory of forcing the resignation of the French foreign minister. But, unwisely, Bülow sought to push his advantage too far, for, if the French yielded to the German demand for an international conference, they were resentful of what they considered undue foreign interference in their domestic affairs.

The Conference of Algeciras in 1906 was a surprise to Germany, which found itself virtually isolated, and the resulting Act of Algeciras, in effect, enhanced the French position in Morocco. Bülow had been all too successful in conveying abroad the impression of German aggressive intent, implemented by tactics of intimidation, the mailed fist, and sabre rattling. One result of the episode was the strengthening of the Entente, the opposite of his intentions; Anglo-French military discussions, again a contingency plan, began in 1906.

After a few years of quiescence, the Moroccan issue flared up again when the French moved in to restore order in the country. The Agadir Crisis of 1911 bears close resemblance in its unfolding to the one just outlined. Sending a German gunboat (the "Panther") to Agadir was a way of asserting a German claim equal to that of the French. From the very outset of the crisis, Britain took a much stronger position in support of France than it had in 1905, and after a tense passage the issue was finally resolved as it might have been in the first place, through a bilateral Franco-German understanding: Morocco to France in exchange for a substantial section of the French Congo to Germany. But the manner in which that result had been procured left a legacy of bitterness and suspicion, especially in Franco-German relations; in France there soon followed the *réveil national*, a resurgence of nationalistic feeling. Between the two Moroccan crises a good deal else had happened.

The impact of the impression created by German tactics in the First Moroccan Crisis and the continued growth of the German navy are to be largely credited for another development. Britain would maintain its naval superiority but proceeded further to reduce the range of its commitments. The Anglo-Russian agreement of August 1907 brought into existence the Triple Entente. The Anglo-Russian understanding is comparable to the Anglo-French one of 1904, dealing with Asiatic rivalries in this case. The heart of it was a division of Persia into three zones—British, Russian, and neutralized. The cry went up in Germany of *Einkreisung*, and indeed encirclement it was, though not with aggressive intent but rather as a response to German impatience expressed in tactics that to others seemed aggressive. Appearance can count for more than reality, and a vicious circle of misunderstanding was closing.

German encirclement

The forces of nationalism, meantime, continued to threaten both the Ottoman and the Habsburg empires. The policy of the latter oscillated between the grant of belated and insufficient concessions and repression. The South Slavs in particular were a troublesome problem, which the shifting of Serbia into the Russian camp in 1903 served to accentuate.

In 1897 Austria and Russia had together "put the Balkans on ice," but the Russian defeat in the Russo-Japanese War shifted the focus of Russian policy back toward Europe. In 1906 the Russian and the Austrian foreign ministers, Aleksandr Izvolsky and Alois Aehrenthal, met and agreed on a scheme for certain changes in the arrangements of 1878. Where Austria was concerned, it would put an end to the ambiguity of its position in Bosnia–Hercegovina by outright annexation of these provinces. The looseness of the understanding and the uncertainties arising from a successful Young Turk takeover in Constantinople—formal title to Bosnia–Hercegovina was still Turkish—induced Austria to create the *fait accompli* of annexation in October. This was the opening gun of the Bosnian annexation crisis of that year and the following. Russia this time, as Germany had done in the case of Morocco, demanded a European conference, which Austria, like France in 1905, was seeking to avoid. The tension lasted until the following spring, when Russia, militarily unprepared and but lukewarmly supported by France, felt that it could not face the prospect of hostilities when Germany took a determined stand in support of its own ally.

The outcome was an undoubted success for the Austro-German combination—a preview in many respects of July 1914—emphasized by a Serbian surrender to an Austrian ultimatum that demanded formal acknowledgment of satisfaction with the annexation. Germany had been the key factor in the resolution of the crisis, but it had acted in support of an initiative that was Austrian; the Bismarckian direction of the alliance had been reversed, in part because Germany considered that in the existing constellation of powers it could not afford to allow its one dependable ally to suffer a setback. But, as a consequence, the distrust and resentment of which Germany was the focus had been deepened.

There followed a two-year lull in the accelerating tempo of crises, during which, however, the arms race was intensified. In 1912 Viscount Haldane, the British war minister, visited Germany in yet another attempt to moderate the naval competition. The effort foundered on the usual reef of the German demand for a political commitment on England's part. Meanwhile, the Agadir Crisis had again raised the spectre of war. Though also resolved, as has been indicated, that confrontation may be regarded as the opening of the last chapter in the story of nearly a half-century of peace.

The Franco-German agreement had not even been concluded when Italy declared war on Turkey for the purpose of acquiring the vilayets of Tripoli and Cyrenaica. The Italians' diplomatic preparations had been adequate, all the powers having at various times acknowledged the validity of the Italian claim. With some difficulty Italy had its way, and the war was concluded with the surrender of Libya by Turkey in October 1912.

The Italo-Turkish War and the Balkan Wars

The Italo-Turkish War was unpopular with the powers, who feared that an attack on the Ottoman Empire might lead to larger complications. These fears were wholly justified. In March 1912, the Russians scored a notable diplomatic success in contriving a Serbo-Bulgarian alliance, which was joined by Greece within two months. It was clearly an offensive alliance, and in October the Balkan allies went to war against Turkey. A final attempt by Austria and Russia to restrain them came too late.

The allies were surprisingly successful. By December, save for a few isolated centres of resistance, the Turks were defeated and surrendered the entirety of their Balkan holdings with the exception of their capital and the straits connecting the Mediterranean with the Black Sea.

The prearranged division of the spoils, underwritten by Russia, proved unacceptable to some of the other powers. Austria, supported by Italy in this case, interposed a veto on Serbian access to the Adriatic. The Balkan allies

fell out. Bulgaria, unwisely, attacked the Serbs and the Greeks and found itself confronted with the Turks and the Romanians as well and shortly went down in defeat. The Treaty of Bucharest (August 1913) settled matters for Bulgaria, but in the western Balkans the powers imposed their own will by creating the state of Albania.

This settlement may be considered as the last instance of the successful operation of the Concert of Europe. The outcome of the Balkan wars was, nevertheless, regarded on the whole as a setback for the Austro-German combination, and the specific form of the settlement had the effect of exacerbating Serbian feeling toward Austria, now the exclusive focus of South Slav irredentism.

Outbreak of World War I

The explosion of the powder keg of Europe, the Balkans, in 1912–13 had failed to ignite the whole continent. But it was only six months before Austria decided to give its own South Slavs a demonstration of its power and determination. On June 28, 1914, the heir to the Habsburg crown, Archduke Francis Ferdinand, visited Sarajevo, the capital of Bosnia, to review army manoeuvres. Some young and irresponsible nationalists, idealistic martyrs to a cause in another view, succeeded in carrying out their plot. Francis Ferdinand and his wife fell victims to their pistol shots—shots that, more than an earlier one, truly were destined to ring around the world. The opening week of August saw all the major powers of Europe, save one, and two minor ones at war. The last chapter of the European Age had begun. (R.A.-C.)

BIBLIOGRAPHY

1500–1648: The following general works include their own bibliographies: *The New Cambridge Modern History*: vol. 1, *The Renaissance, 1493–1520;* vol. 2, *The Reformation, 1520–1559;* vol. 3, *The Counter-Reformation and Price Revolution, 1559–1610;* and vol. 4, *The Decline of Spain and the Thirty Years War, 1609–1648/59* (1957–70), the best recent textbook of general history written in English; *Handbuch der mittelalterlichen und neueren Geschichte*, vol. 2, *Geschichte des europäischen Staatensystems von 1492–1559* (1919; Italian trans., 1932); WALTER PLATZHOFF, *Geschichte des europäischen Staaten-Systems, 1559–1660* (1928, reprinted 1968), a concise, useful account of international relations; GASTON ZELLER, *Les Temps modernes*, vol. 1, *De Christophe Colomb à Cromwell* (1953), a critical and stimulating study; and HENRI HAUSER, *La Prépondérance espagnole, 1559–1660* (1933, reprinted 1972), most illuminating.

1648–1815: Two multivolume series cover this period in its entirety: *The Rise of Modern Europe*, ed. by W.L. LANGER, 20 vol. (1936–71); and *The New Cambridge Modern History*, planned by SIR GEORGE N. CLARK, 14 vol. (1957–). Each set deals at length with diplomatic and military developments. The annotated bibliographies in the Langer series are useful and valuable. For special studies of war and diplomacy, see GASTON ZELLER, *Les Temps modernes*, vol. 2, *De Louis XIV à 1789* (1955); and ANDRE FUGIER, *La Révolution francaise et l'Empire Napoléonien* (1954), both diplomatic and military histories at their broadest, linking them to economic organization and social forces. JOHN U. NEF, *War and Human Progress* (1950); and SIR GEORGE N. CLARK, *War and Society in the Seventeenth Century* (1958), convincingly refute the traditional thesis that war and militarism were the matrix of cultural progress. The later-17th century is treated in two admirable studies by JOHN B. WOLF: *The Emergence of the Great Powers, 1685–1715* (1951) and *Louis XIV* (1968). Both are broadly conceived, rich in details, and persuasively written. Also balanced and judicious is PIERRE GOUBERT, *Louis XIV et vingt millions de français* (1966; Eng. trans., *Louis XIV and Twenty Million Frenchmen*, 1970), which makes a strong case for the greatness of Louis XIV. For Marlborough there is WINSTON CHURCHILL's six-volume authoritative tribute to his ancestor, *Marlborough: His Life and Times* (1933–38). The old study by ALFRED ARNETH, *Prinz Eugen von Savoyen*, new ed., 3 vol. (1864), is still basic. Three consecutive volumes in the Langer series deal with the period from the 18th century to the French Revolution: PENFIELD ROBERTS, *The Quest for Security, 1715–1740* (1947); WALTER L. DORN, *Competition for Empire, 1740–1763* (1940); and LEO GERSHOY, *From Despotism to Revolution, 1763–1789* (1944)—all have chapters on the central themes of balance of power diplomacy, competitive militarism, and the global war for markets and colonial goods. Prussia's place in the wars is discussed acutely in PIERRE GAXOTTE, *Frédéric II* (1938; Eng. trans., 1942). V.O. KLIUCHEVSKY, *A History of Russia*, vol. 5 (Eng. trans., 1931), deals with the diplomacy and wars of Catherine II. For orderly introductions to the troubled years from 1789 to 1815, there are LEO GERSHOY, *The French Revolution and Napoleon*, rev. ed. (1964); and GEOFFREY BRUUN, *Europe and the French Imperium, 1799–1814* (1938). Commercial war is discussed with fresh acumen by ELI F. HECKSCHER, in *The Continental System: An Economic Interpretation* (1922, reprinted 1964). There are many reprints of the once authoritative *The Influence of Sea Power upon the French Revolution and Empire, 1793–1812*, by ALFRED T. MAHAN, 2 vol. (1892). ROBERT B. MOWAT, *The Diplomacy of Napoleon* (1924, reprinted 1971), is a useful brief introduction; while EDOUARD DRIAULT, *Napoléon et l'Europe*, 5 vol. (1910–27), is a storehouse of details. Still quite useful and authoritative is THEODORE A. DODGE, *Napoleon: A History of the Art of War*, 4 vol. (1904–07). The settlement at the Congress of Vienna is studied in CHARLES K. WEBSTER, *The Congress of Vienna, 1814–1815* (1919, reprinted 1963); and, in a pragmatic temper, in HENRY A. KISSINGER, *A World Restored: Metternich, Castlereagh and the Problems of Peace, 1812–1822* (1957).

1815–1914: The best comprehensive treatment of the subject is contained in the two volumes by PIERRE RENOUVIN: *Le XIX^e Siècle:* vol. 1, *De 1185 à 1871* (1954) and vol. 2, *De 1871 à 1914* (1955). A.J.P. TAYLOR, *The Struggle for Mastery in Europe, 1848–1918* (1954), covers most of the period in the author's provocative, if unorthodox, fashion. RENE ALBRECHT-CARRIE, *A Diplomatic History of Europe Since the Congress of Vienna*, 2nd ed. (1973), which covers the 20th century as well, is a dependable, balanced treatment; it can be supplemented by the same author's *Concert of Europe* (1968), a critical documentary collection that deals with the major episodes and settlements of 19th-century international affairs of Europe. ROBERT B. MOWAT, *A History of European Diplomacy, 1815–1914* (1922), is still useful though dated. Two works dealing with British policy, both of them of high quality, deserve special mention: ADOLPHUS WARD and GEORGE P. GOOCH (eds.), *The Cambridge History of British Foreign Policy, 1783–1919*, 3 vol. (1922–23, reprinted 1970); and ROBERT W. SETON-WATSON, *Britain in Europe, 1789–1914* (1937, reprinted 1968). The second half of the first, and the second and third volumes of a collective Russian work, published under the editorship of VLADIMIR P. POTEMKIN and available in French translation, *Histoire de la diplomatie* (1946–47), covers the 19th century; it has the advantages and the limitations that can be expected of a Soviet treatise in this domain. Two books with a revisionist approach are the pioneering work of SIDNEY B. FAY, *The Origins of the World War*, 2nd ed. rev., 2 vol. (1966); and of ERICH BRANDENBURG, *Von Bismarck zum Weltkriege* (1924; Eng. trans., *From Bismarck to the World War*, 1927). The two masterful volumes by WILLIAM L. LANGER, *European Alliances and Alignments, 1871–1890*, 2nd ed. (1950), and *The Diplomacy of Imperialism, 1890–1902*, 2nd ed. (1951), which are also somewhat informed by the revisionist tendency, are very detailed and dependable. Representing a different orientation, but of similarly high quality, is BERNADOTTE E. SCHMITT, *Triple Alliance and Triple Entente* (1934, reprinted 1971). PIERRE RENOUVIN, *Les Origines immédiates de la guerre, 28 juin–4 août, 1914*, 2nd ed. rev. (1927; Eng. trans., *The Immediate Origins of the War, 28 June–4 August 1914*, 1928) deals exclusively with the July 1914 crisis. The single best and most balanced work is LUIGI ALBERTINI, *Le origini della guerra del 1914*, 3 vol. (1942–43; Eng. trans., *The Origins of the War of 1914*, 3 vol., 1952–57), which has the merit of having had at its disposal a fuller documentation. The last two volumes of it, like the second of Fay's work, deal with the July crisis.

(R.E.M./L.Ge./R.A.-C.)

European Law, Medieval

In medieval Europe, law was the expression of a common civilization. This essential unity was not inconsistent with an extraordinary variety of local law at the level of the kingdom, the town, and the village, for embodied in the diversity of individual laws was the reality of a social and religious life common to medieval Europe.

In the Middle Ages a new science of law was born, together with a class of professional jurists to cultivate and practice it. The foundations of monarchical states were laid, and the traditions of the early Middle Ages were adapted to the exigencies of a society in transformation. A complex of public and private rules of law were thereby generated that, for variety within a framework of unity, represent a high achievement of civilization. In no small measure they survive to the present day.

THE SOURCES OF LAW

From personal to territorial law. One may begin by inquiring how law was made in medieval Europe, by

whom, and for whom. In this respect the Middle Ages span a period of profound change. In the 9th century in the Carolingian era, one finds a system of personal laws still operative throughout Europe. Within each kingdom, just as within the Carolingian Empire as a whole, there was not one legal system but a multiplicity of legal systems, written and unwritten, originating in custom or laid down by authority. Every free man, wherever he found himself, could appeal to the laws of his own kin, who might be Romans, Franks, Visigoths, Lombards, Bavarians, Saxons, Angles, or Alamanni.

The Carolingian era

Under the emperor Charlemagne this system began to show signs of strain. Frequent conflicts of law arose as a consequence of increased trade between men of different groups living in the same territory. More and more often it happened that one of a number of laws applicable to the same issue was found preferable because it was clearer or more equitable or better adapted to the region and the times, and it would be singled out by all those in the territory who knew of it, regardless of their origin. Racial mixing, although gradual and uneven, favoured these tendencies. Over a period of about three centuries, a transition took place from a personal system of law to a territorial system, in which the laws a man had to observe were determined no longer by his ethnic origin but by the territory he lived in.

This evolutionary process did not at the outset produce a greater uniformity of law. It is true that the territoriality of written law—of the Carolingian *Capitularia generalia*, for example—brought uniformity within a given kingdom or even throughout the empire; but, in the many important fields of community life not previously disciplined by written law, the establishment of the territorial principle signified rather the triumph of local law. In each separate territory the plurality of coexisting laws resolved progressively into a single common custom. Each land thus had its own *consuetudo loci*, perhaps with an imprint of Roman law, perhaps Germanic, but more often the result of a synthesis of various elements that gave the local system features of its own.

By the 11th century this process had ended in most of Europe. At this moment, when the fragmentation of the law was at its highest, a new development began that was destined to re-establish a common juridical system in continental Europe.

The glossators and the revival of Roman law. The flowering of intellectual life was accompanied by profound changes in economic and social life: the development of trade and craftsmanship, the rebirth of cities, and the appearance of the first universities. All of these played a decisive part in the revival of juridical science. In the city of Bologna, at the end of the 11th century and the beginning of the 12th, scholars began rereading and interpreting the emperor Justinian's 6th-century compilations of Roman law: the *Digest*, the Codex, the *Institutes*, and the *Novels*. A judge and master of the liberal arts named Irnerius, soon to become legendary, laid the foundations of the new school. Its members were known as glossators from the marginal or interlinear annotations (*glossae*) they made in the texts of Roman law. These original texts were incomparably more complex and enlightening than the extracts and summaries that had circulated in the early Middle Ages.

The Bolognese school

The method of painstaking and subtle analysis adopted by the glossators enabled them to demonstrate the inner consistency of the classical texts and their applicability to the most various cases. Though there had been an important revival of juridical science at Pavia in the 11th century, this new advance was on a different plane. The rich juridical literature produced by the glossators took a variety of forms; the individual glosses were assembled in *apparatus*, along with collections of *dissensiones, distinctiones, quaestiones, brocarda, casus, lecturae*, and finally *summae*—a synthesis of the whole system of law. Their activities lasted from the beginning of the 12th century to the middle of the 13th, through five generations of scholars. The terse, concentrated thought of Irnerius was followed by an equitable tendency in Martinus, contrasting with the interpretative rigour of the latter's contemporary Bulgarus. The development proceeded through the realism of Johannes Bassanius and the profundity of Azo to the giant tapestry of law compiled by Accursius in the middle of the 13th century from a selection of his predecessors' glosses. Accursius' *Glossa magna* remained for about 500 years the indispensable complement to the texts of Roman law.

The spread of civil and canon law. The mainspring of all this work was the glossators' conviction that the great Justinian collection—rounded off with the feudal customs gathered in the *Libri feudorum* and with other minor additions—constituted the only true source of *civilis sapientia* for the whole of Christendom, standing above all other laws and customs. Corresponding to civil law was canon law, the province of which was religious. The monk Gratian, a contemporary of the glossators, completed his great compilation of the sources of canon law, the *Concordia discordantium canonum*, about 1140. This was also known as the *Decretum Gratiani*. In ensuing decades there was much papal legislation in the form of rules and principles, often innovatory in their effect, which were issued in deciding the cases that streamed to Rome from all over Christendom. These judgments and decisions were brought together in numerous collections from the 12th century onward, the most famous being that authorized by Pope Gregory IX in 1234 (*Liber Extra*), which proved of fundamental importance in the history of medieval and modern canon law. The works devoted to Gratian's *Decretum* by Rufinus and Huguccio in the 12th century and those devoted to the papal decretals by Henricus of Susa and Pope Innocent IV in the 13th century and by Johannes d'Andrea in the 14th century indicate the intensity of the research that grew up around these sources. The influence exercised by canon law was very profound throughout medieval Europe not only in the ecclesiastical sphere but also in the development of secular law, public and private, penal and procedural. Some important examples are cited below.

The new juridical science spread with amazing speed. Around the masters of Bologna there gathered an ever-increasing number of pupils, *cupida legum juventus*, who were the founding members of what was, with Paris, the oldest of the medieval universities. They came from all over Europe, and, on returning to their native lands, they spread the knowledge of Roman law. At the university they had acquired a capacity to interpret juridical sources and a sureness and logical rigour in argument that were entirely foreign to the jurists of the preceding centuries. The chief reason for the success of the school of Bologna lay in the fact that the political, social, economic, and intellectual awakening of the 12th century had created a need for trained judges, advocates, notaries, and administrators that only the sound and supple teaching of the glossators could ensure. A new figure appeared on the scene, unknown to the society of the early Middle Ages: the professional jurist.

The centres of research soon multiplied. The glossator Vacarius took the new methods to Oxford in England, and Placentinus took them to Montpellier in southern France. Independent centres of scholarship arose in the Rhineland, in Normandy, in Paris, and in England, for canon law, and in Provence and the Dauphiné, for Roman law. By the middle of the 13th century, in Accursius' time, the school of glossators was already in decline. For a few decades, however, some jurists addressed themselves to certain aspects of the law that had been overlooked, such as penal and statute law, which were studied in a practical way quite free from theorizing. Contemporaneously, in the University of Orléans, a second great school of law grew up—that of the commentators.

The commentators. The commentator's new approach was characterized by a greater concern with the law as practiced and by a marked analytic rigour in the interpretation of sources. Their knowledge of the daily business of the law is attested by the vast collections of legal opinions (*consilia*) drawn up by them with reference to particular suits. The commentators made use of this practical experience in their theoretical studies, which they illus-

The new juristic school

trated with hypothetical or actual cases (*quaestiones*). The novelty of this lay not in the literary form of the *quaestio*, already familiar to the glossators, but in its widespread use in the interpretation of Roman texts. The laws commented on were illustrated by means of the questions (*quaestiones*) they might give rise to in practice. But a solution was often impossible within the letter of the Roman law, and, therefore, the commentators were induced to inquire into the principles inspiring the law (*ratio legis*); in this way they succeeded in extending by analogy the validity of the ancient texts to new kinds of cases coming up in the courts.

Primacy in juridical science passed from France back to Italy in the 14th century. Cynus of Pistoia, a poet and contemporary of Dante, was the first to introduce the new method into Italy. One of his pupils was Bartolus of Saxoferrato, whose commentaries excel in their clarity and profundity of thought. No other jurist of the Middle Ages achieved such authority as Bartolus enjoyed for about three centuries throughout continental Europe. The work of Bartolus' greatest pupil, Baldus de Ubaldis, was also very influential.

The commentators dominated juridical science in Europe for two centuries. The training of jurists in the universities of Italy, France, Spain, Germany, and central and eastern Europe followed the method of commentary. The only exception was England, where the education of professional jurists was done in law offices and neither Roman studies nor the universities had any real influence on them.

Influence of custom

Local law. The ascendancy enjoyed by Roman law in European universities did not mean that it was the only form of law. The basis of private law, even in the era of the rebirth of juridical science, was still to be found in the local customs dating from the establishment of territorial law. In the 12th century, throughout Europe, these customs were established in writing. This both strengthened its validity when disputed by sovereigns, lords, or neighbouring cities and satisfied the need for clarity in the law. Many new rules of law were, of course, introduced.

Spain. From the 11th and 12th centuries on, many towns in Castile, León, Navarre, Aragon, and Catalonia obtained from their respective sovereigns grants of privileges and the recognition of local customs (*fueros*). In addition to these municipal *fueros*, regional *fueros* developed, chiefly in the 13th century; they were private collections of laws applying over a larger area. In Castile a central role was played by the legislative work of Alfonso X the Wise, who sought to give statutory form to the juridical principles of Roman and canon law as elaborated by the glossators.

France. At the beginning of the 13th century two distinct juridical areas emerged in France. The customs prevailing in the north (*pays de droit coutumier*) were inspired by Germanic tradition, chiefly Frankish. Some treatises by private individuals at the end of the 13th century (as the *Établissements de St. Louis* and the *Coutume de Beauvaisis*, by Philippe de Beaumanoir) give a vivid picture of them. Only two centuries later, sovereigns promoted an approved collection of customary laws. But in the south (*pays de droit écrit*), Roman law prevailed almost everywhere. The kings resisted it, fearing that open recognition of Roman law would diminish their independence of the empire, since the emperors were considered the personification of Roman law as legitimate heirs of their classical predecessors.

Italy. In Lombardy between the 11th and 12th centuries, a collection of feudal customs was put together that was to form the basis of the *Libri feudorum*, the first organic text of feudal law. This was destined to spread widely and achieve great authority in every part of Europe. The cities of northern and central Italy, on obtaining political independence, wanted an explicit recognition of their customs from the emperor. Subsequently, in the 12th to the 14th centuries, every commune adopted its own municipal law, partly new and partly traditional. Later, the formation of principates led to a revision of municipal statutes, with the aim of bringing them into line with the principles of absolute government. In southern Italy, however, the Norman and Swabian kings were energetic in their policy of centralization; they issued general laws among which Frederick II's *Liber constitutionum* of 1231 stands out for its originality and perfection.

Germany. Territorial law had already become established in Germany when, in the course of the 13th century, certain works of synthesis appeared that had a lasting success. The most famous was written by the judge and knight Eike von Repkow and was renamed the Sachsenspiegel ("Mirror of the Saxons") for its brilliant and lucid exposition of the traditional nucleus of Saxon customs. Although a private work, it was soon invested with the authority of law in northern Germany. A similar work, the *Schwabenspiegel* (*Swabian Mirror*), appeared in southern Germany half a century later. Equally noteworthy was the contemporary flowering of German municipal law; sovereign privileges, autonomous statutes, and private collections of law (*Stadtrechtsbücher*) multiplied in many 12th-century German cities. Later, from the end of the 15th century onward, the widespread acceptance of Roman law stimulated the revision of many city statutes in an attempt to adapt them to Roman principles.

England. The Norman conquest did not destroy the vigour of Anglo-Saxon laws and customs. But the powers of the king's courts of justice and the authority of their decisions so increased as to become the creative source of a new system of law, common to the whole kingdom and prevailing over all other laws and local courts. It became known as common law. A clear picture of the system and its rapid evolution emerges from two famous works, both of which came out of the circle of the royal judges: the *Tractatus de legibus* (*c.* 1190), attributed to Ranulf de Glanville, and the treatise *De legibus et consuetudinibus Angliae* (*c.* 1250), by Henry de Bracton. Coming half a century later, Bracton's tract paints a much fuller and more complex picture; the author cites numerous actual cases in his exposition and exploits the logical and systematic framework elaborated by the jurists of Bologna. Aside from this influence, Roman law did not penetrate England, the effect of Vacarius' teaching having been ephemeral. An essential role was played, on the other hand, by royal legislation. The granting of Magna Carta (1215–25) by John and Henry III was followed in the first period of Edward I's reign (1272–1307) by the publication of legal texts such as the second Statute of Westminster (1285), the influence of which has been felt to the present day in many fields of English public and private law. This was historically all the more remarkable since legislative interference by kings in the field of private law was rare in the Middle Ages.

Common tendencies. Underlying the immense local variety, certain common tendencies can be detected in European law. About one century after Bologna saw the revival of the scientific study of law, the need to provide a clear and organic framework for the various juridical systems was felt almost everywhere. The need was the same, whether the system in question was Roman law (as in the works of Azo and Accursius), canon law (as in the clear syntheses of Henricus of Susa and Pope Innocent IV), Saxon law (as in the *Sachsenspiegel*), English law (as in Bracton's treatise), or French customs (as in the *coutumiers*). All these works were to have centuries of success and in many cases were still very much alive after the end of the medieval era. It seems no mere coincidence that they appeared in the course of a few decades of the 13th century. Without the spread of Bolognese teachings throughout Europe, none of them, not even those furthest from the principles of Roman law, would have been written in quite the same form.

CONSTITUTIONAL LAW

Changes in political structure

The king. The 13th century was also a crucial period for legislation by sovereigns. Royal lawmaking included Frederick II's constitution for the Kingdom of Sicily, Alfonso the Wise's compilation for Castile, and Edward I's legislation in England, not to mention the great collection of decretals authorized by Pope Gregory IX in canon

law. As the age of feudalism waned and the national monarchies began to consolidate themselves, the judge-king of the high Middle Ages gave way to the legislator-king.

At the summit of the medieval order stood the two universal authorities of pope and emperor. The thorny problem of their ideological, political, and juridical coexistence was brought into the open in the 11th century by radical differences over the investiture of bishops. The papacy won its struggle with the Holy Roman Empire, freeing the church of feudal obligations both centrally and in the provinces. The empire remained politically powerful, however, until the middle of the 13th century. In its relationships with the cities and kingdoms, it retained for a long time the prestige of being at the summit of human affairs in the temporal sphere. The grandiose concept of the politico-juridical unity of humankind survived the political failure of the empire.

Even when a monarchy succeeded, as in France or England, in asserting its complete independence from any higher authority, the medieval sovereign always considered himself bound by observance of divine law in dealing out justice or dictating his own laws. This was symbolized in the ceremony of investiture of the king by the church at the moment of coronation.

The question of succession to the throne was resolved in a variety of ways. Wherever the monarchy was weak, as in the post-Carolingian era or in late medieval Germany, the king was nominated by a small group of powerful men of the kingdom. The powerful Capet and Norman dynasties of France and England succeeded, however, in imposing the hereditary principle on their feudal lords, the throne passing undivided to the eldest son upon the death of the king.

At the same time, there was a transformation in the relationship of the king to his people. In barbarian kingdoms the king had been essentially the chief among men of his own kin: *rex Francorum* or *rex Langobardorum*. Later, when victor and vanquished had merged into one people, they became equally bound to their king by an oath of allegiance.

Feudal institutions. Around the 9th century, as feudalism spread throughout Europe, powerful lords interposed themselves between the king and his subjects. Feudalism was based upon the institutions of vassalage and the benefice, or fief. Those who served the king in peace and war received grants of land, in return for which they became his vassals, doing him homage and swearing loyalty to him. The rights and the duties of each party were gradually defined. For a long time, the feudal bond retained a distinctly military character, in that the essential obligation of the fief holder was to support the sovereign in war with cavalry and foot soldiers. From the juridical point of view, the vassal's devotion and loyalty to his lord were of fundamental importance; the gravest crimes in the feudal world were those involving the breaking of faith with the lord (the English word felony originally meant such a breach of faith).

Growing power of the lords

In time, the network of feudal relationships became more complex as the principal fief holders passed on a part of their possessions in fief to others, creating a hierarchy of feudal lordships at the summit of which stood the king. Fiefs were originally revocable at any moment by the king; but between the end of the 9th century and the beginning of the 11th century, they became hereditary everywhere. The precise rules varied. In Germany the sovereign was obliged to make a new grant whenever the benefice reverted to him for lack of heirs or for any other reason.

As the powers of the sovereign declined, those of the feudal lords grew. In the 10th century it even became normal for the sovereign to grant immunity (*immunitas*) to the great lords, giving them the right to exclude the king's judges and tax collectors from their fiefs. The basic functions of government thus passed gradually into the hands of the lords. Along with the power to collect a variety of direct and indirect taxes, they also acquired the right to sit in judgment over their vassals. In its most complete form, this right covered the whole range of jurisdiction, including civil and criminal cases; but in many countries, including England, Germany, and Italy, the feudal lords only exceptionally achieved such wide powers and then only by citing that characteristic institution of medieval law, prescription *ab immemorabili*—the long exercise of such jurisdiction. In cases concerning fief holders, a limitation of the powers of the lord in relation to his vassal grew out of the principle, which spread throughout feudal Europe, that a fief holder had the right to be judged in his lord's court by his peers only.

Communal institutions. The feudal order was eventually upset by new forces, those of the towns and the rising national monarchies. The growth of towns, closely linked to the revival of trade, brought pressures for local independence as townsmen everywhere claimed, with varying success, the rights to self-government, exemption from the lord's justice, nomination of their own judges, and preservation of local customs. An outgrowth of this movement was the city commune.

The characteristics of the commune varied greatly. Some towns remained under the jurisdiction of their lords or of the sovereign himself, although with a limited degree of self-government. In others, after a hard struggle, the citizens were able to form a commune that had the power to make its own laws, mint money, declare war, and enter into alliances. Between these extremes, of course, there was a wide range of local situations.

Between the 11th and the 14th centuries, communal institutions underwent profound changes, especially in Italy and Germany where they flourished the most. The Italian commune abandoned the democratic council form of government (*consoli*) and entrusted the running of affairs to an outsider (*podestà*) appointed every year by a more limited group. Later, in both Italy and Gemany the merchants and artisans formed associations (*arti, Zünfte*), the political weight of which grew until it often achieved the leadership of the commune. In Italy a further stage was reached in the 14th century with the passage of all power into the hands of a single ruler (*signoria*). In the period of the 14th to 16th centuries, the rulers became hereditary princes.

The kings' courts

Monarchical institutions. The 12th and 13th centuries brought, among the monarchies, a progressive strengthening of the king's position. In England, France, and Spain the sovereigns strove tenaciously and long to win direct control of their territory by appointing their own officials throughout the country, free of feudal ties to local lords. The authority of the king's officials gradually increased, notwithstanding the opposition of the lords, especially in the administration of justice.

At the centre was the sovereign, surrounded by a small group of advisers. For financial, legislative, and judicial problems, special bodies of counsellors and magistrates were formed, which acted in the name of the king and were constantly consulted by him. The king's courts of justice acquired increasing influence. In England as early as the 13th century, the more important civil and criminal cases were already being heard in the central courts. Later, in 15th-century England, the expanding jurisdiction of the chancellor's court led on to the development of the system of equity. France developed, from the middle of the 13th century onward, the Roman–canonical appeal system that permitted the highest court of justice in the realm, the Parlement of Paris, to reconsider decisions by the king's local judges; later on, it was able to reconsider sentences passed by the feudal lords as well. A class of professional judges developed in the central courts whose work was to be decisive in the elaboration of a uniform law throughout the kingdom.

The sovereign's power over his subjects was, although ample, not unlimited. His subordination to the higher principles of divine law was recognized in rights such as *Widerstandsrecht* (right of resistance), which could be invoked to resist the sovereign's unjust actions. Above all, the monarchs were restrained by certain representative bodies, the history of which varied in different countries. In 13th-century Spain the king's legislation could not override the deliberations of the Cortes, which already included, in addition to the nobles and ecclesiastics, repre-

sentatives of towns and boroughs. In France the representative assemblies of the three estates (*États généraux*) partially established themselves only later, in the 14th and 15th centuries; the king succeeded in preserving for himself a maximum of autonomy, in some degree tempered by the courts (*parlements*), which were not representative national bodies. Only in England had a constitutional balance between the monarchy and other forces in society already been reached. The king was bound by Magna Carta to seek the lords' assent in the crucial matters of taxation and war. In 1295 representatives of the middle classes and landed commoners entered Parliament, and in 1433 the assent of both houses of Parliament became necessary for the passage of bills.

THE RULES OF LAW

Personal and family law

Persons. Social stratification in the medieval world took not only economic forms but juridical ones as well. The law recognized differences in personal status, generally as a function of birth. On the lowest rung of society were the serfs, tied to the land, subject to the lord, and without property rights. Not until the 13th century did large numbers of serfs begin to win enfranchisement in various parts of Europe. A class of *coloni* enjoyed better status, cultivating land that was not their own and paying a rent to the lord. But in some countries, such as England, the distinction between these two classes had lapsed by the 13th century, and they acquired rights against all but their lords. Under feudalism the nobles occupied the highest rungs of the ladder of power and privilege, at the top of which was the sovereign. With the rebirth of towns, a new class arose, chiefly in Italy, Flanders, and Germany, concerned with trade and crafts. To pass from one class into another, difficult enough in the towns, was almost impossible in the rural areas.

In a society so full of juridical distinctions, minorities often found themselves disadvantaged. Foreigners were in a difficult position. Jews were insecure. Persons holding heretical opinions were threatened by the severest penalties, especially from the 13th century onward.

Women were juridically inferior, as they had been for centuries. Slowly their position changed, and they became able to inherit property. In places where commercial life developed, women acquired, as early as the end of the 12th century, the power to draw up and sign legal documents, such as wills, bills of sale, and company agreements, even without the presence or agreement of a legal guardian.

The family. Almost everywhere the old Germanic traditions of matrimony faded away, and the church's conception was gradually superimposed; as a sacrament, marriage belonged to the spiritual sphere and was therefore regulated by canon law. While tradition gave primacy to the wishes of the betrothed's families, the medieval church vigorously upheld the free consent of the betrothed themselves as essential to the validity of a marriage. Solemn rites were not required until the 16th century.

Within the family the father was pre-eminent, although the classical Roman concept of paternal power had been moderated by medieval tradition and Christian views. In regions where the influence of Roman law was still felt, children remained permanently under paternal control unless they had obtained emancipation. Elsewhere, the Germanic principle prevailed, by which children passed out of paternal control on coming of age. Until they became independent, the father was responsible for his children's acts and administered their property. In the countryside it was still normal in this period for family holdings to be kept undivided and farmed in common. A harsh strengthening of paternal power, particularly in the upper classes, was to come about after the end of the medieval period, in the era of absolute states.

Laws relating to family property were diverse and complex. In certain regions, such as northern France, the possessions of husband and wife formed a single estate. Elsewhere, their property remained juridically separate; the husband contributed various possessions subject to limitations (*donatio propter nuptias, morgengabe, tertia, quarta*), and the wife contributed her dowry, retaining her rights in it, although its administration was her husband's task.

Inheritance. In the Middle Ages the laws of inheritance followed the Germanic tradition for a long period. Heirs were drawn from the family circle according to rigorous rules of succession. No one had the free disposal of his property after death; it was destined for his legitimate offspring or other relatives, and their consent was needed for any alienation of property during his lifetime. Only if there were no descendants could the estate pass to an outsider, but such a person had to be previously adopted and thus juridically a member of the family.

A breach was opened in these principles by the practice of making bequests to churches and religious bodies. The right to do so had been admitted in the early Middle Ages. Free disposal of property after death was fully admitted around the 12th century with the return in many regions of Europe to the Roman *testamentum*, which differed from a donation in that it could be freely revoked by its author.

The laws governing the inheritance of feudal property remained more rigid. By the 11th century it had been established that a fief could be inherited. A little later, the principle gained ground in England, Normandy, and elsewhere that a fief must pass undivided to the eldest son. In England this rule of primogeniture was soon extended to commoners' property, while in other countries it was not. In Italy the fief itself could be divided among the sons at the death of the lord.

Property and business law

Property and possession. The variety of differences in personal status in the Middle Ages was reflected in a diversity of legal rules relating to property, both movable and unmovable. Land, the most precious possession in a predominantly rural society, carried with it a great variety of rights and duties. Traces of the ancient practice of common property in land, pasture, and woods lingered for a long time. Common family rights in property frequently survived, even though individual ownership had become more usual. There was also "servile" land, generally cultivated by men who were not free, in which rights of possession were weak; it was rarely inheritable. The *coloni* held land belonging to a lay lord or an ecclesiastical body, to whom contributions were due in money or in kind fixed by various forms of contract (*census, libellus, precaria,* and others). On a very large scale, land was held of the king by feudal grant or of great lords by feudal lords of lesser rank. Finally, properties might be held outright, free of any dependent tie (*allodio, alleu*); this was not the case everywhere, for England had no absolute ownership.

Whereas in the Roman world the principle had prevailed that property was an absolute and unlimited *dominium*, medieval society developed the different principle of divided ownership. There was, on the one hand, the *dominium directum* of the man who held the land but leased it to another and, on the other, the *dominium utile* of the man who, by virtue of his lease, made use of the land. As the practice of leasing spread widely, the juridical position of the lessee grew stronger to the point of recognizing in many cases a power not originally accorded him—that of ceding his own right in the land to others, either during his lifetime or at his death. But the restrictions on free exchange of property remained considerable, partly because of the principle of inalienability governing the vast ecclesiastical holdings.

In the course of the Middle Ages, the primitive Germanic concept of property was superseded by the Roman concept, in which mere possession, a person's use of things, was distinguished from property, the rightful ownership of things. The principle that possession deserved primary protection also made headway; anyone forcefully dispossessed had the right to repossess through the courts independently of any inquiry into title. The first complete statement of this principle was in canon law, which thus provided an effective means of discouraging the age-old practice of despoilment. Protection of possession was given clear recognition in England from the 12th century by the action of "assize of novel disseisin."

Contract and commercial law. The development in this period in the vast field of contract law was no less profound. In the early Middle Ages agreements had been considered binding if certain rigorous and solemn forms were observed. But the basis of obligation gradually changed. The fundamental element of consent regained full weight, so that the forms observed by the contracting parties were of importance only in proving that a contract existed. The most consistent development of this principle was in canon law, where the effort was made to bring the precepts of the law into line with morality; in canon law, even a simple informal agreement (*pacta nuda*) was considered fully binding. Civil lawyers, however, following Roman law more closely, distinguished between simple agreements and contracts; full protection was given only to agreements entered into under the terms of the law (*pacta vestita*). Both the Roman and the canonical concepts had great influence throughout Europe from the 13th century on and were not absent even in English law, where the history of contract followed a course of its own. In fact, during the age of the English legal writers Glanville and Bracton in the 12th and 13th centuries, the royal court protected formal contracts only, by means of specific actions of the creditor (*writ of debt, writ of covenant*), while a wider protection was assured by ecclesiastical and merchant courts. English common law reached maturity in this field later, at the end of the 15th century.

The law of agency developed at the end of the Middle Ages when it was admitted that a person could assume obligations and rights in the name of another, who was then accountable in law for his agent's actions without any subsequent transfer of rights being required. This fundamental principle, novel to Roman law, made juridical relationships easier in an increasingly dynamic economy. It was partially recognized by 13th-century canon law and subsequently by secular law.

The growth of commerce brought a need for new legal forms. Canon law had rigorously forbidden the charging of interest upon loans, considered to be usury. This was gotten around, sometimes by subtle juridical expedients, and, in the business practice of northern and central Italy, a complex of institutions developed that were to spread throughout Europe. These included the first limited liability companies, from which the joint-stock company, or corporation, later developed. Bills of exchange began to be used, soon becoming an indispensable instrument of credit. The first insurance contracts were drawn up to protect shipowners against losses at sea. Among the compilations of customs relating to commercial maritime contracts, the Barcelona collection was well-known; a 15-century revision of this, the *Consulado del Mar*, or "Consulate of the Sea," became the standard text of maritime common law until the 18th century.

Criminal law. In criminal law the last centuries of the Middle Ages in a large part of Europe may be considered a period of transition. Certain principles rooted in the early Middle Ages were still very much alive. Collective responsibility for the actions of an individual was frequently attributed to a family group, to the inhabitants of a village, or to the culprit's fellow citizens. The outstanding phenomenon of the 9th to 11th centuries was the revival of the practice of self-protection. The absence of authority led men to administer their own justice, falling back upon family solidarity in the form of the vendetta (*faida, Fehde*). Private wars spread in the feudal world.

In the 12th century, renewed attempts were made by the state to eliminate the vendetta by requiring an offender to pay money to the injured party (*compositio, Sühne*). The institutions of the *pax Dei* and the *treuga Dei* were established by the church and the emperors in an effort to ensure brief periods of truce. The principle that crime should be punished by the state rather than by the victims or their families gained ground. States with strong monarchies—England or 13th-century Sicily, for example—made the greatest progress in this direction, although in Germany and Italy an abundant legislation multiplied the punishments inflicted by the state.

Penalities were generally very severe, ranging from imprisonment or retaliatory corporal punishment to banishment or death. Discriminatory treatment accorded to different social classes accentuated this severity, and the moderation of many judges served to temper it only partially. A judge could, however, exercise considerable discretion, as the principle subordinating every crime to a relevant law had not yet matured. The harshness of the penalties could frequently be mitigated by the payment of a fine or by agreement between the accused and the accuser—a survival from the system of private justice. A more general affirmation of the state's exclusive power to punish criminals was established only in the 16th century.

Procedure. During the 12th century a fundamental change occurred in trial procedure. Feudal Europe had reverted to the primitive method of the ordeal or judgment of God, in which a dispute was won by the person who remained unharmed after ordeal by fire and water or who defeated his adversary in a duel. In the 11th century, distrust in this method of proof grew, along with increasingly frequent denunciations of its injustice. In 1215 the church expressly condemned ordeals. Canon law and the revived Roman law required trial to be founded on the rational evidence of written or oral testimony or, alternatively, admitted testimony under oath, widely employed in medieval law. The judge's function was to weigh the evidence according to certain rules.

From ordeal to jury

A different system was adopted in certain northern countries: in Normandy, Flanders, and England the decision was entrusted to a group of men who, by their origin and status, were qualified to determine the truth and to formulate a just verdict (*inquisitio, veritas scabinorum*). In this way, the cardinal institution of the jury developed in the English royal courts.

The parties to an action were represented in court by attorneys and advocates. The Romano-canonical trial was a series of long and complicated procedural stages, and summary procedures were introduced in the 14th century. The right of appeal won wide recognition throughout Europe; it was favoured by sovereigns and utilized very fully by litigants.

Procedure in criminal cases came to differ from that in civil cases. In both canon and secular law the public authorities might intervene and make an investigation (*inquisitio*) into certain crimes and punish culprits even in the absence of an accusation. On the Continent the two procedures, accusatory and inquisitory, survived together for a long time, and only slowly did the latter prevail. In English law, on the other hand, the accusatory procedure remained the only one.

BIBLIOGRAPHY. The classic work on the history of Roman law in the Middle Ages is F.C. VON SAVIGNY, *Geschichte des römischen Rechts im Mittelalter*, 7 vol. (1834–51). A work by various authors designed to provide an updated account is *Ius Romanum Medii Aevi* (1961–); part 5 deals with the spread of Roman law in individual countries. PAUL VINOGRADOFF's lucid outline, *Roman Law in Medieval Europe*, 2nd ed. (1929), still deserves to be read as does PAUL KOSCHAKER'S very fine *Europa und das römische Recht*, 3rd ed. (1958).

Research into the comparative history of medieval European law is still in an initial stage. By 1971 about 30 volumes had been published of the *Recueils de la Société Jean Bodin pour l'histoire comparative des institutions*. Each volume is the work of various authors and deals by country and period with such historical aspects of public and private law as *Le Servage*, *La Tenure*, *La Ville*, *La Femme*, *La Paix*, *La Prevue*, or *La Monocratie*, with large sections on the Middle Ages. Another series *Vorträge und Forschungen*, of which 12 volumes had been published by 1971, is exclusively concerned with the Middle Ages and follows a similar formula, with volumes devoted to kingship, feudalism, and the towns. The Centro Italiano di Studi sull'alto Medioevo in Spoleto, has been publishing annually since 1954 *Atti delle Settimane di Studio*, each volume devoted to a topic of medieval European history.

Outstanding among monographs using the comparative method are those of MARC BLOCH, *La Société féodale*, 2 vol. (1939–40); and HEINRICH MITTEIS, *Der Staat des hohen Mittelalters*, 8th ed. (1968). EDITH ENNEN, *Frühgeschichte der europäischen Stadt* (1953), should be consulted for the

history of towns, and a lucid survey of feudal institutions may be found in ROBERT BOUTRUCHE, *Seigneurie et féodalité*, 2 vol. (1968–70). Notable essays on the comparative history of the medieval theory of the state include ERNST H. KANTOROWICZ, *The King's Two Bodies: A Study in Mediaeval Political Theology* (1957); and GAINES POST, *Studies in Medieval Legal Thought* (1964). Few attempts have been made at a comparative history of private law in the late Middle Ages; one is EDUARD M. MEIJERS' bold and stimulating *Le Droit ligurien de succession en Europe Occidentale*, 4 vol. (1928–36).

Fundamental for the history of political and juridical ideas in the Middle Ages are OTTO GIERKE, *Das deutsche Genossenschaftsrecht*, vol. 3 (1881), trans. in part by FREDERIC W. MAITLAND as *Political Theories of the Middle Age* (1900); and ROBERT W. and ALEXANDER J. CARLYLE, *A History of Medieval Political Theory in the West*, 6 vol. (1903–36).

For medieval canonical sources three classics suffice: JOHANNES F. VON SCHULTE, *Die Geschichte der Quellen und Literatur des canonischen Rechts*, 2 vol. (1875–77); PAUL FOURNIER and GABRIEL LE BRAS, *Histoire des collections canoniques en Occident depuis les fausses Decretales jusqu'au Décret de Gratien*, 2 vol. (1931–32); and STEPHAN KUTTNER, *Repertorium der Kanonistik (1140–1234)* (1937).

The vast *Continental Legal History Series*, 10 vol. (1912–18), consists of translations of histories of national law with additions and notes. A useful guide to further literature is the *Bibliographical Introduction to Legal History and Ethnology*, ed. by JOHN GILISSEN (1963 *et seq.*).

It is necessary to limit suggestions in the field of national histories of medieval law. The classic work for England is FREDERICK POLLOCK and FREDERIC W. MAITLAND, *The History of English Law Before the Time of Edward I*, 2nd ed., 2 vol. (1898), reissued in 1968 with a new introduction and updated bibliography by S.F.C. MILSOM. T.F.T. PLUCKNETT, *A Concise History of the Common Law*, 5th ed. (1956) is a rich general account. For France, an excellent treatment is FRANCOIS OLIVIER-MARTIN, *Histoire du droit français des origines à la Révolution*, 2nd ed. (1951); for private law, PAUL OURLIAC and JEHAN DE MALAFOSSE, *Histoire du droit privé*, 3 vol. (1961–68). For Germany, HEINRICH BRUNNER, *Deutsche Rechtsgeschichte*, 2nd ed., 2 vol. (1906–28) is fundamental, though it stops short at the Carolingian era. RICHARD SCHROEDER and EBERHARD VON KUNSSBERG, *Lehrbuch der deutschen Rechtsgeschichte*, 7th ed. (1932), is particularly informative, while an up-to-date picture is offered by HERMANN CONRAD, *Deutsche Rechtsgeschichte*, 2nd ed., vol. 1 (1962). For Italy, the foremost work is FRANCESCO CALASSO, *Medioevo del diritto*, vol. 1, *Le fonti* (1954), to which may be added the analytical exposition of ENRICO BESTA, *Fonti:legislazione e scienza giuridica*, 2 vol. (1923–25). ARRIGO SOLMI, *Storia del diritto italiano*, 3rd ed. (1930) is a useful synthesis. For Spain, ALFONSO GARCIA GALLO, *Manual de Historia del Derecho español*, 3rd ed., 2 vol. (1967); and LUIS G. DE VALDEAVELLANO, *Curso de historia de las instituciones españolas* (1968), provide lucid general accounts.

(A.P.S.)

European Peoples and Cultures

Among the continents, Europe is unique in several respects: its coast is the most indented of all, its climate the most temperate, and it is the only continent the autonomy of which is culturally rather than geographically determined. Europe is, in fact, only a peninsula of Asia, and, but for its historical and cultural peculiarities, there would be no reason to view it as a separate continent.

In Europe, as in other areas of the world, geography is merely a conditioning and not a determining factor in cultural evolution; this is reflected by the fact that its major cultural areas do not correspond with its main topographical divisions. Not one of the large, more or less isolated geographical divisions—the British Isles or the Scandinavian, Iberian, or Balkan peninsulas, for example—is culturally homogeneous, for any tendency toward such cultural polarization has been negated by numerous other influences, including linguistic, religious, and political expansions and repeated technical and social revolutions.

Elements of European ethnocultural uniformity

The ethnocultural structure of Europe nevertheless shows more uniformity than is found in other parts of the world. None of the other four continents succeeded as Europe did in realizing a racial homogeneity such as the Europoid (Caucasoid), a linguistic predominance such as Indo-European, a religious and moral order such as Christendom, an almost unanimous defense of common values such as the Crusades and the campaigns against the Turks, Tatars, and Moors, a common intellectual ethos such as Latinity, a uniform economico-political structure such as feudalism, or such a great similarity of social structure, based on economic classes.

GEOGRAPHIC INFLUENCES ON CULTURAL EVOLUTION

Explanations of this European cultural unity and of Europe's technical and economic leadership tended in the 19th century to focus on the idea of the superior racial characteristics of Europeans. Geographical and ecological explanations, however, seem at present to have a better scientific basis. Among Europe's geographical advantages have been listed its variety of landscapes, its climatic mildness, its wealth of natural resources, and a location that has left it open to influences from Asia and Africa, thus minimizing the possibility of insular stagnation.

Natural geographical thoroughfares have played a leading part in the shaping of European cultures, prescribing, on the one hand, the intensity and direction of the main demographic streams and, on the other, permitting contact with extra-European groups. Fishermen, hunters, and, later, cattle breeders and agriculturalists followed the large river valleys, so that, since Paleolithic times, the Volga, Dnepr, Danube, Rhône, and Rhine rivers have been the main highways of European peoples and the most important factors influencing culture contacts and diffusion.

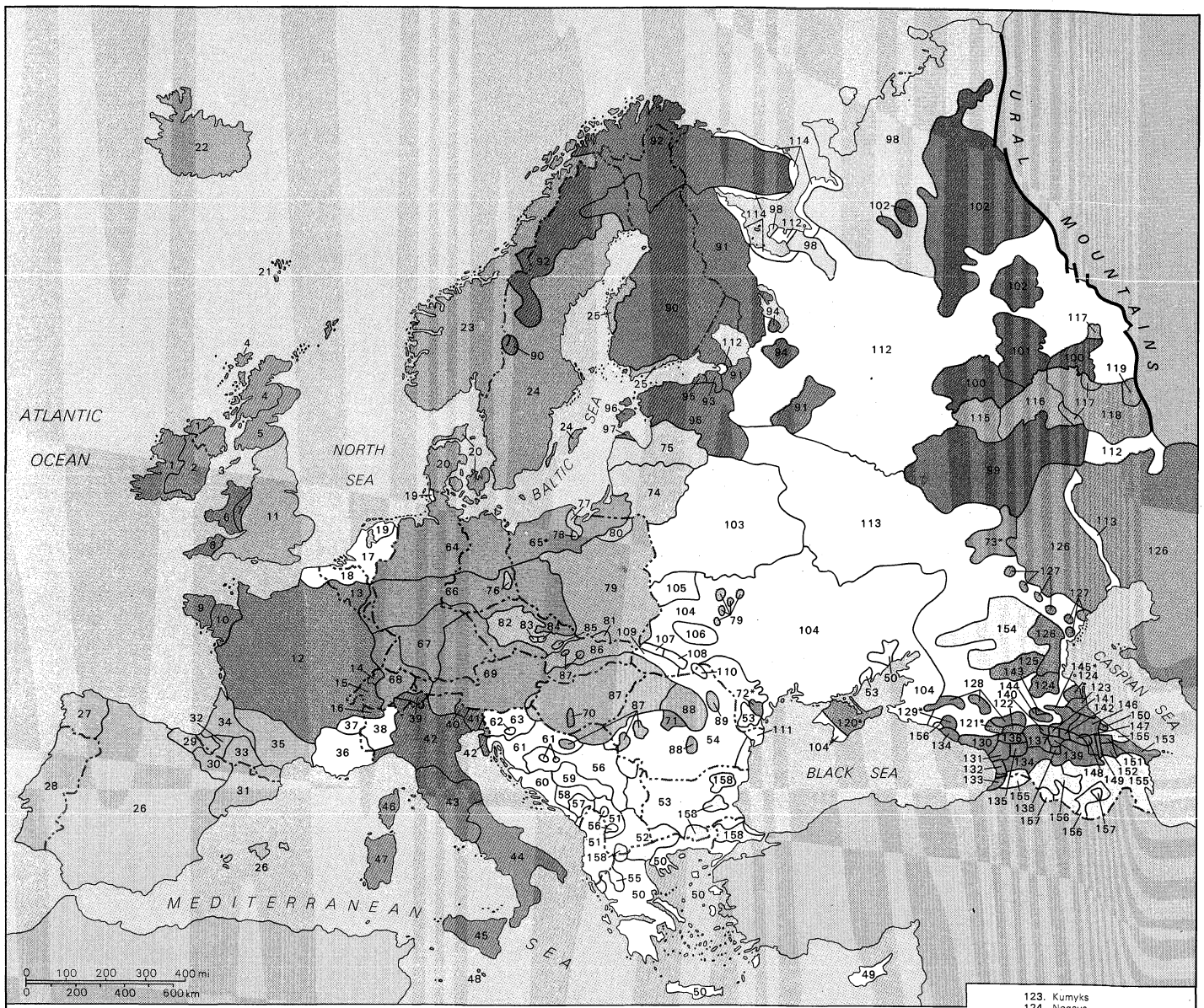
Its geographical location gives Europe three communication junctions with the neighbouring continents: the Iberian Peninsula provides a link with Africa, the Balkan Peninsula connects Europe with Asia Minor, and the large, marshy plain that forms an aperture between the south end of the Ural Mountains and the Caspian Sea acts as a gateway to Central Asia. Ethnic migrations have repeatedly taken place via these passages since the appearance of man in Europe and especially since the Neolithic. During prehistory the migratory direction was almost always centripetal, the refugees or invaders who came from Asia and Africa entering Europe through one of the above routes. But the more evident the economic, political, and military superiority of Europe becomes, from protohistorical times on, the more centrifugal the direction of cultural influence becomes. This process began with Hellenistic colonization in Asia Minor but attained a considerable volume only during the Renaissance. Then European man rapidly discovered and conquered remote territories, with a resulting demographic diffusion that shook the old historical boundaries of Europe and transplanted substantial masses of its inhabitants to other continents. Within four and one-half centuries the temperate zones and large river valleys of North Asia, the eastern Mediterranean, southern Africa, the Americas, Australia, and Oceania were populated by European white men.

PEOPLES AND LANGUAGES

Europe may be ethnologically divided into four major culture areas, then into 22 culture provinces that contain numerous cultural subdivisions from peoples to enclaves (see legend to map).

Difficulties of cultural delimitation in Europe

It should be noted that every cultural division of Europe, above the level of the basic ethnic group termed by anthropologists ethnos (a historically formed autarkic, often endogamous group), is to some extent arbitrary. In primitive societies an ethnos is generally represented by a tribe or a quite distinct tribal unit; in Europe the delimitation of ethnoses is more difficult because of numerous interethnic influences that blur clear boundaries and distinctions. Even so, it is possible to trace the origin of almost every European ethnos to a prehistorical tribe, from which it has usually evolved into a group of villages or hamlets often corresponding to a distinct area of local jurisdiction such as a canton, *Landschaft*, or barony. Ethnoses generally possess their own particular dialects and can be grouped together linguistically into larger units that ethnologists call culture regions. European nations are in most cases divided into several such culture re-



PRIMARY EUROPEAN ETHNIC CULTURE AREAS

I. The Atlantic fringe

Celtic culture area
1. Gaelic-speaking Irish
2. English-speaking Irish
3. Manx
4. Gaelic-speaking Scottish
5. English-speaking Scottish
6. Celtic-speaking Welsh
7. English-speaking Welsh
8. Cornish
9. Celtic-speaking Britons
10. French-speaking Britons

English culture area
11. English

French culture area
12. French
13. Walloons
14. Jurassians
15. Franco-Swiss
16. Valdotans

Netherlandish culture area
17. Dutch
18. Flemings
19. Frisians

Scandinavian culture area
20. Danes
21. Faeroese
22. Icelanders
23. Norwegians
24. Swedes
25. Finlander Swedes

II. The Mediterranean

Iberian culture area
26. Spaniards
27. Galicians
28. Portuguese

Pyrenean culture area
29. Basques
30. North Aragonese
31. Catalans
32. Béarnese
33. Bigordans
34. Gascons
35. Lower Languedocians

Mediterraneo-Alpine culture area
36. Provencals
37. Savoyards
38. Piedmontese

Italic culture area
39. Rhaeto-Romans
40. Ladins
41. Friulians
42. North Italians
43. Middle Italians
44. South Italians
45. Sicilians
46. Corsicans
47. Sardinians
48. Maltese

Balkan culture area
49. Cyprians
50. Greeks
51. Albanians
52. Macedonians
53. Bulgarians
54. Romanians
55. Aromuns
56. Serbs
57. Montenegrins
58. Hercegovinians
59. Bosnians
60. Dalmatians
61. Croatians
62. Chich
63. Slovenes

III. The Carpatho-Alpine-West Plain Climax

German culture area
64. West Lower Germans
65. East Lower Germans*
66. Middle Germans
67. Upper Germans
68. Alemanni
69. Bayuvars
70. Transdanubian Swabians
71. Transylvanian Saxons
72. Moldavian Germans*
73. Volga Germans*

Baltic culture area
74. Lithuanians
75. Latvians

West Slavic culture area
76. Wends
77. Kashubes
78. Slovincz
79. Poles
80. Mazurians
81. Górale
82. Czechs
83. Horaks
84. Zahoraks
85. Hannaks
86. Slovaks

West Ugric culture area
87. Hungarians
88. Siculi
89. Csangos

IV. The East Plain and the Caucasus

West Finno-Permian culture area
90. Finns
91. Karelians
92. Lapps
93. Ingrians
94. Veps
95. Votes
96. Estonians
97. Livonians

Samoyed culture area
98. Nenets

East Finno-Permian culture area
99. Mordvinians
100. Cheremis
101. Votyaks
102. Komi

East Slavic culture area
103. Belorussians
104. Ukrainians
105. Volhynians
106. Podolians
107. Lemks
108. Boyks
109. Ruthenians
110. Huczuls
111. Lipovans
112. North Great Russians
113. South Great Russians
114. Pomors

Pre-Uralian Turko-Tatar culture area
115. Chuvash
116. Kazan Tatars
117. Tepters
118. Bashkirs
119. Misars

Ponto-Caspian Turko-Tatar culture area
120. Crimean Tatars*
121. Karachays*
122. Taulu
123. Kumyks
124. Nogays
125. Turkmens
126. West Kazakh-Kirgiz
127. Lower Volga Tatars

Caucasian culture area
128. Circassians
129. Balkars*
130. Abkhaz
131. Mingrelians
132. Gurians
133. Las
134. Georgians
135. Ymer
136. Svan
137. Ossetes
138. Ingushi
139. Pskhav
140. Mtiul
141. Khevsur
142. Thusk
143. Lezgians
144. Kabarda
145. Chechens*
146. Avars
147. Dido
148. Andi
149. Lak
150. Tabarasans
151. Dargin
152. Kyrinans
153. Tat

Sporadic ethnic units of extra-European origins
154. Kalmyks
155. Azerbaijan Tatars
156. Armenians
157. Kurds
158. Osmano-Turks
Gypsies
Jews

*Peoples partly dispersed by the Soviet government since 1944 (Volga Germans since 1941), but preserving the right to their native homes.

Distribution of European ethnic culture areas.

gions, which may, in turn, be grouped into larger units termed culture provinces or culture sub-areas according to historical, linguistic, and other criteria. This hierarchical classification is illustrated by the following example of a Hungarian village (one asterisk [*] denotes a basic ethnic group; two asterisks [**] denotes other principal ethnic units): village Mezőkövesd, ethnos* Matyók (Mezőkövesd and other villages), culture region the Felföld (Matyók and other ethnoses), culture subprovince Hungarians in a narrow sense (the Felföld and other culture regions), culture province** West Ugrics (Hungarians and other culture subprovinces), culture sub-area the Carpathian Mountains and the Carpathian Basin (West Ugrics and other culture provinces), culture area** the Carpatho-Alpine-West Plain Climax (the Carpathian Mountains and the Carpathian Basin and other culture sub-areas), continental framework Europe (the Carpatho-Alpine-West Plain Climax and other culture areas), intercontinental culture complex** European culture complex (Europe and Europeans in other continents). Linguistically, Europe is dominated by Indo-European languages (*q.v.*) believed to have evolved first on the steppe lands of the Black Sea area or, perhaps, farther east in Central Asia. Archaeology illustrates the general movement westward of people from this area, and some of these movements in the 3rd and 2nd millennia BC resulted in Indo-European forms being modified by local contacts and intermarriage. Proto-Celtic was spoken on the southern German loess lands soon after 1800 BC and spread westward to survive as Gaelic and neo-Celtic (perhaps much modified by contact with some earlier forms) in Ireland, Scotland, the Isle of Man, Wales, and Brittany.

Of non-Indo-European languages, Basque and some Caucasian idioms represent the few survivals of paleo-European cultures. The 14 living Finno-Ugric languages are also autochthonous European, and, among them, Hungarian, Finnish, Mordvinian, Estonian, Votyak, Mari (Cheremis), and Zyrian are each spoken by more than 500,000 people. Turkish languages, in the Middle-Volga-Kama, the Ponto-Caspian, and here and there in the Balkan regions, are results of ancient invasions from Asia.

CULTURE PATTERNS

Religion. Christianity has conquered almost the whole of Europe, with the exception of the Islāmic area in the Balkans and the Turco-Tatar southeast.

The three great branches of Christendom—Catholicism, Protestantism, and Orthodoxy—are linked with Latin, German, and Slavic culture, respectively. Further, almost everywhere in Europe, religion is closely related to ethnic consciousness. In Ireland, Catholicism acts as a shield against English assimilation, and in Finland, Estonia, and Latvia, Protestantism is a fortress against Russification. The spread of Orthodoxy proved to be the first step in the Russian assimilation of the Karelians and other Finno-Permian, Baltic, and Turkish peoples. But it may also be observed that ethnic minorites often survive more because of their religious particularity than because of their language. The Catholic Germans of the Volga, at least until their extermination after World War II, thus preserved their ethnic traditions among Orthodox Russians; the Protestant Saxons of Transylvania likewise survived among Catholic Siculi, as did the Catholic Csangos of Moldavia among Orthodox Romanians. Greek minorities living under Islāmic domination defend their culture first of all through their Orthodoxy.

Religion and ethnic consciousness

The faithfulness to Rome of the Hungarians and Poles is a consequence of their particular devotion to Latinity. In Hungary, including Transylvania, the Latin language was for 1,000 years an intellectual shelter against the permanent threat of Germanization. In the same way, in Poland, it was a defense against Russification. These two countries, actual confines of Western civilization facing Eastern ideology, kept Latin alive longer than any other people. In the 17th century, in Poland, Latin literature reached its highest level of development, and Latin remained the official language at Hungarian diets until 1825 and still later in the universities.

European sociocultural dualism also affects religion. The liturgy, interpretation of dogmas and symbols, and religious practices of the city man come nearer to the official ideas of the high clergy than do those of the common country people. Peasant religious activity is based upon agropastoral life, and pre-Christian ideas penetrate every custom and feast. Fertility rites and invocations for favourable weather and plentiful harvests are prominent in the half-pagan, half-Christian rites and magic still surviving in many areas of Europe, and the calendar of popular religious rites contains feasts grouped around solstices—a practice surviving from prehistoric times. Practices such as noise to drive away harmful jinns, masks to transform the bearers into supernatural beings, ceremonial fire, and various individual abstentions to ensure the success of magic acts are also common in many areas.

Shamanism, which prevailed in many Siberian cultures, may be found still deeply rooted among groups living from the sub-Arctic to the line of the Volga and Kama rivers, where it is strongest among the Nenets (Yurak) Samoyeds. Old pre-Christian ideas and rites are particularly alive among the Cheremis, three-quarters of whom are more or less converted to Christianity, while the remainder continue to practice their ancestral faith, which is linked to the Siberian cults, and refuse any kind of Russification.

The survival of pre-Christian religions in the large eastern plain of Europe is made possible by the fact that in these isolated areas Orthodoxy has always been represented by almost illiterate priests without any theological training. In the Catholic regions of Europe, the clergy demands a stricter break with pre-Christian ideas and practices but often permits customary ceremonies and symbols to survive with a Christianized interpretation. Because of its ascetic principles, Protestantism simplifies to the utmost not only the official liturgy but also the folk ritual. The cult of saints, a foundation of the Catholic tradition calendar, is, for example, almost entirely lacking in most of the Protestant regions. Among Christmas traditions, the Nativity crib is typically Catholic and scarcely appears in Protestant homes. In Switzerland the two confessional groups have different ways of greeting. Diversities also appear in clothing. In Hessen, for example, the traditional costume of Protestant women is dark, whereas the Catholic women's is colourful.

Industrial and traditional ethnic cultures. Superficial observation of the European urban milieu may give rise to a misleading impression of uniformity. In proceeding from the major cities to the villages, burghs, small provincial towns, and even the outskirts of the large, modern cities, the variegated cultural inheritance of the pre-industrial past becomes increasingly evident.

This contrast between urban industrial civilization and traditional ethnic culture is a manifestation of a very old sociocultural dualism that has been dividing Europe more and more deeply since Neolithic times and that is based mainly on differences in life-styles, in means of communication, and, more fundamentally, in the dynamism of the entire culture system. In illustration of the latter dualism, one might point to the role of deliberate and consciously applied individual initiatives in the intricate mechanism of the city, and, on the other hand, to the role of collectively determined and controlled behaviours moved by unconscious, traditional forces in the village. Collective memory may also be a basis of this dualism: in a traditional village, anywhere in Europe, everybody knows everything. Collective knowledge penetrates every social relation and every technical field. Everybody knows the lineage, social connections, daily and seasonal occupations, religious and political beliefs, and working methods of everybody else. Nothing escapes this interknowledge. A peasant, for example, is perfectly acquainted with every technical step of horseshoeing or cart making, even though it is performed by the blacksmith or the cartwright. In the same way, agricultural tasks are familiar to the village craftsmen, even though the latter take no part in farming. This complex mechanism of collective memory is quite similar to that found in the tribal societies that preceded European feudalism (see below

Collective memory and rural traditionalism

Prehistoric cultures) and still exists in some eastern ethnic enclaves of the continent. In urban societies the phenomenon of collective memory disappears or at least is very weak. In large city buildings, even neighbours ignore one another, and extreme professional specialization tends to restrict the individual's knowledge and social relations.

Cultural change. Certain generalizations may be made concerning the processes of cultural change and acculturation in Europe. First, there has been a clear tendency for traditional culture patterns to disappear more rapidly in west central Europe than in the eastern, northern, and southern regions. West central Europe (focussing, say, on the area included in the quadrangle of Paris, London, Amsterdam, and Frankfurt) is also the most economically developed, so that one may point to a certain coincidence between economic and cultural geography; in moving away from this central area, the population's material standard of living gradually declines, but traditional cultural elements are much more in evidence. In other words, the geography of material poverty coincides with the geography of ethnic traditionalism.

In the acculturation process the economically stronger groups are invariably the givers and the weaker the receivers. It is also true, however, that the farther an area is from west central Europe, the more fiercely it protects its cultural inheritance. The real strongholds of traditional cultural elements are, in fact, the islands, such as Iceland, the Faeroes, the Hebrides, the Orkney Islands and the Shetland Islands, the thousands of islets off the Norwegian and Finnish coasts, the Channel Islands, the Balearics, Corsica, Sicily, Malta, and the Greek archipelago, and mountain and moorland areas, such as the Scottish Highlands, the Pyrenees, and the deltas of the Danube and the Po. In these peripheral areas survive linguistic, social, religious, and technical relics of European prehistory. In the Caucasus and the western Pyrenees, for example, were preserved elements of pre-Indo-European languages, while Celtic traditions, once dominant in the Middle Danube Valley, have receded to the farthest coasts and islands of Ireland, north Scotland, Wales, Cornwall, Brittany, and Iberian Galicia. Finno-Ugric peoples, who until the Middle Ages occupied the greater part of the eastern half of Europe, were first driven back and partly assimilated by the Bulgaro-Turks and Turko-Tatars, then squeezed toward the Barents Sea and the Urals by Russian invaders. Some of them, such as the East Ugric Voguls and the Ostyaks, fled toward the Asian slopes of the Urals and are now settled in the basins of the Irtysh, the Ob, and their tributaries.

Strongholds of ethnic traditionalism

Whereas traditional cultural elements move back toward the borders, a dynamical demographic countercurrent flows toward the centre: great wars and political and economical commotions result in migrations mainly directed toward the economically strong countries. The Bolshevik Revolution drove about 2,500,000 Russian citizens westward, and, between World Wars I and II, streams of Polish (450,000), Czech (60,000), Jewish (170,000), and other immigrants rushed into western-European countries. After World War II, 16 million Germans arrived in West Germany from the regions occupied by the Red Army; 200,000 Poles fled to France and England, 415,000 Karelians to Finland, and some 500,000 Balkan and central European people to western Europe. The Hungarian Revolution of 1956 and the Prague revolt of 1968 similarly resulted in a westward flow of population. Workmen's migrations also show a tendency toward west central Europe: in the 1950s and 1960s 1,500,000 Italians, 1,000,000 Yugoslavs, and as many Iberians settled in western Europe. Internal migrations from rural areas, known in every European country, are also directed toward economically prosperous areas. One of the most remarkable examples of the influence of the westward trend of population groups on social morphology is the orientation of the various class districts in the large European capitals. Residential and fashionable districts in Moscow, Berlin, Budapest, Vienna, Rome, Paris, London, and other cities are invariably situated in the "west end," whereas almost everywhere the "east end" is the slummy district.

Such demographical tides have important consequences. On one hand, they increase the density of population in the west central regions, where the natural rate of increase is generally low, and, on the other hand, they deprive the surrounding areas chiefly of their youth and of their most dynamic, transformer elements, thus contributing to their cultural traditionalism.

ETHNOGENIC AND PROTOHISTORICAL BACKGROUND

Bio-racial diversities. There have been numerous attempts by anthropologists to classify European racial types and to determine their geographical distribution. Among the best known classifications of modern times, that of the Russian Joseph Deniker (1900) identifies six main European races and four secondary ones: (1) Nordic, with the Sub-Nordic subrace (tall, long skull, blond); (2) Littoral or Atlanto-Mediterranean, with the northwest subrace (tall, flat head, dark); (3) Adriatic or Dinaric, with the Sub-Adriatic variety (tall, round head, dark); (4) Ibero-Insular (small, long head, dark); (5) Occidental or Cevenol, also called Slavic, Rhaetic (Raetic), or Ligurian (small, round head, dark); (6) Oriental, with the Vistula subrace (small, round head, blond).

Despite his being a sociologist, the American W.Z. Ripley influenced a whole generation of physical anthropologists with his original tripartite racial division (1899): (1) Teutoni, limited to the northwest (tall, long head and face, short nose, blond); (2) Celtic or Alpine, occupying central Europe (small and thick-set, round head, wide face, gray or hazel eyes, short nose, brown hair); (3) Mediterranean, south of the Pyrenees, Provençal coast, and southern Italy (long head and face, wide nose, brown eyes and hair).

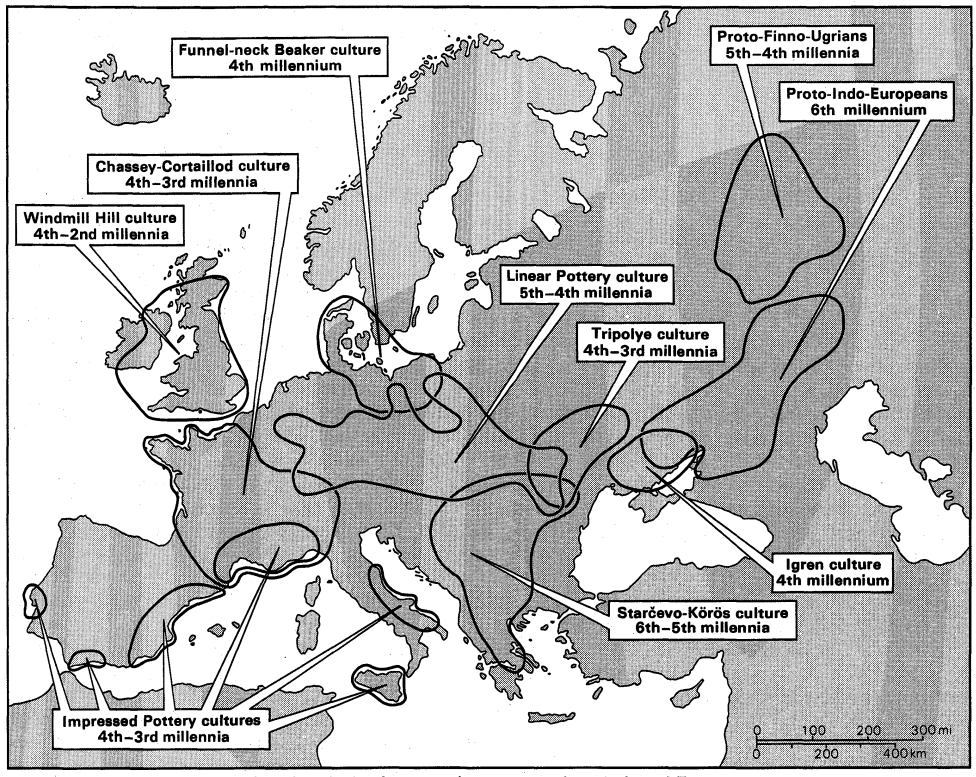
Another American, Carleton S. Coon, is perhaps the most often quoted and his racial classification (1939) probably the most appreciated by his contemporaries. He counts ten European races divided into four major groups: (1) large-headed Paleolithic survivors (fossils); (2) pure and mixed Paleolithic survivors of moderate head size; (3) pure and mixed Mediterranean derivatives; (4) brachycephalic (broad and short-headed, with a cephalic index of over 80) Mediterranean derivatives, probably mixed.

A Russian, N.N. Cheboksarov, on the other hand, divides (1951) the Europeans into only two principal races, with local variations: (1) south European, or Indo-Mediterranean, including the Mediterraneo-Balkanics, the Atlanto-Pontics, and the east Europeans; (2) north European, comprising the Atlanto-Baltics and the Balts of the White Sea.

More recently (1969), an Englishman, John Geipel, has argued the impossibility of establishing a racial classification of Europeans that can be related to definite geographic areas. The dispersal of distinct physical types is, he asserts, so wide that it is impossible even approximately to identify racially homogenous groups.

Prehistoric cultures. Agriculture and the technical, economic, and social developments associated with it first appeared in Europe in the context of Neolithic culture. The first rudimentary origins of agropastoral techniques appeared in the Balkans toward the end of the 7th millennium BC, but, until the beginning of the 2nd millennium, most of the continent was inhabited, for the most part, by gatherers, hunters, and fishermen of a Mesolithic culture stage. Numerous isolated agropastoral communities were in existence as early as 4500 BC, however, from the Lower Danube to the Atlantic. Throughout this period, new waves of migrants from the Near East were arriving in Europe, sometimes via North Africa and the Iberian Peninsula.

Domestication of new types of plants and animals and the introduction of new labour techniques, architecture, and pottery marked the progressive stages in the development of this early European culture. Europe was then covered with dense forests, but the slash-and-burn technique, common among Amerindian primitive agriculturalists and in medieval Europe, was apparently not applied; rather, the first European farmers tended to sow in woodland glades. In the Balkans, the first Neolithic



Presumed locations of the major identified cultures and agropastoral societies of Europe from the 6th to 2nd millennium BC.

houses, around 6000 BC, were small, rectangular, single-roomed, mud-wall constructions. It was among these Balkan people that the earliest distinctive, identifiable European culture—termed Starčevo-Körös—emerged. Based on Near Eastern and Cretan Neolithic traditions, its architecture was poor, and what is known of its culture from clay and pottery figures indicates an earth-mother-goddess cult and a matriarchal social organization.

Linear Pottery culture and its influence

During the 5th millennium, the now so-called Linear Pottery colonists, migrating from the southeast, built ordered villages of huge longhouses with massive timber framing, averaging about 100 by 25 feet (30 by 8 metres). The Linear Pottery culture spread through the river valleys of present north Hungary, Czechoslovakia, Poland, and Germany. Simultaneously with the middle period of this culture, an eastern European agropastoral group, the Funnel-neck Beaker culture, began to expand from the Dnepr River and in the Pontus area. Since the early 4th millennium, its influence has been strongly felt to the north on the Polish plateau as far as the mouth of the Vistula and, later, in the Baltic regions, Jutland, and the Danish islands. In the Middle Danube Valley, around 3800 BC, the Linear Pottery culture evolved into the so-called Note-Head Pottery culture, and between 3500 and 3000 BC the same area was occupied by the Dotted Pottery culture, which expanded as far as northeast France. Simultaneously, in the 4th and the 3rd millennium, agriculturalists settled in the western Mediterranean area and spread into France, Switzerland, and to the north Italian lakes (Chassey-Cortaillod cultures). Their settlements included the so-called Swiss Lake-dwellings, considered for a long time by specialists to have been built on wooden piles over open water. Recent researches have revealed, however, that they were situated on the shore and that their foundations simply served as a protection against ground dampness. At the same time, around 3100 BC, in the Atlantic and west Mediterranean littoral zones, appeared Megalithic cultures characterized by stone-chambered tombs (dolmens) and anthropomorphic stone blocks (menhirs), isolated or in ordered alignments. The orientation of these Megalithic monuments suggests a solar cult, the propagation of which was probably due to the activities of missionary migrants. Around 3000 BC another agrarian culture developed in the Dnepr Valley. Called the Tripolye culture, after the first site discovered near Kiev, it is characterized, like the Linear Pottery culture, by ordered villages of long rectangular houses but arranged in concentric circles. In the centre the cattle, the most valuable asset, were fenced in an enclosure.

Since linguists and ethnologists date the appearance of Indo-European languages in the Neolithic, the question arises as to which of the above described peoples might be identified as the first Indo-Europeans. One of the best known authorities on the problem, the Mexican P. Bosch-Gimpera, has identified the primitive Indo-Europeans as the bearers of the Linear Pottery, the Funnel-neck Beaker, and the Tripolye cultures and has argued that it is impossible to trace Indo-European ethnogenesis back to a single ethnical and geographical unit. The opinion of other scientists is strongly divided on the geographical location of the early Indo-Europeans.

The ethnogenesis of the other important European linguistic family, the Uralians (Finno-Ugrics and Samoyeds), gives rise to less debate: modern authorities agree that their native land is the central and northern part of the large east European plain, somewhere west of the Ural Mountains. A divergent opinion based on scientific evidence suggests an origin east of the Urals.

BIBLIOGRAPHY. ALAN G. OGILVIE, *Europe and Its Borderlands* (1957), is one of the best syntheses of Europe's human geography. The movements of linguistic majorities and minorities is clearly presented by ALBERT DAUZAT, *L'Europe linguistique* (1953); and GUY HERAUD, *Peuples et langues d'Europe* (1968). For a general survey of the ethnocultural structure of the whole continent, see SIGURD ERIXON, "West European Connections and Culture Relations," in *Folk Liv* (1938); M. KENNY (ed.), *Europe and Its Cultures*, a special issue of *Anthropological Quarterly* (1963); ANDRE VARAGNAC, *Civilisation traditionnelle et genres de vie* (1948); and GEZA DE ROHAN-CSERMAK, "La Notion de complexe ethnique européen," in *Ethnologia Europaea* (1967). The best summaries of ethnocultural features according to culture areas and provinces are published in German: GEORG BUSCHAN (ed.), *Illustrierte Völkerkunde*, 3rd ed., pt. 3 (1926); HUGO A. BERNATZIK (ed.), "Europa," in *Die grosse Völkerkunde*, vol. 1 (1939); and TORSTEN GEBHARD and JOSEF HANIKA (eds.), *IRO-Volkskunde: Europäische Länder* (1963). Europe's prehistory is described in several excellent handbooks, including GRAHAME CLARK, *Prehistoric Europe* (1952); VERE G. CHILDE, *The Prehistory of European Society* (1958); STUART PIGGOTT, *An-*

cient Europe (1965); and JAN FILIP (ed.), *Enzyklopädisches Handbuch zur Ur- und Frühgeschichte Europas*, 2 vol. (1966–69). For migratory movements, see VERE G. CHILDE, *Prehistoric Migrations in Europe* (1950); GEZA DE ROHAN-CSERMAK, "Le Principe du 'pont cultural' dans l'ethnologie européenne," in *Actas do Congr. Int. de Etnografia, Santo Tirso* (1963); and EUGENE M. KULISCHER, *Europe on the Move: War and Population Changes, 1917–47* (1948). Among the numerous works on physical anthropology CARLETON S. COON, *The Races of Europe* (1939); and JOHN GEIPEL, *The Europeans* (1969), may be profitably consulted.

(G.de R.-C.)

European Plain, Folk Cultures of the

The European plain is one of the most productive areas in the world and supports one of the greatest population densities. It covers a vast area. East to west it stretches from the Ural Mountains in the Soviet Union to the Atlantic shores of Ireland. In western Europe, it covers a territory that extends from the Pyrenees and the Alps in the south through Denmark and the southern part of Sweden in the north. The north–south spread of the plain is even greater in eastern Europe, where it extends from the Black Sea to the Arctic Ocean.

Landscape and climate of the European plain

Because it covers so much territory, the plain gives Europe the lowest average elevation of any continent. The flatness of this enormous lowland, however, is broken by hills, particularly in the west. Forests and woods also diversify the landscape, although present forests are no more than remnants of a thick mixed forest of oak, elm, ash, linden, and maple, which, since the Middle Ages, has given way to villages and fields in most places.

The climate on the whole is characterized by marked seasonal changes, with cold winters and warm summers. The west has a maritime climate very favourable to agriculture. It has enough rain in all seasons to keep fields green. Summers are warm, but not hot, and winters are cold, but not freezing. As one moves eastward into the zone of continental climate, rainfall is concentrated in the warmer months, and winters become extremely cold. Spring and fall nearly disappear as separate seasons, and the greenness of the summer gives place abruptly each year to the gray drabness of a frozen winter. Agriculture in eastern Europe tends to be more difficult and less productive than in the west.

A variety of languages is spoken on the plain. As in the rest of Europe, almost all belong to the Indo-European family. The one exception is Magyar (Hungarian), a Finno-Ugric language. Four major branches of Indo-European speech are represented. The Celtic branch is found in the form of Irish (Gaelic) still spoken in the western part of Ireland. The Germanic branch is represented by English on the British Isles and by Dutch, Flemish (in part of Belgium), German (including the dialect of Austria), Danish, and Swedish elsewhere. The major representative of the Romance branch is French, along with some inhabitants of the eastern plain who speak Romanian. In most of the east, however, people speak languages of the Slavic branch, of which Polish, Russian, Belorussian (White Russian), and Ukrainian are the most widespread, but also including Czech, Slovak, Slovene, Serbo-Croatian, and Bulgarian.

The European plain includes people as diverse in personality and culture as Irishmen, Russians, Hungarians and Swedes. In spite of important differences, many of these peoples traditionally shared underlying similarities that derived in part from a common pattern of village life and agricultural routine. This pattern was called the open-field system because several landholders owned unfenced plots in the same field.

Open-field communities covered large parts of England in a band running diagonally from the North Sea coast through the Midlands to the English Channel. Across the Channel, they covered major parts of France from the mouth of the Seine to the northernmost bend of the River Loire (particularly the Beauce, Champagne, and Picardy). In the north they were characteristic of Denmark and southern Sweden (Scandia), whereas in central Europe they covered most of Germany as well as the lowlands of Austria, Hungary, and Bulgaria. Eastward, open-field villages were typical of Poland and Great Russia.

TRADITIONAL CULTURE PATTERNS

Economic systems. *Settlement patterns and housing.* In the classic form of open-field husbandry, the land of a village was divided into three fields, and every landholder in a village owned parcels in each field. Use of the fields was rotated so that, in any given year, one was planted in winter wheat or rye, a second was planted in spring oats or barley, and the third field lay fallow to regain its fertility. Only the main fields were fenced in, and then only at times. The three-field system appears to have originated among the Franks when they were ruled by Charlemagne. Late in the 8th century, a number of imperial manors were reorganized on this principle. It seems probable that the Franks established still other such villages in conquered territory, particularly where they put new land under cultivation.

Patterns of village settlement

It took several centuries for the open-field village system to spread. In the 11th, 12th, and early 13th centuries, however, the population of Europe was growing rapidly. The greater part of western Europe, which until then still had a primeval-forest cover, came under the plow. When new villages were established, they were set up on the open-field plan. At the same time, many old villages appear to have been reorganized so that they, too, could be worked in what became the dominant pattern of the plain.

During the Middle Ages, then, a particular pattern of village settlement spread widely over the heart of Europe. It had certain widespread features. Often it focussed on a single street, though sometimes it centred on an open space or plaza. Along the street or around the open space, each landowner had his house.

Typically, the house was single-storied, with a gabled roof of straw, slate, or tile. Anciently, it probably was constructed of wood, but, as the forest thinned out, half-timbering became widespread, the frame of heavy beams filled in with wattle and daub, stone, adobe, or, the most expensive, ceramic brick. Where the forests were completely gone, masonry walls became the norm.

In the Middle Ages, indoor living centred on the hearth. This changed during the late Renaissance (16th century), when the living room was provided with glass windows and chimneyed fireplaces. Because smoke no longer lingered in the room as it sought its way out through cracks, crevices, or holes, interior ceilings could be constructed. The Renaissance room was easier to keep warm and had a new source of light in the daytime. Activities shifted from the hearth to a bench and table by the window. This pattern endured for centuries.

Part of the main building might function as a stall or barn, but often separate buildings were constructed, each serving a special purpose. Whether one building or several, however, they normally had the street or plaza to the front and a kitchen garden to the rear. The relatively small fenced-in garden was largely under the care of the housewife and provided vegetables for the family table.

Open-field villages

At the centre of the village, one generally would find a village pond from which people and livestock got their drinking water. Nearby, a place for village meetings might be indicated by a circle of stones or a large shady tree. Many villages, however, had their meetings on the steps of the church, which normally was located in this part of the village. The village commons, where livestock could be grazed, was usually located near the village if not in its centre. The same was true of small cottages to house landless labourers, the village poor.

As one looked across the plain, the buildings and church of a village stood out like an island in a sea of fields, each community with its fields separated from others by woods and untilled waste. The fields tended to form a crazy quilt, each patch cut into strips that were long and narrow (see Figure). In part, at least, the shape of fields was determined by the equipment used for plowing. Fields were cultivated with a wheeled plow. Wheels made it possible to construct a heavy implement

that supported a knifelike coulter to cut a trench and a deep-cutting moldboard to lift and turn the soil. One reason that medieval farmers were able to spread over the plain and to root out so much of the forest was that they possessed this plow capable of turning the heavy, moist soil. To pull such a plow required from two to six or even eight horses, with at least two people to assist, one to guide the plow and the other to manage the animals. The resultant rig was long and cumbersome and required long fields with plenty of turning room at each end in order to operate efficiently.

Patterns of production and technology. As noted earlier, only the main fields occasionally were fenced in. This saved expense and trouble, but it also had important implications for the farmer who practiced mixed farming, a technique that joined crop cultivation with animal husbandry. Under the mixed farming regime, livestock raised for meat and milk were fed in part by grazing on the commons or on hay harvested from natural meadows. But in part they fed on the stubble of harvested crops. In return, animals provided rich fertilizers for the soil, thus increasing the production of cereals for human consumption.

Open-field farming

The two types of production were mutually supporting. But cattle could not be turned loose to feed on the stubble of one man's strips if the crop on neighbouring strips had not yet been harvested. So long as individually owned holdings were not separately fenced, all harvests had to be brought in before any livestock could graze. This meant that all of the major decisions about cultivation and cattle raising had to be subject to communal sanctions. They had to be synchronized. All had to agree to plow and sow at the same time as well as to cultivate the same crop. Only in that way were they all able to bring in their harvests at about the same time.

The open-field system of land management, then, was associated with a relatively high development of communalism. Communal habits were further encouraged by use of the wheeled plow. Such a plow was expensive, and many farmers did not own one. It also required more horses than one man was likely to own. Villagers found it convenient to share horses and plows and to cooperate in their use.

Property and exchange systems. Men in open-field villages fostered habits of cooperation and sharing in their work. Their communalism was balanced, on the other hand, by a keen sense of individual responsibility and individual property ownership. Most villagers were not completely free, however, and much of the wealth that they produced was not theirs to use, whether individually or communally. A wealthy elite of landlords was legally entitled to part of what was produced. Throughout the plain, particularly in earlier centuries in the west and in the last century in the east, this relationship was shaped by the custom of serfdom.

The landlord and serfdom

The open-field village did not stand alone on the plain. Whether in the village or off by itself, a manor house usually provided separate, more elaborate accommodations for the local landlord. The landlord might also own strips in the fields. Often his fields were consolidated off to one side. Either way, he was entitled to have them worked by villagers, for, in addition to rents and dues of various sorts, the landlord was entitled to free labour by serfs who supplied their own tools. At times, the landlord's entitlement was unlimited. He could demand as much free labour as he needed. But even when limited, it was a difficult burden for peasants, for it required their services at precisely the time that their own fields needed to be worked.

Full serfdom largely disappeared during the late Middle Ages in western Europe. It still existed in the 18th century, but by that time most serfs were neither bound to the land nor subject to unlimited demands for labour. In Russia, serfdom was later to appear and was not officially abolished until 1861.

Villagers, then, lived under a complex system of property and exchange. Under private ownership, they possessed their houses, furnishings, clothing, tools, livestock, gardens, and strips in the three village fields. Under communal sanctions, they coordinated their labour, cooperated in plowing, shared rights to the church and cemetery, shared the village commons, and shared the village forest and waste, where they could gather firewood and graze pigs. Finally, under the law of serfdom, they might be obliged to contribute labour and the use of their tools without remuneration, in addition to making payments of land rents and other dues. Under forms of land tenure other than that of serfdom, they also customarily paid rents, dues, and taxes.

Social structure. *Local and territorial organization.* Some forms of tribal and clan organization existed far back in time, at least in some parts of the plain. But such groupings did not survive the Middle Ages. It seems likely that under the conditions of serfdom, when landlords manipulated peasant families at will, larger forms of social solidarity were destroyed.

Family and kinship patterns. The typical family of the plain was usually the small domestic (nuclear) family. There were exceptions. In parts of Germany, for example, the inheriting son was expected to continue to live with his parents, even after he married and had children. The result was a three-generational family known as a stem family. In Russia and other parts of eastern Europe, several sons might remain in the family after marriage to form a larger, so-called joint family.

Beyond the household of nuclear, stem, or joint family, wider kinship ties were important. Since relationships were traced bilaterally (through both the male and the female lines), every individual had a large number of relatives (the kindred). These ties could be especially strong when they were close in both genealogical and territorial terms. Yet, their importance was limited. Kindred solidarity found expression primarily in the ritual of funerals, weddings, and baptisms, when more distant relatives were invited to the festivities. The individual, however, could not normally expect significant economic help from distant relatives, and even married siblings severely limited the assistance they usually would give to one another. In the economic context, households had to be self-reliant.

Self-reliance of households

Ties of ritual kinship (godparenthood) were universally established at baptisms. Individuals considered it an honour to be asked to become godparents, and acceptance brought with it certain minor obligations, such as to provide gifts for the child. In general, however, people of the plain did not use the institution of godparenthood as a way to establish patron–client relationships or to build a network of ties with individuals from whom economic benefits might be expected.

Stratification. Social stratification existed on two levels in traditional Europe. On the one hand, villagers comprised a class of peasants who had a very different life-style from that of burghers (craftsmen and artisans) or the upper class. Not only did these people have different cultures, they also were ranked, with peasants regarded as belonging at the bottom of the scale.

Within each class, further ranking existed. Among peasants, landless labourers (cottagers) had a much lower standard of living than did landowners. This constituted a fundamental dichotomy in the social structure of the village. Within each of these two basic strata, further ranking was customary. But always, the major division was that which separated the landed from the landless.

Socialization and education. Socialization practices varied greatly from one place to another. Everywhere, children were welcomed, perhaps because everywhere children, and especially sons, added to the work force of the family when they got older. Children also provided support for their aging parents.

Child care was highly variable. The swaddling of infants was widespread but varied as to whether the arms were bound or not and whether the bindings were tight or relatively loose. Because infant mortality was high, child rearing was surrounded with mystical beliefs and practices, particularly during and just after birth. The mother, for example, was often not allowed to leave the dwelling until 40 days had passed, after which she formally was readmitted to the village congregation in a

ceremony presided over by the priest, the so-called rite of churching. The baby itself was baptized shortly after birth, and this normally provided the occasion for a family celebration.

The life of the young

Peasants in most of the plain were uneducated and illiterate. There were exceptions. In some villages, church schools had existed for centuries, so that at least some village children learned to read, write, and do sums. In Denmark, schools could be found in most villages as early as the 18th century. Even where schools were not normally attended, a child generally was taught the catechism prior to his confirmation (Protestant) or first Communion (Catholic). The main education of every peasant child, however, was in the apprenticeship that he did on the farm.

Children learned by doing. As young as seven or eight years of age, they were given chores commensurate with their abilities; they might spend their days tending geese, sheep, or cows, for example. It was customary, though, to send young teen-agers to work as servants or hired hands on other farms, in town, or on manorial estates. In this way, they learned something of how other people lived, though on the whole they remained markedly parochial in mentality.

Religious beliefs and aesthetic systems. The peoples of the plain were Christians. They varied greatly as Orthodox Catholics, Roman Catholics, and Protestants, yet they remained similar in sharing some knowledge of biblical stories and homilies. Everywhere they shared some features of the Christian calendar, including Sundays as a day of rest, and Christmas and Easter, especially, as major annual holidays. Being Christian also meant that village solidarity was increased by the habit of meeting in church as a congregation. Village business often was conducted on the steps of the church or inside. The priest or minister tended to be a man of leadership and high status in the community.

Pre-Christian beliefs survived throughout the centuries. Even though it was contrary to Christian belief, villagers continued to rely on magical potions, omens, and rites to cure disease, help crops, or dispel bad fortune. Christian and pagan beliefs tended to blend, as in the custom of using holy water or the sacred Host as part of a magical substance concocted to improve fertility or cure illness.

Some folk art originated locally, but much of it was built on aristocratic or urban styles. Yet, folk art remained distinctive, because modifications were made during the slow process of borrowing. Peasant styles in the Middle Ages, for example, tended to reflect concepts of Gothic art as these were reproduced by village craftsmen. In the late 16th and early 17th centuries, when peasants were unusually prosperous and open to innovation, elements of Renaissance art largely replaced Gothic elements that still survived in the countryside. Subsequently, one can trace influences of later art styles in painting and woodwork design. Yet, 18th- and 19th-century folk art retained its basic Renaissance inspiration.

General character. On the whole, the folk culture of the plain was that of peasants resident in open-field villages. People varied greatly from one place to another, yet everywhere they shared certain fundamental assumptions as Christians. These included Christian teachings that supported the low status of peasants on the ground that a man's place in life was ordained by God. This apparently made it easier for the lowly to accept their lot, particularly since they were assured rewards in heaven if they were faithful to their calling, just as the landlord would be punished if he were evil.

> O let us love our occupations,
> Bless the squire and his relations,
> Live upon our daily rations,
> And always know our proper stations.

Attitudes and outlook

Peoples of the plain were superstitious. Although they lived in a relatively productive area, there were times of bad harvests, epidemic diseases, and warfare when the peasants were helpless and without defense. Perhaps for this reason they supplemented reliance on hard work and an agrarian technology with belief in miracles and magical interventions.

Open-field villagers lived by an ethic of communalism. They were accustomed to subordinate their wishes to decisions of village ("town") meetings. Yet, a communal orientation was balanced by a keen sense of individual integrity. Each family had the ultimate responsibility for its own affairs, and, though families contributed to cooperative ventures, they did not do so at the cost of compromising their fundamental status as independent social units.

Finally, countrymen of the plain were down-to-earth villagers in orientation. They lived out their lives within the narrow confines of a village and its locality. Often, their personal experience extended only as far as the nearest small market town. They were parochial in the sense that the world they knew was essentially the world of their own parish. And within that world, they were absorbed by the demands of agriculture. Cultivation was not an occupation for them but a way of life. The activities and interests that filled their days, as well as the values that gave meaning to their lives, were shaped by a massive preoccupation with livestock and fields. The folk culture of the plain was shaped by this preoccupation with the soil and the particular form it took under an open-field system and a peasant form of Christianity.

EVOLUTION OF THE CULTURES TODAY

Heritage of integrating and disintegrating factors. The folk culture of the plain was never static. In clothing, art styles, dialect, and other aspects of custom, minor changes were normal. Yet, at times, changes of greater magnitude took place, and this especially was so during the 18th and 19th centuries. It is so now.

Enclosure. During the 18th and 19th centuries, folk life was thoroughly changed by the consolidation of farm plots and the division of communal land. The intent of consolidation was to remove the difficulties of land fragmentation. Each peasant owned strips in various parts of each of the three main fields. He might possess anywhere from 20 to 100 or more small plots. To work them entailed much wasted energy, since a man had to spend time going from one plot to another. The need for access paths also wasted space that could otherwise have been put to the plow.

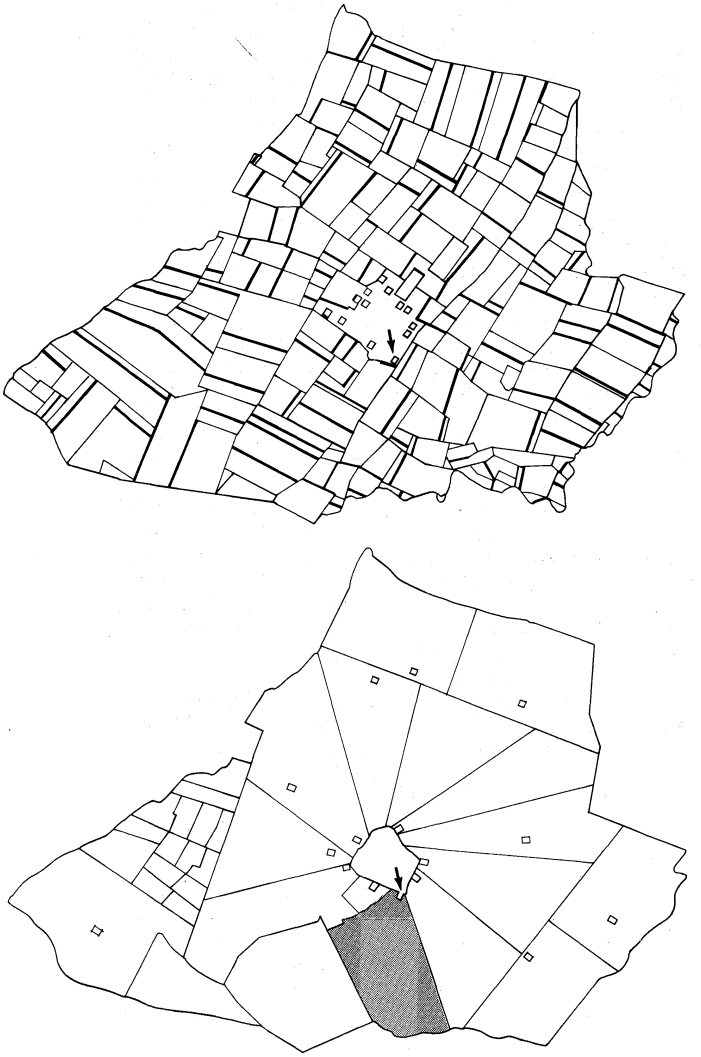
Villagers resisted consolidation. They were aware of the advantages it offered, but they feared the possible disadvantages. The scattering of farm plots usually meant that a peasant had some good land as well as some poor land. Scattering also provided some insurance against catastrophic damage by the weather. A hailstorm, for example, might destroy the crop on some plots but spare others, so the entire harvest was not lost. Consolidation, it was argued, would remove this insurance against the weather. It also might leave a man with fields of poorer quality than he already had. Or, he might be assigned a larger field or two at the margin of village territory, far removed from his house and the community.

In places, it took a century or more to overcome this aversion to consolidate. Legislation made it possible for a minority, perhaps no more than the local landlord and his allies, to force the change. At the same time, communal land was divided up for private ownership among landholders, including the local gentry. This move proved disastrous for smallholders and landless labourers who depended for a major part of their subsistence on the right to graze livestock, for they did not share in the division. Many became paupers. Others found work in new factories.

The move to consolidate fields and divide the commons is termed enclosure, because its visible effect was to put fences around fields as each man enclosed his own property. On the whole, the result was beneficial for landowners. With the opportunity to work their fields individually, they became more open to innovation. The movement prepared the way for the adoption of new kinds of seed, new methods of land use, and new farm machinery. Each family could manage its land more efficiently, but social life and ethos changed.

Results of enclosure

With enclosure, communalism declined dramatically. Each family tended to operate independently of others.



(Top) A village before enclosure and consolidation. Dark lines are scattered strips of farm at arrow. (Bottom) The same village after enclosure. The farm at arrow now has its fields consolidated rather than fragmented.
From H. Thorpe, "The influence of inclosure on the form and pattern of rural settlement in Denmark," *Transaction and Papers*, pub. no. 17 (1951)

The ethic of communalism disappeared, and when 19th-century reformers attempted to persuade farmers to participate in the cooperative movement, they found the farmers reluctant to commit themselves to new group arrangements. Cooperatives now are widespread in western Europe, but they were slow to come.

Communal solidarity was undermined not only by the end of open-field communalism but also by the appearance of a new settlement pattern. Farmers with holdings at the margins of village territory moved their homes to their fields. On the plain today, the territory between villages now commonly is scattered with individual homesteads, each isolated on its own land.

The question of enclosure was clearly decided in England by the end of the 18th century. On the Continent enclosure got under way later and was finished during the 19th century. The movement lagged even further in eastern Europe. Villages in Russia operated under a form of open-field communalism known as the *mir* until the October Revolution of 1917. During the 1920s, the period of the so-called New Economic Policy (NEP), most Russian farming was under a regime of small family farms. Then, beginning in 1928, Joseph Stalin enforced a massive program of collectivization, and within a few years all farming was under one of two forms of socialist farm, the collective (*Kolkhoz*) or the state industrial farm (*sovkhoz*).

Other parts of the east European plain developed much as did western Europe during the period between the two world wars. After 1945, when these nations became satellites of the Soviet Union, they were pressed to convert to socialist forms of farming. By 1962, this was accomplished throughout the plain, with but one exception. In Poland, where today only 13 percent of arable land is in the socialist sector, most of the land is under a regime of small family farms.

Indirect industrialization. By the latter half of the 19th century, a new form of massive change began to take place as an indirect result of industrialization. New products replaced the products of folk arts and crafts. Countrymen of the plain began, for example, to acquire factory-loomed cloth and ready-made clothing to replace homespun, hand-sewn garments. By the end of the century, peasant styles of clothing had been replaced by new urban modes.

Effects of modernization

Cast-iron cookstoves and factory-produced petroleum lamps were widely adopted in these decades. New sources of light and heat were associated with more individual freedom within the family, as evening activities no longer tended to locate the whole family around the fireplace and the table with its oil lamp. Other rooms could be heated and lighted. People stayed up later in the evening, as the old custom of going to bed with the sun came to an end. Since these developments were associated with massive efforts to bring modern schooling to villages, daily activities came to include evening hours in which children did their lessons and parents might read a book or paper as well as knit or do minor repairs on equipment.

An openness to change in the latter 19th and early 20th centuries found expression in part in the modernization of farm technology. In most areas, the old wheeled plow had been replaced by improved implements. Now these too were replaced by still better steel plows and by early forms of machinery for harvesting and threshing grain. In the last 50 years, tractors have replaced horses as the main source of energy. New crops and changed methods have left almost nothing of the old folk economy behind. Subsistence agriculture has been replaced by commercial production, and, with that, commercial attitudes have become widespread.

Current cultural conditions. Although the eastern part of the plain remains old-fashioned in many ways, the western part since the early 1950s has moved importantly toward still one more new style of life. Recent decades have seen urban centres grow into the surrounding countryside. Modern transportation has made it easy for farmers to get regularly to town. In many places, members of farming families, released now from working on the land by the effectiveness of modern machinery, find jobs in town, but without moving from the farm. They commute. Conversely, urbanites find that they can live in villages and work in town. Where this has occurred, the centuries-old distinction between urban and rural cultures (or subcultures) has been obliterated. Even where developments have not gone that far, to the extent that the farmer is no longer parochial, the old distinction is broken down. Folk culture on the plain disappeared in the latter 19th century as an indirect result of industrialization. In recent decades, even the subculture of farmers has tended to disappear as farmers and townsmen have merged.

BIBLIOGRAPHY. LYNN T. WHITE, JR., *Medieval Technology and Social Change* (1962); and ROBERT T. ANDERSON, *Traditional Europe* (1971), provide general information on historical periods, though neither deals exclusively with the plain. The best general work that includes some material on the modern plain is S.H. FRANKLIN, *The European Peasantry: The Final Phase* (1969). England is the most written about in English. GEORGE CASPAR HOMANS, *English Villagers of the Thirteenth Century* (1941 and 1960); and GEORGE G. COULTON, *The Medieval Village* (1925), offer classic accounts of life in open-field villages; modern villages are described in VICTOR BONHAM-CARTER, *The English Village* (1952); and RONALD FRANKENBERG, *Communities in Britain: Social Life in Town and Country* (1966). For the continent, ROBERT T. and BARBARA GALLATIN ANDERSON, *Bus Stop for Paris: The Transformation of a French Village* (1965), describes an open-field community and the changes that took place with urbaniza-

tion. The best work in English on Germany is a chapter in Franklin's book cited above, while for Denmark the most useful remains PETER MICHELSEN and HOLGER RASMUSSEN, *Danish Peasant Culture* (1955). As concerns eastern Europe, the transformation of Russian open-field villages into collective farms is documented in STEPHEN P. and ETHEL DUNN, *The Peasants of Central Russia* (1967); and SULA BENET (ed. and trans.), *The Village of Viriatino: An Ethnographic Study of a Russian Village from Before the Revolution to the Present* (1970; orig. pub. in Russian, 1958), but since both books depend upon official Soviet sources they are somewhat romanticized. EDIT FEL and TAMAS HOFER, *Proper Peasants: Traditional Life in a Hungarian Village* (1969), is thoroughly objective, and includes a brief section on collectivization. Also thoroughly objective and well documented is WILLIAM I. THOMAS and FLORIAN ZNANIECKI, *The Polish Peasant in Europe and America,* 2nd ed., 2 vol. (1972).

(R.T.A.)

Eusebius of Caesarea

The most important 4th-century historian of the Christian Church, Eusebius of Caesarea, lived in that momentous period in Christian history during which toleration was attained for Christianity through the conversion of the Roman emperor Constantine I (died 337) and the doctrine of the incarnation (God made flesh in the Person of Christ) was authoritatively defined at the Council of Nicaea (325). He not only recounted these events but he preserved for posterity a knowledge of some of the personages, events, and movements of Christianity from its beginning until the 4th century. While bishop of Caesarea Palestinae he was involved in the Arian controversy, a dispute over the relationship of Christ as Son to God the Father, and therefore took a relatively active part in the Council of Nicaea.

Early career

He was baptized and ordained at Caesarea, which had become a centre of Christian scholarship through the long residence (*c.* 232–254) there of Origen, the former head of the Catechetical school of Alexandria. Alexandrian learning, doctrinal views, and theological methods that stemmed from Origen were mediated to Eusebius through the teachings of the learned presbyter Pamphilus, to whom he was bound by ties of respect and affection and from whom he derived the name "Eusebius Pamphili" (the son or servant of Pamphilus). Eusebius was an enthusiastic disciple and collaborator.

In 303 the greatest persecution of the church was undertaken by Diocletian (reigned 284–305) and his fellow emperors. In such a widespread persecution it was natural that the Christians of a provincial capital such as Caesarea should suffer, and Eusebius later commemorated them in his *Martyrs of Palestine.* The most poignant event was the long imprisonment of Pamphilus, ending in his martyrdom in 310. Eusebius may himself have been imprisoned at Caesarea or, later, in Egypt. Long afterward he was accused of having secured release by some act of recantation, but such accusations were widespread, and it is unlikely that a willing traitor would have become bishop of Caesarea so soon after persecution ended.

The work of the scholars of the Christian school at Caesarea extended into all fields of Christian writing. They corrected manuscripts of Scripture, an activity that extended even into the period of Pamphilus' imprisonment. They were alive to current theological trends in writing a *Defense of Origen,* who was now subject to attack from various quarters. Eusebius himself wrote a lengthy refutation of the Neoplatonist philosopher Porphyry's attack on Christianity (written after 271), and also against Hierocles, governor of Bithynia, a prominent official and persecutor, who had made a comparison of the Neo-Pythagorean philosopher and reputed wonder-worker Apollonius of Tyana (1st century AD) with Christ. Throughout his life he also produced commentaries on the Bible and works explaining the parallels and discrepancies in the Gospels. His method of exegesis (interpretive criticism) was partly literal and partly allegorical, with a tendency to seek where possible messianic interpretations.

But for subsequent centuries his fame was to rest on his *Ecclesiastical History* (probably begun during the persecution and revised several times between 312 and 324). Here he was conscious of being a pioneer, as indeed he was, for the writing of church history would have been exceedingly difficult before the settled conditions of the late 3rd century and the availability in one place of an adequate library. Eusebius is therefore a chief primary source. He produced what may be called, at best, a fully documented history, and, at worst, collections of passages from his sources. He had already compiled his *Chronicle* (later revised and then expanded by Jerome), an epitome of universal history, and he carried this annalistic method over into his *Ecclesiastical History*, constantly interrupting his narrative to insert the accession of Roman emperors and of the bishops of the four great sees (Alexandria, Antioch, Jerusalem, and Rome). He enlarged his work in successive editions to cover events down to 324, the year before the Council of Nicaea. Eusebius, however, was not a great historian. His treatment of heresy, for example, is inadequate, and he knew next to nothing about the Western Church. His historical works are really apologetic, showing by facts how the church had vindicated itself against heretics and heathens. Of more direct apologetic import were his vast *Praeparatio* ("Preparation") and *Demonstratio evangelica* ("Demonstration of the Gospel"). In the former he refuted pagan mythology, oracles, and astrology, expounded the Jewish scriptures, and argued that whatever was best in Greek thought coincided with or was derived from Jewish thought. In the course of this work he quoted Plato over 250 times and made some use of Neoplatonic writings. The *Demonstratio* dealt with the fulfillment of Hebrew prophecy in Christ.

Election as bishop

The duties of bishop of the metropolitan see of Palestine, assumed by Eusebius in *c.* 313, must have been onerous, and there is little wonder that his literary output slackened. Moreover, the period after the persecution of Diocletian and his successors—which ended in 313—was one of reconstruction, during which multitudes sought admission to the church. He was in Tyre about 316, preaching at the dedication of a new church. Paulinus of Tyre was one of his closest friends; another was Theodotus, bishop of Laodicea in Syria, to whom he dedicated his *Praeparatio* and *Demonstratio.*

When (about 318) the theological views of Arius, a priest of Alexandria, became the subject of controversy because he taught the subordination of the Son to the Father, Eusebius was soon involved; he had already discussed the subjects at issue, for example, in the *Demonstratio.* Expelled from Alexandria for heresy, Arius sought and found sympathy at Caesarea, and, in fact, proclaimed Eusebius as a leading supporter. Eusebius did not wholeheartedly support either Arius or Alexander, bishop of Alexandria from 313 to 328, whose views appeared to tend toward Sabellianism (a heresy that taught that God was manifested in progressive modes). He wrote to Alexander, claiming that Arius had been misrepresented, and he also urged Arius to return to communion with his bishop. But events were moving fast, and at a strongly anti-Arian synod at Antioch, about January 325, Eusebius and two of his allies, Theodotus of Laodicea and Narcissus of Neronias in Cilicia, were provisionally excommunicated for Arian views. When the Council of Nicaea, called by Constantine, met later in the year, Eusebius had to explain himself, and did so on the basis of the baptismal creed of the Church of Caesarea. He was exonerated with the explicit approval of the Emperor. Perhaps Eusebius never forgot that, and though he at first denounced the term *homoousios* ("of one substance") as the key word in the Creed of Nicaea in expressing the relation of the Father to the Son, once more the Emperor's influence, and his own desire for peace, caused him to accept it.

Eusebius and Constantine seem to have met only twice, at Nicaea in 325 and ten years later at Constantine's *tricennalia* (30th anniversary of accession). The years that followed the Council of Nicaea showed how much the Emperor was bent on unity, and the true supporters of

Relations with imperial and ecclesiastical authorities

Nicaea soon found themselves forced into the position of dissidents. Eusebius took part in the expulsion of Eustathius of Antioch (*c.* 337), Athanasius of Alexandria (335), and Marcellus of Ancyra (*c.* 336). But he refused to accept the position of bishop of Antioch, a refusal for which Constantine praised him. He was present, and may have presided, at the Synod of Tyre in 335, where Athanasius was condemned (though not on doctrinal grounds). Summoned, with others, to Constantinople when Athanasius appealed to the Emperor, he did not fail to continue in the Emperor's favour and delivered an oration on his sovereign's *tricennalia*—in which he developed a Christian theory of the empire, to meet the changed conditions of the new age. His ideas on the relation of God, the empire, and the church were based on the view that the emperor ruled by God's grace and was analogous to earlier deified rulers. This view was to collapse when the empire and a large part of the church came into collision under the Emperor's son, Constantius II (337–361). After Constantine's death (337) Eusebius wrote his *Life of Constantine*, a panegyric which possesses some historical value, chiefly because of its use of primary sources. Eusebius was not simply a servile flatterer; he once rebuked Constantia, the Emperor's half sister, for her desire to possess a picture reputed to be of Christ, but she had little political influence.

About 336 Marcellus of Ancyra, a leading supporter of Nicaea, was exiled for supposed Sabellianism. Eusebius and Marcellus had long been hostile toward each other, and Eusebius attacked Marcellus and defended various friends in his "Against Marcellus" and "On the Theology of the Church." These works show that he was still not a defender of the doctrine of Nicaea, but he had moved nearer to that position. Eusebius died about 340.

Acacius, his pupil and successor at Caesarea, wrote a life of Eusebius, which is not extant. But if Acacius and another pupil, Eusebius of Emesa, can be taken to represent their master's theology, it is of a semi-Arian type. Later generations were puzzled by his services to Christian scholarship coupled with his Arian leanings. His only wholehearted supporter was the church historian Socrates (*c.* 380–*c.* 445). Eventually, the seventh ecumenical council (787), at Nicaea, condemned him, finding him double-minded and unstable in all his ways.

BIBLIOGRAPHY. All existing works may be found in J.P. MIGNE, *Patrologia Graeca*, vol. 19–24 (1857); the critical edition of most works in *Die griechischen christlichen Schriftsteller der ersten drei Jahrhunderte*, 8 vol., pub. by the Berlin Academy (1902–56). English translations include the text of the *Church History* from this edition, trans. by K. LAKE and J.E.L. OULTON, 2 vol. (1926–32); *Eusebius: The Ecclesiastical History and the Martyrs of Palestine*, trans., with introduction and notes, by H.J. LAWLOR and J.E.L. OULTON, 2 vol. (1927–28); the text of the *Praeparatio evangelica*, trans. and notes by E.H. GIFFORD, 4 vol. (1903); and the Eng. trans. of the *Demonstratio evangelica* by W.J. FERRAR, 2 vol. (1920). Other translations may be found in *Nicene and Post-Nicene Fathers of the Christian Church* (1890). For editions and studies to date, see B. ALTANER, *Patrologie*, 6th ed. (1963); H.J. LAWLOR, *Eusebiana* (1912); D.S. WALLACE-HADRILL, *Eusebius of Caesarea* (1960); J.N.D. KELLY, *Early Christian Doctrines*, 4th ed. (1968); R.P.C. HANSON, "Biblical Exegesis in the Early Church," *The Cambridge History of the Bible*, vol. 1, ch. 13 (1970); and the bibliographies in J. MAROUZEAU, *L'Année philologique* (annual).

(J.St.)

Evans, Oliver

Oliver Evans was one of the most versatile geniuses of the Industrial Revolution, yet one of the least known in modern times. He pioneered the high-pressure steam engine, an improvement on the inventions of Newcomen and Watt that made steam power widely applicable in industry and transportation; and he created the first continuous production line, an innovation 50 years or more ahead of its time.

Born in Newport, Delaware, September 13, 1755, 21 years before the American Declaration of Independence, Evans was apprenticed to a wheelwright at 16. Observing the trick of a blacksmith's boy who used the propellant force of steam in a gun, he began to investigate ways to harness steam for propulsion. Before he could successfully pursue this line of research, however, he became involved with a number of other industrial problems. Carding, or combing, fibres to prepare them for spinning was a laborious process constituting a bottleneck in the newly mechanized production of textiles. To speed this operation Evans invented a machine that cut and mounted 1,000 wire teeth per minute on leather, the teeth serving as an improved carding device.

Evans, portrait by W.G. Jackman, 19th century. In the Metropolitan Museum of Art, New York.

By courtesy of the Metropolitan Museum of Art, New York, gift of Mrs. Van Santvoord Merle-Smith, 1933

Work on automated production

In 1784, at the age of 29, he attacked another major industrial production problem, the age-old process of grinding grain. Building a factory outside Philadelphia, and adapting five machines, including conveyors, elevators, and weighing scales, he created a production line in which all movement throughout the mill was automatic. Labour was required only to set the mill in motion; power was supplied by waterwheels, and grain was fed in at one end, passed by a system of conveyors and chutes through the stages of milling and refining, emerging at the other end as finished flour. The system, which reduced costs by 50 percent according to Evans' calculations, was widely copied in United States flour milling.

When Evans applied for patent protection, first to state governments (1787) and later to the new United States Patent Office (1790), he added a third invention, his high-pressure steam engine. He continued to work on this for the next several years, envisioning both a stationary engine for industrial purposes and an engine for land and water transport. In 1801 he built in Philadelphia a stationary engine that turned a rotary crusher to produce pulverized limestone for agricultural purposes. The engine that became associated with his name was an original adaptation of the existing steam engine; Evans placed both the cylinder and the crankshaft at the same end of the beam instead of at opposite ends, as had been done previously. This greatly reduced the weight of the beam. An ingenious linkage, which became world famous as the Evans straight-line linkage, made the new arrangement feasible. He saw at once the potential of such an engine for road transportation but was unable to persuade the authorities to permit its use on the Pennsylvania Turnpike—not unnaturally, since it might well have frightened the horses, which at that time provided the main form of transport. Within a few years he had engines doing several other kinds of work, including sowing grain, driving sawmills and boring machines, and powering a dredge to clear the Philadelphia water frontage. Completed by June 1805, his new type of steam-engine scow, called the Orukter Amphibolos, or Amphibious Digger, was 30 feet (nine metres) long by 12 feet (four and seven-tenths metres) wide. In its machinery it embodied the chain of buckets principle of his automatic

flour mill. Equipped with wheels, it ran on land as well as on water, making it the first powered road vehicle to operate in the United States.

Building of steam engines

In 1806 Evans began to develop his noted Mars Works, where, over the next ten years, he made more than 100 steam engines that were used with screw presses for processing cotton, tobacco, and paper. The Navy Yard in Washington bought one of Evans' engines, and when the War of 1812 broke out Evans and a partner proposed to build a powerful steam warship with a large gun at the bow, thus anticipating John Ericsson's "Monitor" of 50 years later; but the proposal was not accepted.

Evans' last great work, completed in 1817, was a 24-horsepower high-pressure engine for a waterworks. He died on April 15, 1819, shortly after a disastrous fire that destroyed his Mars Works, including his valuable patterns and molds.

His *Young Mill-Wright and Miller's Guide*, which he had written back in 1792, continued to sell and had gone through 15 editions by 1860. In another work, *The Abortion of the Young Steam Engineer's Guide* (1805), he forecast the need for government subsidization of technological advances.

Vested interests in horses, poor roads, steep gradients, inadequate springing, and an inadequate technology of materials hindered the adoption of his ideas for steam engines on roads; and also because of the slowness of later manufacturers to make use of his innovative manufacturing techniques, Evans was long a somewhat neglected figure. More recently, however, in the allocation of priorities for the development of the high-pressure steam engine, the simultaneity of Evans' work with that of the British genius Richard Trevithick has been established, and historians have accorded proper credit for his pioneering of the assembly line.

BIBLIOGRAPHY. GREVILLE and DOROTHY BATHE, *Oliver Evans: A Chronicle of Early American Engineering* (1935), a massively authoritative study that is not likely to be superseded, since it is based on meticulous historical research, with copious references; SIEGFRIED GIEDION, *Mechanization Takes Command: A Contribution to Anonymous History* (1948), a stimulating impressionistic account of the development of mechanization from Evans' day to the end of World War II, providing a framework in which Evans' ideas can be assessed; as does H.W. DICKINSON, "The Steam Engine to 1830," in CHARLES SINGER *et al.* (eds.), *A History of Technology*, vol. 4, pp. 168–198 (1958), the most recent synoptic view of the steam engine, which places Evans in context with his English and European contemporaries; G. DOORMAN, "Dredging," in the same volume, pp. 628–643, synthesizes the expert studies on this subject with a most helpful bibliography.

(W.H.G.A.)

Evaporites

Evaporites are mineral deposits that form by the precipitation of salts in restricted water bodies such as inland seas, semi-isolated coastal waters, lagoons, and terrestrial playa lakes. When such water bodies are subject to evaporation losses that equal or exceed the water supplied by rainfall, runoff, groundwater flow, or the open sea, the salts in solution become concentrated and, by definition, the salinity increases. Ultimately a saline brine will form and salts will begin to precipitate from the brine. Because evaporation is the principal cause of the salt concentration and precipitation, the individual minerals and the deposits as a whole—whether marine or nonmarine in origin—are termed evaporites, after the process.

Sequence of precipitation from seawater

The precipitation of salts from a brine is governed chiefly by the solubilities of individual salts and by the brine temperature and composition. The first investigation of the order of salt precipitation from seawater was undertaken by the French chemist J. Usiglio, who evaporated a water sample from the Mediterranean Sea in 1849 and periodically analyzed the mineral precipitates and residual brine. When the sample was reduced to about 50 percent of its original volume, calcium carbonate ($CaCO_3$) began to precipitate. When reduced to about 15 to 20 percent of the original volume, calcium sulfate ($CaSO_4$) began to precipitate. Only after reduction to 10 percent of the original volume did sodium chloride (NaCl) precipitate; further evaporation led to the separation of magnesium sulfate ($MgSO_4$) and magnesium chloride ($MgCl_2$), followed by sodium bromide (NaBr) and potassium chloride (KCl).

Usiglio's results are roughly in accord with the sequence of minerals observed in some evaporite deposits. An occurrence of calcite ($CaCO_3$), gypsum ($CaSO_4 \cdot 2H_2O$) or anhydrite ($CaSO_4$), halite (NaCl), and a small quantity of magnesium, sodium, or potassium salts is not unrepresentative. But the deposition of evaporites in nature is not nearly so simple a process as the continuous evaporation of a seawater sample in the laboratory. Changes of temperature, salinity, and brine composition in nature may occur for a wide variety of reasons; such changes may lead to the re-solution (or alteration) of previously deposited minerals or to the formation of new and different ones. Hence, evaporite deposits may consist of repetitive sequences of only two or three common minerals such as calcite, anhydrite, and halite, or of an enormously complex array of 30 or more individual mineral types.

More than 80 minerals have been identified in evaporites of marine origin. These are principally carbonates, sulfates, and chlorides, with some borates, silicates, and other mineral groups. Evaporites formed in nonmarine environments may contain many of the same mineral species, but, as a generalization, they exhibit still greater diversity. This is because the composition of seawater is everywhere more nearly constant than the waters that feed playa lakes, which reflect the vagaries of neighbouring source areas. Borates, nitrates, and sulfocarbonates occur more often in nonmarine evaporites.

Evaporites occur in strata of every geological system, beginning with the Cambrian Period, 570,000,000 years ago. They are widespread, occurring on every continent, but a gradual restriction toward the Equator is evident from consideration of their distribution through time. This fact, combined with the need for high evaporation rates, has led many authorities to believe that evaporites are good paleoclimatic indicators—*i.e.*, that their occurrence signifies the existence of a generally arid climate at the time of their formation.

Sites of modern deposition

Areas of modern evaporite formation tend to sustain this view; noteworthy occurrences include: Great Salt Lake, the Death Valley salt pan, and many other playa lakes in the western United States, Chile, Australia, and other arid areas; Kara-Bogaz-Gol Gulf in the eastern part of the Caspian Sea; coastal lagoons along the margins of the Persian Gulf; the Rann of Kutch in India; Lake Larnaca, Cyprus; Lake Assal, Afars and Issas, formerly French Somaliland; the Dead Sea, between Israel and Jordan; and parts of the Red Sea. In each of these cases, the climate is generally arid, evaporation rates are high, and water salinities are about seven to ten times that of normal seawater, except in the rainy season.

In addition to their possible paleoclimatic significance, evaporites are of much interest for paleogeographic reasons; this stems from the occurrence of thick deposits in nature. It has long been known, for example, that evaporation of a column of seawater 100 metres (330 feet) in height would lead to deposition of a salt bed only 150 centimetres (60 inches) thick. Moreover, only four centimetres (1.5 inches) of this salt would be gypsum (or anhydrite), a mineral that is present in beds whose aggregate thickness is as great as 500 metres (1,600 feet) or more in some sequences. The quantities and depths of water required to produce 500 metres of anhydrite are so vast that it has long been obvious that some resupply mechanism must be involved.

A "bar theory" of evaporate deposition was set forth in 1864 to explain these thick evaporite deposits. Briefly stated, this theory, and its subsequent refinements, simply assumes the existence of an offshore bar that restricts free circulation between the open sea and a coastal lagoon area. The bar would permit inflow of seawater to the lagoon, however, to replenish the water lost through evaporation and the salt lost through precipitation. The principal notion involved is the constant resupply of saline water, which eliminates the need for an initial water body that is nearly oceanic in size.

The paleogeographic or environmental setting is still subject to debate, however. Aside from clearly continental playa deposits, the association of some evaporites with coral-reef deposits, red beds, or fluvial structures suggests shallow-water or terrestrial environments. Indeed, there are many authorities who would interpret most deposits as originating in this way. Others, however, argue for deepwater basins, and still a third idea involves a two-layer aqueous system in which a lower brine layer is exposed intermittently to evaporation.

Importance to man

Evaporites are of considerable importance to man. They are the source for the world's supply of common salt (halite), gypsum, and much potash and magnesium. Borates and nitrates of economic importance also are derived from evaporite deposits. In addition, salt beds that are confined by the weight of overlying sedimentary rocks eventually tend to rise upward through this overburden, forming salt domes. These salt structures are intimately related to petroleum and natural gas occurrences in many parts of the world because they trap or block the subsurface migration of oil and gas and cause its accumulation.

The low permeability of salt, which makes it an effective seal against petroleum, is useful in other ways. Caverns in salt are almost perfectly dry, and it has been suggested that liquefied petroleum gas could be efficiently stored in artificial caverns in salt. Such caverns also might be used for the storage of certain radioactive wastes.

This article treats the general characteristics of evaporite minerals and mineral equilibria, the several theories of origin with regard to depositional environments, and the occurrence and distribution of evaporites in the geological record. For further information on carbonate minerals and rocks, see LIMESTONES AND DOLOMITES and CARBONATE MINERALS. For data on the composition and geochemistry of seawater, see OCEANS AND SEAS and GEOCHEMICAL EQUILIBRIA AT LOW TEMPERATURES AND PRESSURES. The several environments that are related to the occurrence of evaporites are further treated in the articles LAGOONS; PLAYAS, PANS, AND SALINE FLATS; CORAL ISLANDS, CORAL REEFS, AND ATOLLS; PALEOGEOGRAPHY; and SEDIMENTARY FACIES. See also SALT; SALT DOMES; and the several articles on saline water bodies—*e.g.*, RED SEA.

EVAPORITE MINERALOGY

Classification of evaporite minerals

Evaporite mineralogy appears formidable to the uninitiated because of the large number of mineral species involved and the unfamiliar ring of their names and chemical formulas. Examination of some relatively common evaporite minerals, however, reveals that a somewhat restricted set of constituents is present (see the Table). The minerals are categorized in four principal groups: carbonates, sulfates, chlorides, and borates and nitrates. The four principal cations (positively charged ions) associated with these groups are sodium, calcium, potassium, and magnesium, which are, of course, the principal salts in seawater. The Table also reveals that several pairs of mineral species are distinguished solely by reason of different numbers of water molecules contained in their crystal structure. Gypsum and anhydrite; vanthoffite and bloedite; and colemanite and inyoite are paired examples of this fact.

Having said this, it must be admitted that variety is one of the hallmarks of evaporite mineralogy. In addition to the minerals listed in the Table there are perhaps another 70 species that have been identified in marine and nonmarine deposits. They include some fluorides (fluorite and sellaite), iron oxides (magnetite, hematite, and limonite), silicate minerals (quartz, opal, and talc), clay minerals (kaolinite, illite, and sepiolite), phosphates (goyazite), sulfides (pyrite), and even native elements (sulfur).

Some Relatively Common Evaporite Minerals*

	formula
Carbonates	
Calcite	$CaCO_3$
Aragonite	$CaCO_3$
Magnesite	$MgCO_3$
Dolomite	$CaCO_3 \cdot MgCO_3$
Natron	$Na_2CO_3 \cdot 10H_2O$
Trona	$Na_2CO_3 \cdot NaHCO_3 \cdot 2H_2O$
Nahcolite	$NaHCO_3$
Pirssonite	$Na_2CO_3 \cdot CaCO_3 \cdot 2H_2O$
Gaylussite	$Na_2CO_3 \cdot CaCO_3 \cdot 5H_2O$
Sulfates	
Anhydrite	$CaSO_4$
Gypsum	$CaSO_4 \cdot 2H_2O$
Glauberite	$CaSO_4 \cdot Na_2SO_4$
Thenardite	Na_2SO_4
Mirabilite	$Na_2SO_4 \cdot 10H_2O$
Aphthitalite	$(K,Na)_3Na(SO_4)_2$
Vanthoffite	$3Na_2SO_4 \cdot MgSO_4$
Bloedite	$Na_2SO_4 \cdot MgSO_4 \cdot 4H_2O$
Loeweite	$6Na_2SO_4 \cdot 7MgSO_4 \cdot 15H_2O$
Kieserite	$MgSO_4 \cdot H_2O$
Hexahydrite	$MgSO_4 \cdot 6H_2O$
Epsomite	$MgSO_4 \cdot 7H_2O$
Langbeinite	$K_2SO_4 \cdot 2MgSO_4$
Leonite	$K_2SO_4 \cdot MgSO_4 \cdot 4H_2O$
Picromerite	$K_2SO_4 \cdot MgSO_4 \cdot 6H_2O$
Polyhalite	$K_2SO_4 \cdot MgSO_4 \cdot 2CaSO_4 \cdot 2H_2O$
Celestite	$SrSO_4$
Barite	$BaSO_4$
Chlorides	
Halite	$NaCl$
Sylvite	KCl
Bischofite	$MgCl_2 \cdot 6H_2O$
Carnallite	$KCl \cdot MgCl_2 \cdot 6H_2O$
Kainite	$4KCl \cdot 4MgSO_4 \cdot 11H_2O$
Borates and nitrates	
Colemanite	$Ca_2B_6O_{11} \cdot 5H_2O$
Inyoite	$Ca_2B_6O_{11} \cdot 13H_2O$
Priceite	$Ca_4B_{10}O_{19} \cdot 7H_2O$
Kernite	$Na_2B_4O_7 \cdot 5H_2O$
Borax	$Na_2B_4O_7 \cdot 10H_2O$
Ulexite	$NaCaB_5O_9 \cdot 8H_2O$
Searlesite	$Na_2O \cdot B_2O_3 \cdot 4SiO_2 \cdot 2H_2O$
Boracite	$Mg_3B_7O_{13}Cl$
Niter	KNO_3
Soda niter	$NaNO_3$

*The formulas are written as shown for reasons of convenience and ease of distinguishing the cationic and anionic elements and groups involved. Dolomite, for example, is commonly expressed chemically as $CaMg(CO_3)_2$, rather than the form listed above.

Mineral equilibria. The reasons for this impressive mineral diversity can best be understood in terms of the geochemical relations that exist between evaporate minerals and the mother liquor, or brine, from which they precipitate. Usiglio's work clearly showed that calcite and gypsum are the first minerals to precipitate from seawater in the normal course of evaporation. After reduction to 10 percent of the original volume, sodium chloride begins to precipitate from solution, and magnesium chloride, magnesium sulfate, potassium chloride, and a host of other salts do not precipitate from solution until a volume reduction to about 3 percent is attained. Interpretation of the salts associated with this residual hypersaline brine can best be accomplished by considering the geochemical system: sodium–potassium–magnesium–chloride–sulfate–water. The complexities of such a six-component system are legion, but simplification is possible. Much work has been based on consideration of the three-component system, magnesium–potassium–sulfate, when this system is in equilibrium with a solution that is saturated with respect to sodium chloride (*i.e.*, essentially the conditions that exist when sodium chloride begins to precipitate from seawater).

The magnesium–potassium–sulfate system

The geochemical relations among mineral species in equilibrium with the magnesium–potassium–sulfate system (and the associated sodium–chloride solution) are portrayed on the triangular Jänecke-type diagram in Figure 1. The diagram was devised by the German chemist Ernst Jänecke, who studied the geochemical relations of the Braunschweig salt deposits in the early 1900s. Essentially, the area bounded by the triangle represents all possible compositional variations of the three chemical end-members, or components (magnesium, potassium, and sulfate), that add up to 100 percent. The apex designated K_2^{++} (potassium ion), for example, represents 100 percent potassium and no magnesium or sulfate. If a vertical line were drawn from the K_2^{++} apex to the base of the triangle, the point of intersection would represent the composition 50 percent magnesium, 50 percent sulfate, and no potassium. Similar relations prevail for the other corners and

edges. Thus, each edge of the Jänecke diagram is actually a scale reflecting all possible mixtures of the two pure components at its ends; a point located precisely in the centre represents the tripartite composition: one-third magnesium, one-third potassium, and one-third sulfate.

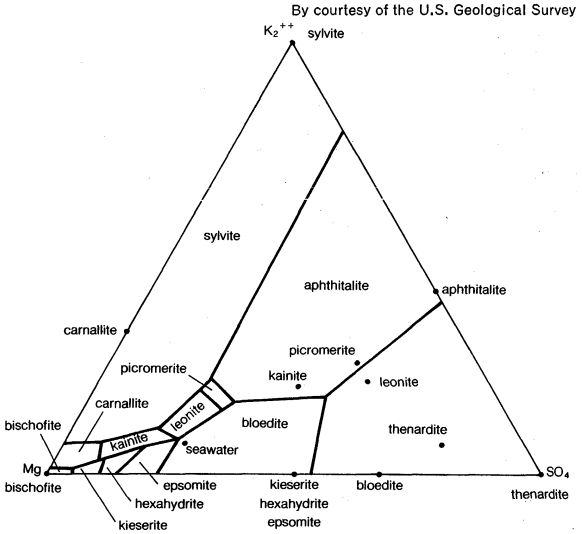


Figure 1: Stability fields of evaporite minerals at 25° C in the system magnesium–potassium–sulfate (see text).

The dots on the diagram indicate mineral compositions, and sylvite (KCl; no magnesium or sulfate), thenardite (Na_2SO_4; no magnesium or potassium), and bischofite ($MgCl_2 \cdot 6H_2O$; no potassium or sulfate) are, accordingly, the mineral end-members of this system; they are plotted at the corners of the triangle. Minerals such as kainite ($4KCl \cdot 4MgSO_4 \cdot 11H_2O$), which contain some of each of the components, are plotted in the triangle interior as dictated by their compositions.

Temperature dependence of stability fields

Finally, it is important to note that the areas given to a particular mineral species cover a much wider range of compositions than those indicated by the plotted dots on the diagram. Thenardite is plotted in the sulfate corner because, as noted above, it contains no potassium or magnesium. The large area labelled thenardite in the lower right part of the diagram refers to the stability field of this sodium sulfate at 25° C (77° F). At any solution composition that falls within this area, thenardite will begin to precipitate at the stated temperature. The lines that bound the thenardite field represent experimentally determined transitions (thermodynamic equilibria) between thenardite and aphthitalite on the one hand, and thenardite and bloedite on the other. These stability boundaries refer, in a gross compositional sense, to the limiting concentrations of potassium and magnesium that can be associated with the precipitation of thenardite from solution.

When precipitation of any mineral species occurs, the overall brine composition changes because some components will have been removed from solution. Figure 1 shows the compositional location of seawater; it falls within the stability field of bloedite. This means that bloedite, a hydrated sodium and magnesium sulfate, will be the first of the minerals to precipitate upon evaporation of seawater at 25° C (77° F). Removal of some sulfate and sodium from solution shifts the composition of the brine, and, in consequence, the next mineral that will precipitate is epsomite. Continued evaporation at this temperature results in the further mineral succession: kainite, hexahydrite, kieserite, carnallite, and, ultimately, to the precipitation of bischofite. This sequence has been experimentally demonstrated. Many authorities believe, however, that bischofite is formed by diagenesis in evaporite deposits (that is, by alteration of previously formed minerals) because solutions rich in the required magnesium chloride are most uncommon in nature.

The temperature dependence of geochemical reactions involving evaporites and brines is great, and it should be emphasized that the minerals, stability fields, and evaporative succession cited above hold only at a temperature of 25° C (77° F). The Dutch chemist Jacobus Henricus van't Hoff devoted many years of research to a systematic examination of mineral equilibria related to the Stassfurt salt deposits in Germany. He might well be called the father of experimental study of evaporites, and among his many fundamental discoveries was the change in mineral assemblage to be anticipated with increasing temperature. The interval from 25° to 37° C (77° to 99° F) was characterized by van't Hoff as one marked by the disappearance of picromerite, epsomite, and hexahydrite—each of which appears on the stability fields (25° C) shown in Figure 1. The temperature interval from 37° to 55° C (99° to 131° F) is associated with the appearance of langbeinite, loeweite, and vanthoffite; and a third temperature interval, from 55° to 83° C (131° to 181° F), is one in which bloedite, leonite, and kainite disappear.

Typical dehydration reactions

Some typical dehydration (water loss) reactions that indicate both the temperature dependence of evaporite assemblages and the cause of much mineral speciation include the following:

$$\underset{\text{mirabilite}}{Na_2SO_4 \cdot 10H_2O} \rightleftharpoons \underset{\text{thenardite}}{Na_2SO_4} + \underset{\text{water}}{10H_2O},$$

which occurs at 18° C (65° F) in the presence of halite;

$$\underset{\text{picromerite}}{K_2SO_4 \cdot MgSO_4 \cdot 6H_2O} \rightleftharpoons \underset{\text{leonite}}{K_2SO_4 \cdot 2MgSO_4 \cdot 4H_2O} + \underset{\text{water}}{2H_2O},$$

which occurs at 26° C in the presence of halite, bloedite, and aphthitalite;

$$\underset{\text{hexahydrite}}{MgSO_4 \cdot 6H_2O} \rightleftharpoons \underset{\text{kieserite}}{MgSO_4 \cdot H_2O} + \underset{\text{water}}{5H_2O},$$

which occurs at 35.5° C in the presence of halite and bloedite;

$$\underset{\text{bloedite}}{Na_2SO_4 \cdot MgSO_4 \cdot 4H_2O} + \underset{\text{thenardite}}{2Na_2SO_4} \rightleftharpoons \underset{\text{vanthoffite}}{3Na_2SO_4 \cdot MgSO_4} + \underset{\text{water}}{4H_2O},$$

which occurs at 46° C (115° F) in the presence of halite and aphthitalite; and

$$\underset{\text{kainite}}{4KCl \cdot 4MgSO_4 \cdot 11H_2O} \rightleftharpoons \underset{\text{kieserite}}{4MgSO_4 \cdot H_2O} + \underset{\text{sylvite}}{4KCl} + \underset{\text{water}}{7H_2O},$$

which occurs at 83° C (181° F) in the presence of halite.

Each of the above reactions occurs only at the temperatures stated and in the presence of the minerals indicated; being reversible dehydration reactions, they are characterized principally by the loss (or gain) of water molecules as shown.

Jänecke diagrams for temperatures as high as 350° C (650° F) have been prepared. They are similar to the diagram of Figure 1, but the minerals present and the area of their stability fields vary because of the temperature dependence noted above. At 110° C (230° F), for example, the stability field of thenardite is increased and occupies much of the right side of the triangle. The other principal mineral species present at 110° C are aphthitalite, sylvite, langbeinite, kieserite, and vanthoffite; small areas represent the stability fields of loeweite, carnallite, and bischofite. Several of these mineral species are, of course, absent from the 25° C diagram, and seawater is in equilibrium with vanthoffite instead of bloedite at the elevated temperature of 110° C.

Knowledge of these mineral relations and equilibria is of great aid in efforts to understand and interpret the origin of evaporites. The presence of certain high temperature species in a particular deposit is obviously of paleo-environmental significance. Similarly, the presence of some of the rarer potassium or magnesium chlorides, for example, is of significance because it indicates either (1) the existence of environmental opportunity for evaporation to proceed nearly to completion or (2) the occurrence of post-depositional alteration of an original evap-

orite assemblage, perhaps by percolation of brines within a groundwater circulation system. In either case, the history of deposition is clarified.

Mineral characteristics. The physical properties of the several mineral groups involved are treated elsewhere; see, for example, the article CARBONATE MINERALS. In general, the most distinctive physical characteristic of evaporite minerals and mineral sequences (aside from salt or bitter taste) is colour. A wide variety of colours has been noted, and some are rather readily explained in terms of inclusions: carbonaceous matter lends black and brown tones; finely disseminated clays produce shades of gray; iron oxides such as hematite are associated with reds and reddish browns; and the presence of small amounts of iron chloride is thought to be the cause of some yellows.

Colour in evaporites

Many investigations of colour in evaporites have been directed toward explanation of blue, violet, and yellow colours in halite, violet in anhydrite and langbeinite, and other evaporite colours that are known or thought to be produced by inclusion of radioactive trace elements. Uranium, thorium, and potassium-40 are among the suggested causes, and a relationship between mineral colour and helium content has been demonstrated in some cases. The helium may be derived from brine inclusions that were trapped when the minerals were first formed. In other instances, however, minerals with equally high helium contents are uncoloured. Hence, colour in evaporite minerals—indeed, in minerals in general—remains somewhat enigmatic. It is a source of intrinsic beauty as well as speculation, however, because the brilliant white of many evaporite beds contrasts markedly with the coloured deposits of some of the minerals mentioned here.

ORIGIN OF EVAPORITES

The origin of evaporites is rather simple in the sense that all deposits are derived by evaporation of water bodies. Beyond this, complexities arise because the variability of conditions in nature can give rise to the great mineral diversity already noted, and because a number of possible environments of deposition are involved. The purposes of this article are not compatible with extensive discussion of the technical arguments on evaporite origin. The topic will here be treated in terms of the principal environments of deposition that are known to exist: namely, the nonmarine closed-basin or playa environment; the shallow-water tidal flat and coastal lagoon; and the deep-basin environment with restricting sill or bar. An alternate hypothesis treating evaporite deposition from layered solutions will be appended in order to indicate some trends of modern thought on the question of origin.

The nonmarine environment. Evaporite deposition in the nonmarine environment occurs in closed lakes—those without outlet—in arid or semi-arid regions. Such lakes form in closed interior basins or shallow depressions on land where drainage is internal and runoff does not reach the sea. If water depths are shallow or, more typically, somewhat ephemeral, the term playa or playa lake is commonly used.

Hydrology of closed lakes

Hydrologic inflow to closed lakes consists principally of precipitation and surface runoff, both of which are small in amount and variable in occurrence in arid regions. Groundwater flow and discharge from springs may also provide water input in some cases, but evaporation rates are high and are always considerably in excess of precipitation and surface runoff. When sporadic or seasonal storms occur, they give rise to a sudden surge in water inflow. Because the closed lake lacks an outlet, it can respond to such circumstances only by increasing its surface area and water depth. Subsequent evaporation will reduce the volume of water present to the prestorm or normal amount; fluctuation of the lake level, therefore, is characteristic of the environment.

Changing lake levels and water volumes give rise in turn to fluctuating salinity values. Salinity is a measure of the total salt content, commonly expressed in units of parts per thousand and represented by the symbol ‰. The values reported for some closed lakes are as high as those existing anywhere on Earth; some examples are: Lake Assal, Afars and Issas, formerly French Somaliland (325 ‰); Elton Lake, Soviet Union (300 ‰); Tuz Gölü, Turkey (250 ‰); Dead Sea, Israel and Jordan (270 ‰); Deep Springs Lake, California (310 ‰); and Lake Eyre, Australia (240 ‰). By way of contrast, the salinity of normal seawater is about 35 ‰ and that of the most saline seawater does not exceed closed-lake values (salinity in the Kara-Bogaz-Gol Gulf of the Caspian Sea is 310 ‰; salinity in the Atlantis II Deep and the Discovery Deep, both in the Red Sea, is 257 ‰). It must be emphasized, however, that salinites are highly variable in closed lakes. Influx of fresh water from storm runoff has produced temporary salinity values as low as 40 ‰ in Lake Eyre and 22 ‰ in Deep Springs Lake, for example.

The salinity variations described can affect equilibrium relations between brine and minerals and lead to much solution and reprecipitation of evaporites in the nonmarine environment. Nevertheless, there is often a striking resemblance between the mineral zonation exhibited by sediments of playa lakes and that which would occur if the natural brines were allowed to evaporate in the laboratory. Indeed, such zonation can be said to be one of the hallmarks of the nonmarine environment. A particularly fine example is shown in Figure 2, the result of a study of evaporites in Deep Springs Lake, California.

Areal zonation of minerals

The relatively flat basin containing this playa lake receives about 125 millimetres (five inches) of precipitation per year, whereas the mean annual evaporation is approximately 1,250 millimetres (50 inches) per year. Water composition reflects the lithologies (rock types) that exist within the drainage basin, and the saline brine of this particular playa lake consists essentially of sodium sulfo-carbonate, with lesser amounts of calcium, magnesium, potassium, and chloride. The areal variation of evaporites from the playa margin to its centre is as follows: calcite, dolomite, gaylussite, and thenardite, surrounding a central zone of burkeite. The composition of the latter mineral, which characterizes the central salt pan, is

$$Na_6(SO_4)_2CO_3.$$

The areal variation shown in Figure 2 is precisely what would be expected of a sequence of mineral precipitates under the prevailing conditions.

Many of the same minerals are present in the neighbouring Death Valley salt pan, where a more complete carbonate–sulfate–chloride areal variation has been demonstrated. The particular mineral suite present in the nonmarine environment is quite variable, however; water composition always varies to some degree with rock type and composition, and the evaporites present may vary from basin to basin, even within the same general region.

The chloride stage of precipitation in Deep Springs Lake is represented by a vertical sequence of evaporite minerals within salt crusts and polygons. Halite and sylvite overlie thenardite, burkeite, trona, and one or two other minerals at depth, but the overall relationship in vertical layering is not nearly so consistent as is areal variation. This is true of closed lakes generally, because the fluctuating water levels combine with very flat topography (note ground elevations shown in Figure 2) to produce extensive shoreline changes. Such changes may have caused considerable variation in brine composition over any fixed point through time, and vertical sequences are, of course, always representative of these variations or changing environmental conditions.

The shallow marine environment. Evaporite deposition in the shallow marine environment (sometimes termed the salina) occurs in desert coastal areas, particularly along the margins of such semi-restricted water bodies as the Red Sea, Persian Gulf, and Gulf of California. Restriction is, in general, one of the critical requirements for evaporite deposition because free and unlimited mixing with the open sea would easily overcome the high evaporation rates of arid areas and dilute these waters to near-normal salinity. The semi-restriction referred to cannot, in fact, prevent a great deal of dilution by mixing; coastal physiography is the principal factor involved in brine production.

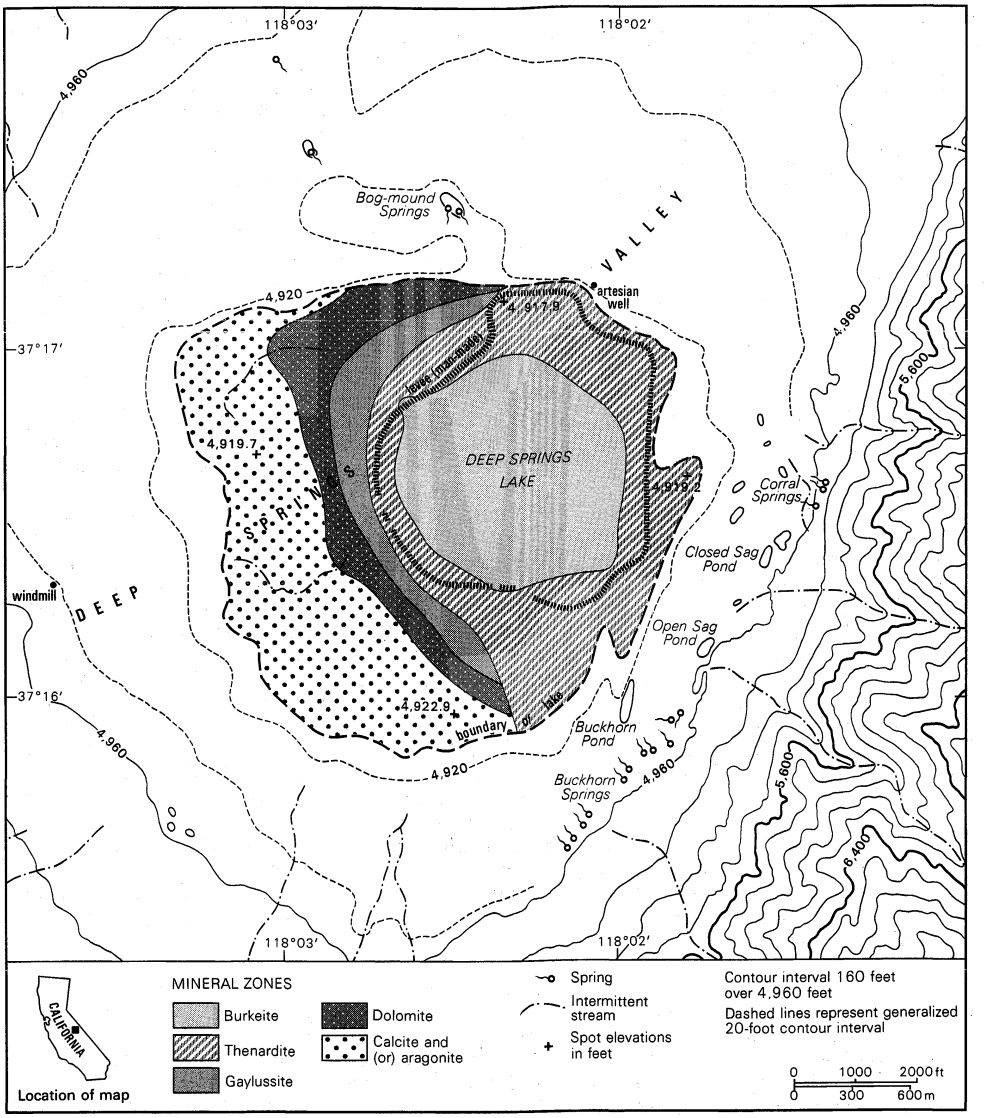


Figure 2: Deep Springs Lake, California, showing the general mineral zonation in the surficial deposits of the playa.
By courtesy of the U.S. Geological Survey

Evaporite formation along the coast of the United Arab Emirates

The United Arab Emirates (formerly the Trucial Coast) of the Arabian Peninsula provide an excellent and much-studied example of this environment. A fringing network of offshore islands between Qatar and Abu Dhabi serves as a barrier for coastal lagoons, within which the waters of the Persian Gulf are further restricted. Salinity values are about 70 ‰, or twice that of normal seawater. One of the largest of the lagoons off the United Arab Emirates is Khawr al-Bazm, and the lagoon and its associated islands constitute the outermost of three distinct but contiguous elements of the shallow marine environment (Figure 3). The second element is the algal flat, an intertidal zone in the vicinity of the high-water mark, and the third element is the coastal salt flat, or sabkha, a supratidal zone that is inundated by seawater only occasionally.

The lagoonal sediments are largely carbonate sands and muds, rich in coral fragments and other bioclastic grains. It is thought that blue-green algae play an important role in the alteration of the carbonate sediments and in inducing precipitation of aragonite rather than calcite in many instances. They are also instrumental in creating the algal-mat area; the mats trap such carbonate sediment and tend to be preserved as distinctive strata within vertical sedimentary sequences. Evaporite formation is largely restricted to the coastal sabkha, however.

The sabkha along the coast attains widths of eight to ten kilometres (five to six miles), and temperature and salinity values more nearly resemble those of continental playas than those of the adjacent lagoon, Khawr al-Bazm. Evaporation of groundwater influx and occasional storm-generated floodwaters has produced an interstitial brine (*i.e.*, within the pore space of the sabkha sediments) about one metre below the sabkha surface. The salinity of this brine may be as great as 350 ‰ in places, and evaporite precipitation, largely in the form of gypsum and anhydrite, occurs within the interstices of the sabkha sediments (Figure 3). Alteration and replacement (diagenesis) of the original sediments also occur by reaction with the interstitial brine. Dolomite, produced in a variety of ways, is fairly common, and halite occurs in the salt crusts of the sabkha surface. Other evaporites present include celestite and huntite; the latter is a relatively rare carbonate of calcium and magnesium.

Shoreline regression and mineral variation

Several kinds of vertical mineral variations occur in this environment, but an areal variation is particularly worthy of note. Data from Baja California and the Persian Gulf suggest that shoreline regression is occurring at the rate of one to two metres per year, which means that a sabkha as wide as 100 to 200 kilometres (60 to 120 miles) could be formed in as little as 100,000 years—a short time span in the geological sense. Whether this occurs without interruption depends upon a number of factors; however, with regard to the areal variation of minerals, gypsum belts will move progressively seaward through time, and the width of belts of anhydrite—which

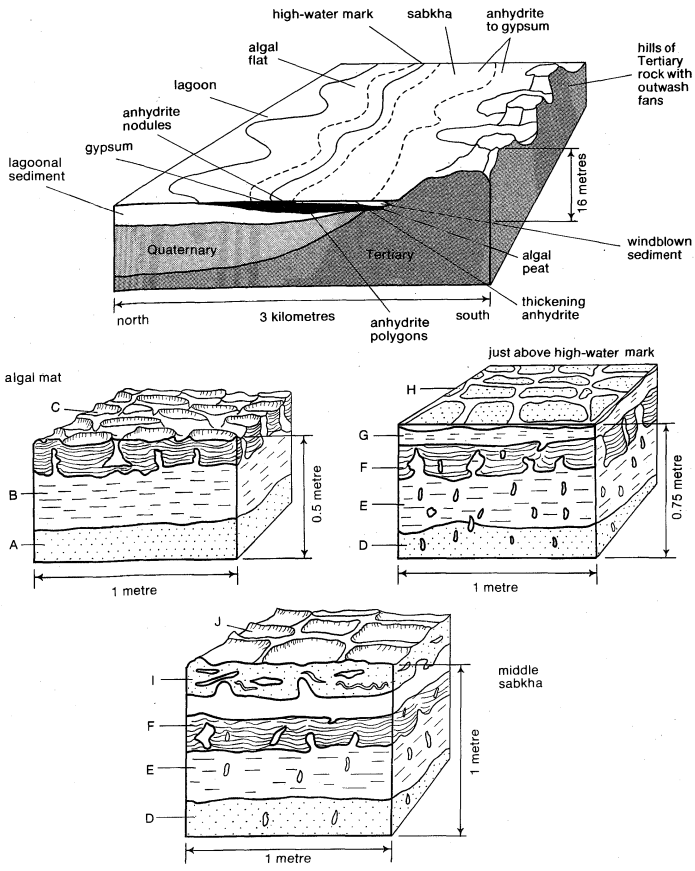


Figure 3: *Distribution of evaporites in supratidal flat lining Khawr al-Bazm, Persian Gulf.*
(A) Lagoonal carbonate sands or muds, or both. (B) Poorly laminated carbonate-rich algal peat. (C) Algal mat formed into polygons. (D) Lagoonal sediments with gypsum crystals. (E, F) Algal peat with large gypsum crystals. (G) Mush of gypsum and carbonate. (H) Anhydrite polygons and windblown carbonate and quartz. (I) Anhydrite layer replacing gypsum mush and forming diapiric structures capped by anhydrite layers and nodules in windblown carbonate and quartz. (J) Halite crust formed into compressional polygons.
From Kendall and Skipwith, *American Association of Petroleum Geologists Bulletin* (April 1969)

forms by dehydration of gypsum under certain conditions—will increase. The sabkha mineral belts always tend to parallel the associated coastal lagoons. When these evaporites are found in the geological record, they provide evidence of the former location of land and sea and of the trends of former shorelines.

The deep-basin environment. There is no doubt that some evaporites in the geological record were deposited under shallow-water conditions that closely resembled those of the modern Khawr al-Bazm. The great thickness of evaporites in some ancient deposits (*e.g.*, approximately 1,200 metres [3,900 feet] thick in the German Zechstein deposits) poses difficulty for the salina model, however. The basic problem is that rates of land subsidence in areas of salt deposition would necessarily have to equal known rates of evaporite accumulation. The required subsidence seems geologically unreasonable in many instances, and a deepwater model provides a more plausible alternative in such cases.

The deep-basin model from the initial to permanent stages

As in all instances of evaporite formation—including nonmarine formation—evaporation must generally exceed precipitation and surface runoff in the deep-basin model. The principal physiographic requirement is the presence of a barrier sill to prevent unlimited mixing of water within the basin with that of the open sea. This physical configuration is quite similar to that of the Mediterranean Sea, most of the Norwegian fjords, and the Red Sea proper.

Figure 4 illustrates the probable sequence of events. In the formative stage (Figure 4A), incoming seawater of density 1.02 grams per cubic centimetre (subsequently abbreviated here as g/cm³) gives rise to water of density 1.025 g/cm³ within the basin because of surface evaporation. Toward the landward margin of the basin, the density further increases, and the circulation caused by the sinking of denser water is as shown. Conditions in the Mediterranean Sea today closely resemble those depicted here for the formative stage.

The continual sinking of dense brines formed at the surface ultimately displaces all of the waters of normal or near-normal density within the basin, and stagnant bottom water, lacking oxygen, is formed. This is the euxinic stage (Figure 4B), in which normal carbonate deposition ceases and is replaced by the accumulation of dark, sulfide-rich, organic muds. Bottom-water densities attain values of about 1.08 g/cm³ and marine life generally perishes.

Continued surface evaporation will lead to precipitation of the first evaporite mineral, gypsum, when densities of about 1.10 g/cm³ are attained in the distal parts of the basin. This is termed the ephemeral stage (Figure 4C) because the gypsum crystals will redissolve as they sink toward the bottom through water of lesser density (1.08 g/cm³) that is not yet fully saturated with respect to calcium sulfate.

The permanent evaporite stage is finally reached when very dense water fills the entire basin to the depth of the sill (Figure 4D). At the density shown (1.25 g/cm³), gypsum, anhydrite, and halite could accumulate on the basin bottom without the occurrence of re-solution. Depending upon the salinity (density) variations that occur through time, a variety of evaporite minerals and mineral sequences could be deposited under the conditions here associated with the permanent stage of development.

Preservation of deposits

Evaporite deposition in the deep-basin environment is therefore limited chiefly by basin depth; the great thicknesses of many known deposits can be accounted for by this model. With regard to the ultimate fate and degree of preservation of deep-basin evaporites after basin filling, several alternatives are possible. Figure 4E shows one such alternative. The deep basin has been completely filled and salt flat or salina-like conditions prevail at the surface. Normal sedimentation, associated with shoreline shifts of the shallow marine environment, will preserve the thick evaporite deposit.

If sea level should fall or if the basin and its sill should rise because of uplift of the Earth's crust, the influx of seawater will be effectively cut off and the deep-basin evaporites easily preserved (Figure 4F). The thick salt deposits of the Ethiopian Danakil Depression may reflect this postulated condition.

Finally, climatic change may interrupt basin filling during the permanent stage. Even in the instance of much freshwater inflow (Figure 4G), the combined effect of the immediately overlying saline brine and the basinward transport of land-derived sediments will be to prevent re-solution and thus preserve the evaporites.

Deposition from layered solutions. It has recently been argued that some evaporites in the geological record accumulate at times and places for which there is no independent evidence of persistently arid climates. Because density stratification is known to exist in some form in all lakes and seas, a simple two-layer aqueous system has been proposed as fundamental to evaporite formation. Basically, the system consists of a dense layer of brine overlain by a relatively dilute water layer and separated from it by a pycnocline—*i.e.*, a boundary across which diffusion of salt is slow. The existence of pycnoclines in nature has been demonstrated; in fact, a permanent pycnocline was established in the Dead Sea about 1,500 years ago.

Laboratory work has demonstrated that evaporite deposition can occur in a two-layer system. This is not a drastically different concept, and several of the stages depicted in Figure 4 are quite similar to it, involving as they do water layers of lesser density and salinity that overlie denser brines. Two aspects of the two-layer system should be noted here, however. First, it is postulated that the brine layer below the pycnocline is sustained and

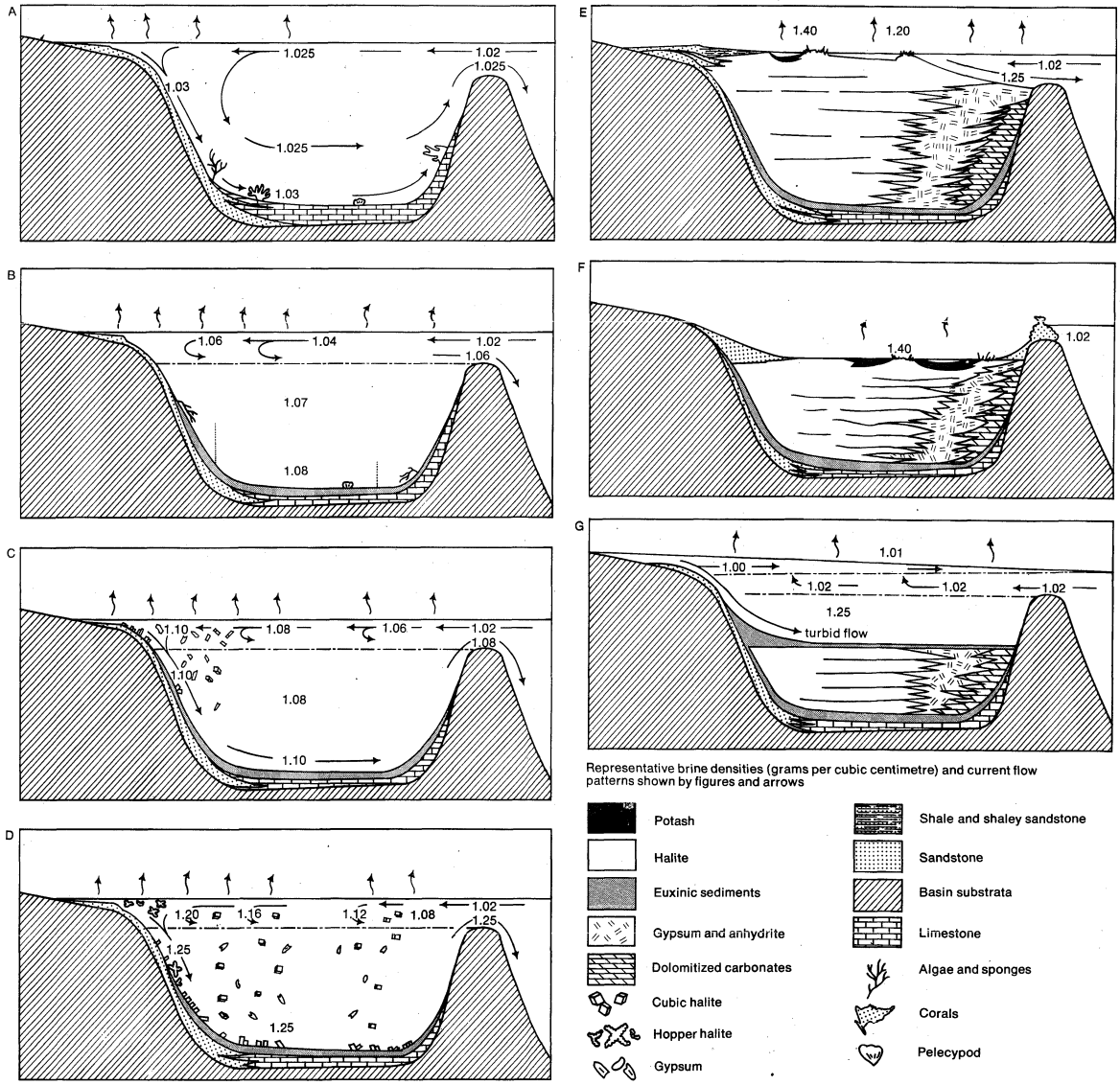


Figure 4: *Model of deepwater evaporite deposition.*
(A) Formative stage. (B) Euxinic stage. (C) Ephemeral stage. (D) Permanent evaporite stage. (E–G) Alternative terminal stages (see text).
From R. Schmalz, "Deep Water Evaporite Deposition: A Genetic Model" (1969); American Association of Petroleum Geologists

replenished by the inflow of saline water derived from the entire drainage area. That is, evaporation over the immediate site of evaporite deposition need not be continuous; local or seasonal changes of climate can be accommodated without effect on the brine layer. Second, inclusion of the entire drainage area reduces the vast quantity of water per unit area that is customarily thought to be evaporated during evaporite deposition.

Brine exposure by seiches

The most tortuous facet of the two-layer hypothesis is a requirement that the brine below the pycnocline be intermittently exposed and subjected to surface evaporation to induce evaporite precipitation. Proponents cite the possibility of surface winds and internal seiches of sufficient amplitude and duration to "pile up" the dilute surface waters, expose the brine to evaporation for periods as long as weeks, and thus permit precipitation of evaporites over the entire basin surface. Internal seiches of Lake Baikal have a period of 30 to 40 days and an amplitude of 150 metres (500 feet) or more; this observation certainly strengthens the plausibility of the two-layer hypothesis.

In conclusion, it can only be added that each of the environments of deposition discussed in this article exists today and that each has been responsible for the formation of some evaporites in the geological past. Evaporites have originated in marine and nonmarine environments, in shallow water and in deep basins, and under arid and somewhat fluctuating climatic conditions. One of the principal virtues of the two-layer model is that it indicates a mechanism whereby the latter circumstances can be accomplished.

DISTRIBUTION IN SPACE AND TIME

Evaporites occur in strata of every geological age, from Cambrian (500,000,000 to 570,000,000 years ago) to Holocene time (the last 10,000 years, approximately). Space limitations preclude description of the mineralogy and stratigraphy of individual deposists; treatment here will be confined to the gross aspects of evaporite distribution and the implications that have been drawn from this distribution.

Evaporites in the geological record

Some of the oldest deposits occur in the Canadian Arctic Archipelago and on the Siberian Platform; Cambrian evaporites in the region of the upper Lena River constitute one of the world's greatest accumulations. This high-latitude occurrence in the Northern Hemisphere persisted through Devonian time (345,000,000 to 395,000,000 years ago), when evaporites were deposited in an enormous zone between northern Greenland and the Rocky Mountains. Devonian evaporites in Alberta and British Columbia are associated with coral reefs and oil and gas fields. Evaporites also were deposited in the northeastern

and north central United States, the Gulf of Mexico region, and in Tethys, the ancient east–west seaway between Europe and Africa that extended across southern Asia, during this general time span of more than 200,000,000 years.

Deposits of Carboniferous age (280,000,000 to 345,000,000 years ago) are not as abundant in the highest northern latitudes, and some of the principal occurrences are in the maritime areas of eastern Canada, and in England and Ireland. Carboniferous evaporites also occur in Brazil; this marks the earliest time of deposition in the Southern Hemisphere.

Evaporite deposition was at a maximum during the Permian Period (225,000,000 to 280,000,000 years ago), and some of the world's most famous deposits are of Permian age. The Arctic Archipelago again became a site of evaporite formation, and deposition occurred in eastern Greenland, Spitsbergen, over much of western Europe, and in the Ural Mountains. The upper Kama Basin in the Urals is thought to contain potash reserves that are several times greater than those in the German Stassfurt evaporites. Other noteworthy Permian deposits occur in the southwestern United States, western Canada, and in the Southern Hemisphere, where evaporites in Australia, South America, and Africa formed at latitudes as far from the Equator as 25° S.

During the Triassic Period (190,000,000 to 225,000,000 years ago) the principal sites of deposition were in western Europe and North Africa, in a broad belt extending approximately from Senegal to Libya. Evaporites of Triassic age also occur in western North America, in east Africa, and in scattered localities in the vicinity of the Persian Gulf, southern Asia, and South America. Some deposits formed in eastern Greenland, but no evaporites were formed in latitudes above 55° N after Triassic time.

Evaporite deposition during Jurassic (136,000,000 to 190,000,000 years ago) and Cretaceous (65,000,000 to 136,000,000 years ago) time was relatively restricted. Occurrences of note include a broad swath of deposits around the Gulf of Mexico, over much of the Arabian Peninsula, and in Argentina and Bolivia. During the Tertiary Period (2,500,000 to 65,000,000 years ago) the areas of evaporite deposition were further restricted. Evaporites formed principally in the Mediterranean region, the Red Sea and Persian Gulf, and in the Gulf of California and the Gulf of Mexico off the Yucatán Peninsula. The glacial conditions of the Pleistocene Epoch eliminated these depositional sites, which have been partly restored under Holocene or modern climates.

Climatic change and continental drift

It seems clear that evaporites were deposited over a much wider latitudinal range in the past, and that evaporite maximums have varied inversely with known glacial maximums through geological time. Combined with the fact that modern limits of evaporite formation are essentially determined by climate, most authorities agree that the distribution of evaporites through time closely reflects worldwide climatic change, specifically coinciding with periods of warmth and aridity. It might be argued that evaporite deposition from the two-layer system described earlier obviates the climatic link; however, Arctic evaporites are associated with coral reefs and plant remains that are independent indicators of much warmer conditions at high northern latitudes.

Because the general zone of evaporite deposition has remained symmetrical about the present rotational axis of the Earth (during the several latitudinal expansions and contractions that have occurred), this poses some constraints on the theory of continental drift. The movement of continents in an east–west direction (*e.g.*, the separation of Europe and North America by sea-floor spreading in the North Atlantic) can be accommodated, but the data on evaporite distribution would appear to conflict with north–south motions. In addition, the relatively sparse occurrence of evaporites in the Southern Hemisphere is most readily explained in terms of the present distribution of land and sea—that is, by the essential oceanic character of the Southern Hemisphere. This is incompatible with the postulated existence of a former great southern landmass that included Antarctica.

Most earth scientists are firm in the belief that continental drift and sea-floor spreading are realities rather than hypotheses, and, indeed, the paleomagnetic and other data from the ocean floors appear to be most convincing. There are, however, a few kinds of geological data that can be reconciled with predrift reconstructions of the Earth's crust only with difficulty. The distribution of evaporites in space and time must be considered in this category, thus providing a fascinating problem of Earth history.

BIBLIOGRAPHY. There are no books in English that treat all aspects of evaporites, but H. BORCHERT, *Ozeane Salzlagerstätten* (1959; Eng. trans. with new material by R.O. MUIR, *Salt Deposits*, 1964), presents a thorough description of mineral stability relations; and J. BRAUNSTEIN and G.D. O'BRIEN (eds.), *Diapirism and Diapirs* (1968), covers salt deposits in all parts of the world that are related to salt domes and piercement structures. Several elementary texts on geochemistry provide abbreviated but basic treatment from the chemical point of view. In this vein are E.T. DEGENS, *Geochemistry of Sediments* (1965); K.B. KRAUSKOPF, *Introduction to Geochemistry* (1967); and B.H. MASON, *Principles of Geochemistry*, 3rd ed. (1966). The relation of evaporites to other chemically precipitated rocks and to the geological record is covered in F.J. PETTIJOHN, *Sedimentary Rocks*, 2nd ed. (1957).

The interested reader must pursue the more specialized technical literature to gain further insight to the subject, however. Basic articles, each of which provides additional references within its bibliography, include: F.H. STEWART, "Marine Evaporites," *Prof. Pap. U.S. Geol. Surv.* 440-Y (1963); H.P. EUGSTER and G.I. SMITH, "Mineral Equilibria in Searles Lake Evaporites, California," *J. Petrology*, 6:473–522 (1965); O. BRAITSCH, "The Temperature of Evaporite Formation," in A.E.M. NAIRN (ed.), *Problems in Paleoclimatology* (1964); B.F. JONES, "The Hydrology and Mineralogy of Deep Springs Lake, Inyo County, California," *Prof. Pap. U.S. Geol. Surv.* 502-A (1965); C.B. HUNT *et al.*, "Hydrologic Basin, Death Valley, California," *Prof. Pap. U.S. Geol. Surv.* 494-B (1966); R. GREEN, "Palaeoclimatic Significance of Evaporites," in A.E.M. NAIRN (ed.), *Descriptive Palaeoclimatology* (1961); and A. LERMAN, "Model of Chemical Evolution of a Chloride Lake: The Dead Sea," *Geochim. Cosmochim. Acta*, 31:2309–2330 (1967). Worthy of special consideration is the entire issue of the *American Association of Petroleum Geologists Bulletin* of April 1969, which is devoted to the theme of evaporites and petroleum, and which covers the several concepts of evaporite formation in a variety of environmental settings; H.P. EUGSTER, "The Beginnings of Experimental Petrology," *Science*, 173:481–489 (1971), which presents a detailed description of the experimental work of van't Hoff on evaporite equilibria; and A.A. MEYERHOFF, "Continental Drift," 2 pt., *J. Geol.*, 78:1–51, 406–444 (1970), which include treatment of evaporite distribution in the geological record.

(L.K.L.)

Everest, Mount

Mount Everest, a peak on the crest of the Himalayan range in Asia, is the highest mountain on earth. It rises above the Himalayas, on the undemarcated Nepal–Tibet border, in the approximate position of 28° N, 87° E. Three barren ridges—the Southeast, Northeast, and West—culminate in two summits at 29,028 feet (8,848 metres; Everest) and 28,700 feet (8,748 metres; South Peak). The mountain can be seen directly from its northeastern side, where it rises about 12,000 feet above the Tibetan Plateau. The lesser peaks of Changtse (north; 24,780 feet), Khumbutse (northwest; 21,785 feet), Nuptse (southwest; 25,726 feet), and Lhotse (south; 27,923 feet), which rise around its base, hide the summit from Nepal (for details of the range of which it forms a part, see HIMALAYAN MOUNTAIN RANGES).

Formation of the Great Himalayan mountain ranges began during the Miocene Epoch (7,000,000 to 26,000,000 years ago), as a result of the shock produced by the convergence of the Indian subcontinent and the Tibetan Plateau. The Kāthmāndu and Khumbu nappes (faulted, overturned rock folds) were squeezed upward and folded upon each other in successive stages, creating a primitive mountain chain. A general uplifting of the northern land mass added further elevation. The whole was again overlaid by a further folding of the nappes, which resulted in the formation of Mt. Everest during the Mahabarat Phase of the Pleistocene Epoch (which lasted from about

10,000 to about 2,500,000 years ago). Its layers of limestone from the Permo-Carboniferous Period (which lasted from 225,000,000 to 345,000,000 years ago), separated by other paracrystalline sediments (sediments that were later metamorphosed), were laid down in synclinal stratification (*i.e.*, folded together with strata dipping inward). Formative forces are still evident in a continuing uplifting, which is counteracted by erosion.

The summit of Everest reaches up two-thirds of the way through the earth's atmosphere to the upper layers, where oxygen is thin. Lack of oxygen, powerful winds, and extremely cold temperatures preclude the development of any plant or animal life on the upper slopes. Precipitation falls, during the summer monsoon only, as snow, which lies in powdery drifts. Because these drifts are above the evaporation line, the large basins of névé (partially compacted granular snow), which usually feed glaciers, do not form. Everest's glaciers are consequently fed only by frequent avalanches. The ice sheets on the flanks of the mountain, separated from each other by the main ridges, cover the slopes down to the mountain's base but are slowly retreating because of gradual climatic change. The summit itself is relatively free of snow cover in the winter months, due to the force of northwest gale winds.

Individual glaciers are the Kangshung Glacier (east); the East, main, and West Rongbuk glaciers (north and northwest); the Pumori Glacier (northwest); the Khumbu Glacier (west and south); and the West Cwm, an enclosed valley of ice between Everest and the Lhotse-Nuptse Ridge.

The mountain's drainage pattern radiates southwest, north, and east. The Khumbu Glacier melts into the Lobujya Khola of Nepal, a river flowing southward as the Imja Khola to its confluence with the Dudh Kosi River. Two Tibetan rivers—the Rong Chu and Karma Chu—originate on Everest's slopes as the Pumori and Rongbuk glaciers and as the Kangshung Glacier, respectively. The Rong Chu and Dudh Kosi river valleys form the northern and southern access routes to the summit.

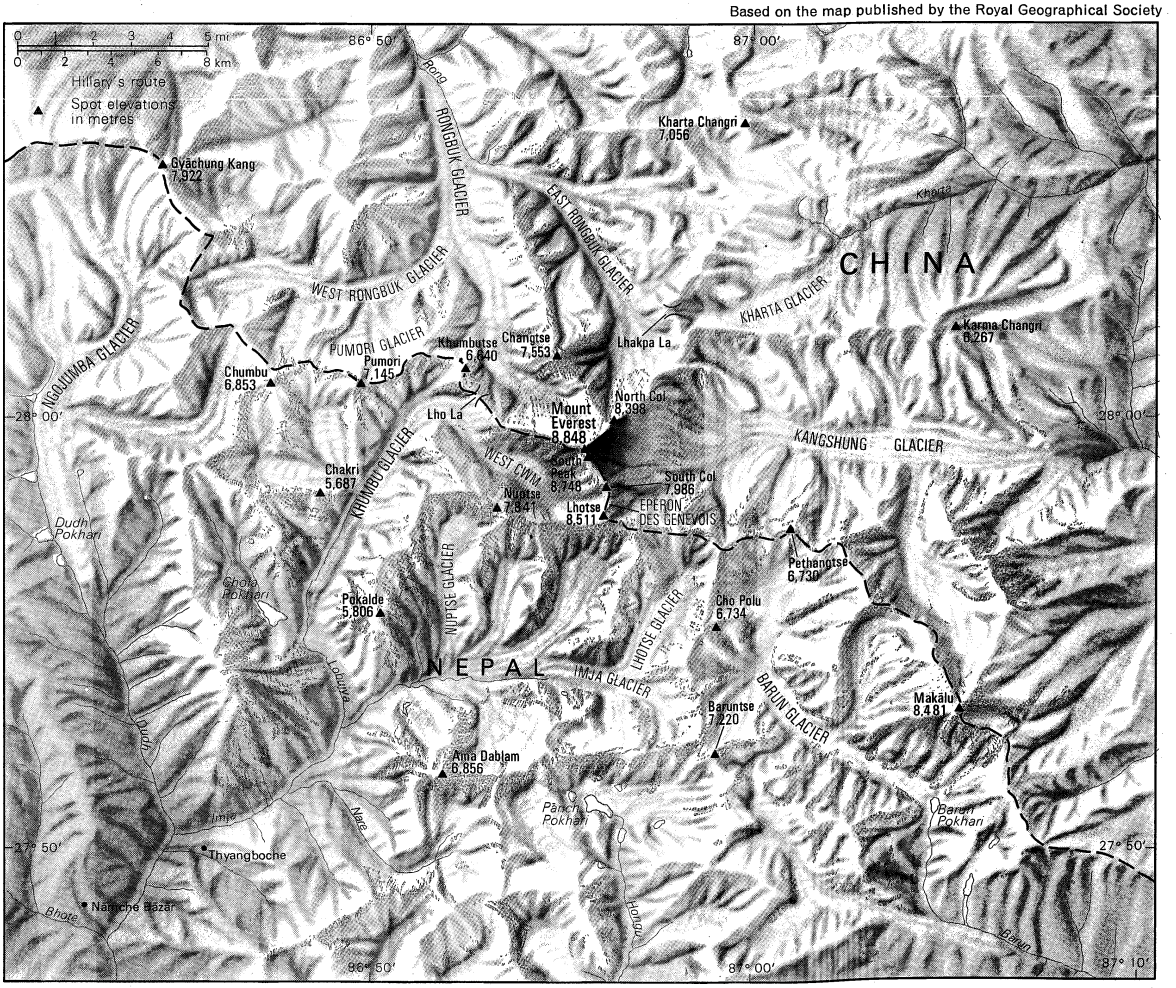
Peoples of the Everest area

Although Mt. Everest itself is void of life, the associated Khumbu Valley is inhabited by the Sherpas, a group of the Bhutia peoples of the Himalayas. Their villages are situated at 12,000 to 13,000 feet and are scattered from the Tibetan border, to the east, to the banks of the Bhote Khose River, to the west. An agricultural people with little cultivable land at their disposal, the Sherpas follow a semi-nomadic life style in their search for open pasturage. Summer grazing is sought as high as 16,000 feet, while winter refuge is taken at lower altitudes on sheltered ledges or on riverbanks. The main crops grown are potatoes and various grains; the yak is raised for its milk, meat, and fur. Many of the men supplement their income by engaging in the transmontane traffic between Nepal and Tibet.

Living in close proximity to the world's highest mountains, the Sherpas traditionally treated the Himalayas with respect—building Buddhist monasteries at their base, placing prayer flags on the slopes, and establishing sanctuaries for the wildlife of the valleys, such as musk deer, monal pheasant, and Himalayan partridge. Spirits and demons were said to live on the high peaks, and the Yeti (the so-called Abominable Snowman) was reputed to roam the lower slopes. Since 1951, when scientific mountain climbing expeditions began to arrive, Sherpas have often taken up employment as porters and have shown a bent for mountaineering.

Because of its commanding size and height, Everest was known in the Tibetan language as Chomolungma, "Goddess Mother of the World." Its identity as the highest point on the earth's surface was not recognized, however, until 1852 when the governmental Survey of India estab-

Based on the map published by the Royal Geographical Society



Mount Everest region.

lished that fact. In 1865 the mountain—previously referred to as Peak XV—was renamed for a British official, Sir George Everest, who was surveyor general of India from 1823 to 1843. Controversy over the exact altitude of the summit developed because of variations due to snow level, gravity deviation, and light refraction. The figure currently accepted—29,028 feet (plus or minus a fraction)—was established by the Survey of India between 1952 and 1955.

Attempts to climb Everest began with the opening of the Tibetan route in 1920. Seven successive attempts on the Northeast Ridge (1921–38), as well as three missions up the Southeast Ridge (1951–52), failed because of the combined difficulties caused by cold arid air, high winds, difficult terrain, and high altitude.

Everest was finally surmounted in 1953, as the result of efforts by an expedition sponsored by the Royal Geographical Society and the Joint Himalayan Committee of the Alpine Club. Open- and closed-circuit oxygen systems, especially insulated boots and clothing, and portable radio equipment were used by the climbers. Eight camps were established on the route that was taken up the Khumbu Icefall and Glacier, the West Cwm, and the face of Lhotse to the South Col, a rocky ridge at 26,200 feet. From there, on May 29, 1953, Edmund (later Sir Edmund) Hillary of New Zealand and Tenzing Norgay of Nepal ascended the Southeast Ridge, past South Peak, to the summit. Since then, four expeditions sponsored by different nations—Swiss (1956), American (1963), Indian (1965), and Japanese (1970)—have scaled the peak. The Japanese party was the first to include a woman.

In 1965 the government of Nepal closed the route to Everest; the ban was, however, lifted in 1968.

BIBLIOGRAPHY. J. HUNT and E. HILLARY, "Triumph on Everest," *Natn. Geogr. Mag.*, 106:1–63 (1954); G.O. DYHENFURTH, *Zum dritter Pol: Die Achttausender der Erde* (1952; Eng. trans., *To the Third Pole: The History of the High Himalaya*, 1955); T.F. HORNBEIN, *Everest: The West Ridge* (1966); T. HAGEN *et al.*, *Mount Everest* (1959; Eng. trans., 1963); Royal Geographical Society Map, "The Mount Everest Region."

(T.N.)

Everglades

Known to the Indians as Pa-hay-okee, "Grassy Waters," the Everglades is a great marsh encompassing nearly 4,000 square miles (10,000 square kilometres) at the southern tip of Florida. Through it, water moves slowly from the lip of Lake Okeechobee to mangrove swamps bordering the Gulf of Mexico and the Florida Bay on the west. On the east the marsh reaches near the narrow, sandy belt that includes the Greater Miami metropolitan area, while on the south Everglades National Park borders Florida Bay at the bottom of the peninsula.

Location

During the 20th century, men have had considerable success in exploiting the rich soil of the Everglades. The region has been the subject of constant conflict between various interests, however, and the delicate ecological balance of the marsh has been threatened constantly by human attempts to manage or transform its resources.

Natural environment. *Natural cover.* The glades occupy a shallow limestone-floored basin that slopes imperceptibly southward at about one-fifth of a foot per mile. Much of it is covered with saw grass (a sedge, the edges of which are covered with minute sharp teeth), which grows to a height of 10 to 15 feet (3 to 5 metres). Where undisturbed, it creates a surface of monotonous sameness. Open water is sometimes found; and there are numerous low islands, or "hammocks," on which a profusion of trees and shrubs grow, including palms, pines, live oaks, cypresses, saw palmettos, and other tropical species. The organic soils, which have been formed from the decay of lush vegetation, range from discontinuous shallow patches to accumulations of peat and muck 8 to 14 feet thick near Lake Okeechobee. The best soils are deep mucks found in a narrow zone along the lakeshore, where a dense tangle of custard apple, or papaw, and moonflower (a morning glory) once grew.

Climate. The climate of the glades is tropical savanna or subtropical and is influenced strongly by the southeast trade winds. Monthly mean temperatures range from 63° F (17.2° C) to 82° F (27.8° C), though winter frosts occur on rare occasions. Rainfall averages over 55 inches (1,400 millimetres) annually, with 75 percent coming between June and October. In winter and spring the glades can be so dry that water disappears into the underlying porous rock.

Wildlife. The mild climate and usual abundance of water provide an environment for myriad birds, particularly such wading birds as herons, egrets, and ibis, as well as for many species of snakes, turtles, and, best known of all, alligators. The "hammocks" served as shelter for deer, wildcats, panthers, bears, smaller mammals, and numerous reptiles. Drainage in the glades has altered the habitats of many species, extending them, as in the case of deer, but shrinking them for water-dependent creatures. An integral part of the ecosystem is a rich variety of insect life.

The Indian inhabitants. The vast areas of open saw grass were used only as passage for dugout canoes and as hunting and fishing territory. The "hammocks," however, were sometimes settled or used as garden sites. Mounds in the vicinity of Belle Glade indicate long-time occupance there. Calusa peoples penetrated from the Gulf Coast, but little colonization within the glades resulted. Adversely affected by European contacts, the Calusa were largely gone by 1800, although the amalgam of refugee Indian groups that became known as Seminole may have absorbed the last vestiges of the Calusa during the early 19th century.

Adaptation to Everglades living

The Seminole found sanctuary in the swamps and marshes because the white man did not covet the glades at the time, and they adapted in several ways. They developed the "chikee," a dwelling without walls, made of a log framework with a thatched roof over a raised platform, that assured maximum ventilation. They wore few clothes and greased their bodies against insect attack. Compromising between agriculture and a gathering economy, they planted corn, beans, melons, and squashes on patches of higher ground, but gathered nuts, greens, and palmetto hearts. The bulbous roots of the coontie plant were the source of a starchy flour, and hunting and fishing provided much of their sustenance.

Of the 1,000 or more Seminole in Florida in the 1970s, scarcely 200 lived within the Everglades. Some lived along the Tamiami Trail or other highways and sold goods or services to travellers and sportsmen. Several villages were open to tourists for a fee, and some were established for that purpose. A few worked as farmhands, at least part of the year, in the Okeechobee farming districts. Though generally these people long resisted cultural change, its impact was revealed in the 1970s in varying degrees.

Development of the Everglades. After the Seminole Indian Wars, which occurred sporadically from 1817 to 1858, interest in the Everglades centred upon exploitation of its wildlife, especially the heron and egret for their feathers and the alligator for its hide. Drastic reductions in wildlife numbers led to legislation after 1920 that protected "plume birds"; more recently, alligator hunting was similarly restricted.

Drainage remained the principal focus of engineering in the glades for many years. The Central and Southern Florida Flood Control District, established in 1947, has emphasized water control. The storage capacity of Lake Okeechobee has been increased behind a larger levee, three diked water-conservancy areas have been constructed along the eastern edge of the glades, and all are linked with an enlarged canal system and a battery of pumping stations.

Agricultural exploitation

A system of canals and lateral ditches to control Lake Okeechobee has converted about 500,000 acres (200,000 hectares) of glades land, about one-fifth of the total, to agricultural use. Muck soils were being cultivated by 1900, but early attempts to farm peat lands were unsuccessful. The discovery, in 1927, that the addition of chemical trace elements made the peat soils amazingly

productive and the development of subsoil channelling were key factors in the growth of the Everglades farming district, which, as the major supplier of many garden vegetables for Northern markets, has come to be known as the "Vegetable Kingdom." Two or three crops a year are grown, and land values are extremely high. Sugarcane growing and milling are also important. Though farms are large and highly mechanized, many labourers are needed during the winter and spring harvests; about 90 percent of the migratory labour used is Negro.

To stabilize the farming economy and utilize sugar-mill wastes, beef cattle first became important in the 1950s. Succulent grasses were planted in carefully drained and fenced pastures, and frequent spraying was necessary to combat insects.

In addition to agriculturally oriented activities, sport fishing has been facilitated by highway construction, and Everglades National Park, established in 1947 to protect wildlife habitats in the central and southern glades, attracts increasing numbers of visitors. A boon for the conservationists is the fact that oil drillers have found brine so troublesome a by-product that large-scale exploitation has been abandoned.

Prospects. On the other hand, reclamation of land in the Everglades has accentuated problems of drought during dry periods, such as that in 1970–71, yet has proved inadequate in times of abundant water. In fact, the entire regional climate may have been altered by removal of water and its stabilizing effects. Numerous muck and saw-grass fires have had disastrous effects, particularly upon wildlife habitats. The drainage of organic soils has caused sinking, and cultivation has led to accelerated oxidation of soils and to wind erosion. The average loss of topsoil from combined causes is about one foot per year, and the rate during the first years of cultivation is even higher. Subsoil drainage becomes impossible in soils less than three feet deep, and, unless new factors are introduced, it is estimated that by the year 2000 about 85 percent of the glades farmland will have become uncultivable.

Problems of ecological balance

Water control is complicated by conflicting demands: the farmer wants both irrigation and drainage; the wildlife manager needs optimum water levels for the diverse ecosystems of plants and animals, including those of Everglades National Park; and Miamians require fresh water. Conflicts of interest also arise when there are proposals for construction of an oil pipeline, an interstate highway, a jetport to serve Miami, or the exploitation of glades petroleum. The total impact of such developments upon the fragile environment of the Everglades is being more carefully weighed than in the past.

BIBLIOGRAPHY. MARJORY STONEMAN DOUGLAS, *The Everglades: River of Grass* (1947), a beautifully written and well-documented volume; ALFRED JACKSON and KATHRYN ABBEY HANNA, *Lake Okeechobee: Wellspring of the Everglades* (1948), a carefully written regional history; CHARLTON W. TEBEAU, *Man in the Everglades*, 2nd rev. ed. (1968), written by a professional historian for the lay reader; LAWRENCE E. WILL, *Cracker History of Okeechobee* (1964), a folksy, humorous, and personalized account of the development of the glades agricultural area.

(B.G.V.)